A CORPUS LINGUISTICS STUDY OF SMS TEXT MESSAGING

by

CAROLINE TAGG

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Department of English
School of English, Drama, and American and Canadian Studies
The University of Birmingham
March 2009

Abstract

This thesis reports a study using a corpus of text messages in English (CorTxt) to explore linguistic features which define texting as a language variety. It focuses on how the language of texting, Txt, is shaped by texters actively fulfilling interpersonal goals. The thesis starts with an overview of the literature on texting, which indicates the need for thorough linguistic investigation of Txt based on a large dataset. It then places texting within the tradition of research into the speech-writing continuum, which highlights limitations of focusing on mode at the expense of other user-variables. The thesis also argues the need for inductive investigation alongside the quantitative corpus-based frameworks that dominate the field. A number of studies are then reported which explore the unconventional nature of Txt. Firstly, drawing on the argument that respelling constitutes a meaning-making resource, spelling variants are retrieved using word-frequency lists and categorised according to form and function. Secondly, identification of everyday creativity in CorTxt challenges studies focusing solely on spelling as a creative resource, and suggests that creativity plays an important role in texting because of, rather than despite, physical constraints. Thirdly, word frequency analysis suggests that the distinct order of the most frequent words in CorTxt can be explained with reference to the frequent phrases in which they occur. Finally, application of a spoken grammar model reveals similarities and differences between spoken and texted interaction.

The distinct strands of investigation highlight, on the one hand, the extent to which texting differs from speech and, on the other, the role of user agency, awareness and choice in shaping Txt. The argument is made that this can be explained through performativity and, in particular, the observation that texters perform brevity, speech-like informality and group deviance in construing identities through Txt.

Acknowledgements

My thanks go first of all to everyone who, by making such an effort in contributing text messages, made this project possible. You remain anonymous and anonymised but you know who you are!

My thanks also to Anne de la Croix, for our runs together and for keeping me sane: as Haruki Murakami (and Anne) say, running certainly teaches you a lot about how to write. And a big last minute thanks to Nikki and Pam Tagg for (among other much more important things) their emergency printing services!

Above all, my thanks go to my supervisors: Pernilla Danielsson for her cheerful advice in all things computational and methodological; and Susan Hunston, for making my supervisions so inspiring, focused, challenging yet fun, and from whom I have learnt so much.

Contents

CHAPTER ONE

INTRODUCTION

CHAPTER TWO

RESEARCH INTO TEXT MESSAGING

CHAPTER THREE

LANGUAGE VARIATION RESEARCH AND THE ROLE OF CORPORA

CHAPTER FOUR

CORPUS COMPILATION: CONSIDERATIONS AND CHALLENGES

CHAPTER FIVE

RESPELLINGS IN TEXT MESSAGING

CHAPTER SIX

EVERYDAY CREATIVITY IN TEXT MESSAGING

CHAPTER SEVEN

WORDFORMS AND PHRASES

CHAPTER EIGHT

SPOKEN GRAMMAR, TEXTED GRAMMAR

CHAPTER NINE

CONCLUSION

CHAPTER ONE
INTRODUCTION

1.1 General aim of the thesis

The general aim of this thesis is to investigate the language used in a corpus of text messages. The thesis aims to go beyond the abbreviations and spelling variation associated with the medium in order to explore other linguistic features which may be said to define texting as a language variety. In achieving this aim, the thesis paves the way for further research by exploring new challenges raised by the compilation and annotation of a corpus of text messages, as well as outlining the benefits of a corpus-driven approach to the study of language variation.

1.2 Background to the thesis

1.2.1 Texting: definition and terminology

Texting refers to the sending of short typed messages between mobile phones using the Short Message Service (SMS), a feature of most mobile phones since the late 1990s. In much of the European literature into texting, the term SMS is also used to describe both the medium and the messages (e.g. Kasesniemi and Rautianen 2002). This thesis prefers the terms *text messaging* and *texting* used interchangeably to refer to both the medium and the language variety, while the term *text message* refers to the individual message sent. Use of either *text* or *message* in isolation to refer to the above is henceforth avoided, due to the varied meanings of both in linguistics, although *text* will be used as a verb to describe the process of writing or sending a text message.

Various terms are used to describe the language of texting: *textese, phone shorthand* (Sutherland 2002), *Txt* (Shortis 2007a,b), *texting* (Crystal 2008), *SMS communication* (Hard af Segersteg 2002). This thesis prefers Shortis's term Txt. Users of texting are labelled *texters*. In keeping with the spirit of Txt, my corpus of over 11,000 text messages is referred to throughout this thesis as CorTxt.

1.2.2 Texting: motivations for the thesis

The motivation for this thesis lies in the observation that, despite the growing significance of texting as a subject of academic interest, there has as yet been little attempt by linguists to describe in any depth or breadth the language used. The significance of texting is evidenced by the growing numbers sent.

Table 1.1 Number of text messages sent across the UK, 2001-2008 (from the Mobile Data Association, www.text.it/mediacentre/)

Month[1]	**No of text messages**
June 2001	0.95 billion
June 2002	1.38 billion
June 2003	1.65 billion
June 2004	2.11 billion
June 2005	2.53 billion
June 2006	3.51 billion
June 2007	4.41 billion
June 2008	6.31 billion

It could be argued on this basis alone that an understanding of how language is being used through the medium is increasingly necessary. Texting is prevalent not only in people's personal lives but also in their professional and commercial interaction. You can now, for example, receive news updates by text message,[2] be reminded of NHS appointments,[3] or be warned that your child is playing truant from school.[4] At the same time, the public debate surrounding the effect of texting on the English language and on children's literacy continues, with media reports comparing texters to 'vandals who are doing to our language what Genghis Khan did to his neighbours eight hundred years ago' (according to John Humphreys, writing in the *Daily Mail*).[5] These negative views are rebutted by Crystal (2008), in his populist although well-referenced *Txtng: the G8 Db8*, who weighs in on the

[1] Figures for December are higher: 1.31 billion in 2001, for example, due to Christmas and New Year.
[2] e.g. http://www.newsweek.com/id/159024, accessed 15th September 2008.
[3] e.g. http://www.oxfordmail.co.uk/news/2436438.patients_messaging_to_go_live/, accessed 15th Sept 2008.
[4] e.g. http://www.publictechnology.net/modules.php?op=modload&name=News&file=article&sid=17159, accessed 15th Sept 2008
[5] 'They are destroying it,' Humphreys explains, 'pillaging our punctuation; savaging our sentences; raping our vocabulary. And they must be stopped'. See http://www.dailymail.co.uk/news/article-483511/I-h8-txt-msgs-How-texting-wrecking-language.html, accessed 16th Sept 2008.

'great debate' claiming media reports to be fostering a 'huge popular mythology … in which exaggerated and distorted accounts of what youngsters are believed to do when they text has fuelled prophecies of impending linguistic disaster' (Crystal 2008: 7). The point addressed in this thesis is, however, firstly that what is being debated in the public arena does not go beyond concerns for spelling; and, secondly, that neither the media nor Crystal base arguments on investigation of substantial data.

Academic research into the medium ranges from the technological (Hasselgren et al 2003) to the ethnographic (Oksman and Turtianen 2004), and documents its mediation and alteration of established social practices such as gift-giving (Taylor and Harper 2003), its transformation of youth culture (Kasesniemi and Rautianen 2002) and its impact on people's lives including changes in interaction patterns motivated by texting (Oksman and Turtianen 2004; Ling 2008). Linguistic interest, however, remains limited with regard to the same aspects cited above in reference to the popular debate. Firstly, its main focus is largely limited to the abbreviations, spelling variation and ellipsis associated with the medium (Hard af Segersteg 2002; Thurlow and Brown 2003).[1] These features, although of interest, only scrape the surface of a true linguistic understanding of texting. Secondly, observations are based largely on interviews, diaries and observation, and only on limited amounts of texting data (Hard af Segersteg uses 1152 text messages; Thurlow, 554 text messages). Texting remains largely unexplored within corpus linguistics, discourse analysis or language variation studies. It is this gap—the lack of thorough linguistic analysis based on large corpora—that the present study aims to fill in order to increase understanding of this prevalent, controversial medium.

1.2.3 Language variation and corpus linguistics

The fields into which this study positions texting include **situational language variation** (the study of how language varies according to the situation in which it is used), within which it aims to start defining texting as a situationally-defined language variety. The importance of recognising variation in any attempt at language description is now widely accepted, with linguists such as Biber et al (1999) and Carter and McCarthy (2006) covering spoken and written varieties in their grammars. This implies the need to base language descriptions on naturally-occurring or attested data and it is for this reason that the

[1] This is added to by some interest also in the functions fulfilled through texting (Doring 2002; Grinter and Eldridge 2001; 2003; Thurlow and Brown 2003).

current study draws on the methodology and tools of **corpus linguistics.** This term covers various approaches which draw on large bodies of naturally-occurring data (corpora) to explore a range of linguistic questions, including those within the field of language variation. Given the growing popular and academic interest in texting, a corpus study can only add to our understanding of language variation, while also bringing this significant medium to the attention of corpus linguists and language variation researchers.

1.3 Objectives of the thesis

1.3.1 Three main objectives

The three main objectives of this thesis are methodological, empirical, and theoretical. In these three areas, the thesis addresses the following main points.

1. Methodological objectives

   To address the challenges and decisions raised by the compilation, storage and exploitation of a text message corpus and explore effective procedures for identifying and retrieving features for analysis;

2. Empirical objectives

   To explore the extent to which previous assertions regarding the language of texting (largely in relation to spelling variation) can be supported by investigation of a large body of text message data;

3. Theoretical objectives

   To describe features which can be said to characterise texting as a language variety.

1.3.2 Methodological objectives

The methodological procedures followed in this study are groundbreaking in the sense that a text message corpus of this size has never before been collected for the purposes of linguistic study in English. However, reasons often given for the lack of existing text message data include complications caused by the personal nature of most text messages and the difficulties which arise in trying to get texters to part with them (Crystal 2008). As well as this initial challenge to data collection, the unconventional nature of the language also complicates attempts to annotate and process text message data. This study aims to

address and resolve these challenges as much as possible and, in doing so, paves the way for future corpus research into the medium.

This study also explores the question of how to initially identify and retrieve features of interest, a question at the heart of the distinction between corpus-based and corpus-driven approaches. In choosing to adhere where possible to the latter approach, this study sets out not only to justify the decision but to serve as a model of an inductive study driven where possible by the data. In practice, the study draws on various resources. Firstly, it draws on existing literature into texting to explore, for example, spelling variation and the ways in which texting is similar to spoken language. Secondly, it uses observations made in the course of investigating the corpus regarding, for example, the somewhat unexpected (in the light of popular and academic research) occurrence of everyday creativity in the corpus. Thirdly, it exploits existing models (including Carter and McCarthy's descriptive model of spoken grammar), which are inductive in the sense of being shaped through investigation driven by other corpora. Finally, it applies corpus tools of word frequency and concordancing in order to 'delay human intervention' (to use Sinclair's 2001 expression) and allow computer-generated frequency counts to identify significant features not necessarily seen in other text-types or discussed in other research. Again, this eclectic and flexible attempt to be driven as much as possible by consideration of the corpus fills a gap in the language variation literature which, despite a growing reliance on corpora, largely comprises corpus-based studies. Whether and if so how a corpus-driven approach could proceed is a key question addressed in this thesis.

1.3.3 Empirical objectives

As mentioned above, previous linguistic investigation of texting is largely restricted to spelling variation and often based on limited datasets. One objective that this study seeks to fulfil is that of providing empirical data to substantiate and refine previously-made assertions in this area. In particular, it seeks to identify and categorise spelling variants according to their form, their dependence or independence on speech, and their presumed function in texting, in comparison to previous research into written texts. As well as having implications for the study of orthographic variation (Sebba 2007), this study may help to

dispel popular mythologies surrounding Txt, reinforced by online dictionaries with extensive lists of spelling variants allegedly used in texting.[1]

1.3.4 Theoretical objectives

The main theoretical objective is that of providing a description of the language used in texting, guided not by pre-corpus assumptions but by observations driven by the data, and drawing on the resources outlined above (section 1.3.2). The questions that emerge from these resources include the following:

- To what extent, and in what ways, does texting (as a written medium) resemble spoken language?
- Which linguistic features distinguish texting from speech, and can be said to define texting as a language variety?
- What are the motivations or situational factors behind the linguistic features?

As shall become apparent in the subsequent chapters, this study reveals a language variety shaped not only by physical constraints of texting but by users who actively exploit the medium to fulfil interpersonal functions. In fulfilling these functions, users draw on their awareness of other language varieties and discourses to construe brevity, perform spoken informality, and indicate group membership and deviance. In terms of the linguistic profile of texting, the choices open to texters will be seen to be realised in spelling variation patterns which reflect those of earlier and contemporary written texts, as well as in the adoption of speech-like grammatical features and the creative manipulation of forms associated with speech. As well as drawing on other text-types, texters also construe a language specific to texting which is, as we shall see, realised in the use of phrases and grammatical features peculiar to texting.

1.4 Outline of the thesis

This thesis begins, in Chapter 2, with an overview of the existing body of literature into texting which not only provides background information for, and possible methods to adopt in, the current study but highlights the gap which the study aims to fill.

[1] e.g. http://www.abbreviations.com/acronyms/SMS

Chapter 3 then positions the current study of texting into the field of situational language variation. It argues that the focus on mode in language variation studies risks exaggerating differences between spoken and written varieties as well as the impact of technological constraints, and neglects the agency of users and the role of function and audience in shaping language varieties. The implication for texting is to focus attention on the influence of the interpersonal over the physical constraints of the medium. The chapter also looks at the role of corpora in descriptions of language variation and highlights the need for inductive, corpus-driven investigation of single varieties, alongside quantitative comparative corpus-based analyses which dominate the current literature on variation. This, it argues, is particularly important given the unconventional nature of Txt and the possibility that features occurring in texting are different from those found in other language domains.

Turning to CorTxt, the text message corpus used in this study, Chapter 4 outlines the methodological considerations and challenges involved in its compilation, storage and exploitation. As well as many issues faced by compilers of other corpora, this chapter covers new challenges raised by the particular nature of the data and so paves the way for future text message corpus processing.

The investigation into spelling variation reported in Chapter 5 is to some extent an extension of the previous chapter, in that it raises methodological questions as to whether variants are to be considered separate entries in, for example, word-frequency counts. However, the chapter also outlines the argument that respellings create social meaning (Sebba 2007) and it explores the extent to which spelling variants in the corpus reflect or extend historical and contemporary respelling practices. In so doing, it raises for the first time in this thesis the suggestion that texters are drawing on their awareness of existing practices in other (written) text types in construing Txt. In particular, it is argued that users can be described as performing brevity, as well as spoken informality and the deviance of the group from the mainstream, through the choices they make in how they spell.

Chapter 6 suggests that creativity in texting is not limited to variation in spelling, but is realised also in the occurrence of everyday creativity such as idiom manipulation and repetition, similar in form and function to that found in spoken interaction. Although this is on the surface surprising, given the constraints of the medium, it will be argued that the

creative manipulation of forms is particularly appropriate for texting, given the dearth of other resources available to users in fulfilling the expressive, interpersonal functions they wish to fulfil through the medium. Nonetheless, this creativity highlights the fact that texters draw not only on a metalinguistic awareness evidenced by this creative manipulation of language but, more specifically, on their awareness of spoken language features. In other words, texters appear to draw on both written and spoken texts in performing texting identities.

In Chapter 7, however, investigation of the word-frequency list compiled for CorTxt highlights differences between texting on the one hand and both spoken and written corpora on the other. The unusual order of the most frequent texted words cannot wholly be explained either by reference to the omission of *the* nor to the speech-like structures of texting, but must instead be addressed with reference to the phrases which distinguish texting from other varieties. This profile or fingerprint, to use O'Keeffe et al's (2007) term, can be explained through the situational factors of texting and in particular the functions to which the medium is put, again highlighting the agency of users in shaping the language and the fact that texting must be considered not only through similarities to speech but as a variety in its own right.

Chapter 8 takes a different approach to the investigation, by exploring texted grammar through the application of existing grammatical models. On the basis of comparisons between texting and speech (identified in the literature and through exploration of CorTxt), this chapter argues for the need for a grammatical model which takes account of spoken features, but also one that is grounded in the data it aims to describe. Carter and McCarthy's (2006) descriptive model, although text- rather than corpus-driven,[1] emerges from close analysis of spoken data. Its application to texting not only further strengthens its validity but highlights similarities and differences between spoken and texted grammar, suggesting again that texting be considered a variety distinct from comparisons with spoken language.

The concluding Chapter 9 draws these findings together, and outlines the methodological and theoretical contributions made by this thesis. In terms of methodology, the thesis

[1] The distinction made in this thesis between a text-driven and a corpus-driven approach is explained in Chapter 3 of the thesis (and draws on Bednarek 2006).

addresses issues faced in the compilation, storage and exploitation of text message corpora, and it serves as an example of a corpus-driven investigation of a language variety as well as an application of Carter and McCarthy's descriptive model of spoken grammar. On a theoretical level, the thesis not only substantiates and refines previously-made assertions regarding spelling variation in texting but argues that texted creativity extends to everyday, speech-like creativity, and goes on to highlight differences between spoken and texted language which suggests that texting be explored not as existing somewhere on a cline between speech and writing but as a variety in its own right which draws on users' awareness of other text-types. The suggestion is then put forward that texting can be described in reference to *performativity*, taken from gender theory (Butler 1990), and the notion that texters are performing brevity, spoken informality and group deviance. This accounts for the linguistic choices and variation seen in texting and the agency of users as well as their creative manipulation of language.

CHAPTER TWO
RESEARCH INTO TEXT MESSAGING

2.1 Introduction

The relevant literature on text messaging can be divided into several broad overlapping areas. One area deals with texting as a feature of an emerging mobile (youth) culture in many European countries (e.g. Taylor and Harper 2003; Oksman and Turtianen 2004). A second set of studies involves data-based investigation of the functions, language and structure of text message interaction (e.g. Hard af Segersteg 2002; Ling and Baron 2007). A third area seeks to determine text message authorship as part of forensic linguistic research (Coulthard 2004; University of Leicester 2006). In the final area explored here, researchers apply the findings of text message analysis to the search for ways to enhance the speed and efficiency of mobile technology (e.g. Hasselgren et al 2003; How and Kan 2005).[1]

Gaps in the above literature, however, point to the need for a corpus study of Txt, the language of texting. Firstly, aside perhaps from Fairon and Paumier's (2006) study into French language text messages, there has as yet been little attempt to conduct a thorough linguistic analysis of texting as an emerging language variety (or to design a corpus to do so), and there has instead been an overriding concern with texting orthography: that is, with the abbreviation and phonetic spelling that characterise the messages. A further noticeable gap is the lack of research into text messages written in English, as most sociocultural and linguistic studies emanate from other European countries. There is also a lack of corpora collected in Britain, as the only large corpus of messages in English is held at the University of Singapore[2] and was collected locally. My corpus and research serve to fill these gaps in the existing research. At the same time, however, the existing literature into texting is useful in that it serves to depict interpersonal features of the situation in which text messages are produced and begins to suggest ways in which these shape the language produced.

[1] Another possibly distinct area is the analysis of texting as part of studies into wider changes in orthographic practice (Shortis 2007a,b), but this is discussed in Chapter 5.
[2] See http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/

2.2 Sociocultural Role of Texting

This multidisciplinary field emerged during the late 1990s, incorporating communication, media and cultural studies, sociology, geography, anthropology, economics, political science and the information sciences in attempts to explore mobile and computer-mediated media and their impact on social practices, interaction, and information exchange.[1] Theoretical perspectives and methodologies taken from both the social sciences and humanities are thus employed in the interpretation of the effects of the new media. The following overview of work by Kasesniemi and Rautianen (2002), Ling and Yttri (2002), Harkin (2003), Ling (2005, 2008), Madell and Muncer (2002), Taylor and Harper (2003) and Oksman and Turtianen (2004) highlights the social impact of texting, as well as the way in which it has been integrated into existing social practices, especially those of young people. This is reinforced by the observation in the literature regarding the integration of texting with other forms of communication. Although not based primarily on text message data, their work also begins to explore how the way in which texting is used shapes the apparently speech-like and abbreviated language which emerges.

[1] See, for example, the journals *New Media and Society* and *Personal and Ubiquitous Computing*.

Table 2.1 Sociocultural studies of text messaging

Author(s)	**Country**	**Subjects**	**Data**	**Points made**
Kasesniemi and Rautianen 2002	Finland	Teenagers	Corpus of nearly 8000 messages collected since 1998 as part of a project into the Finnish mobile culture (drawn on more heavily than in Oksman and Tautianen 2004) Interviews, field notes and journals compiled by the participants	Social-scientific study into the communicative practices of the mobile technology • The adolescent culture of text messaging grows out of the circumstances of teenagers' lives, which the authors describe as 'comparable to the hectic communication routines in the tightly scheduled lives of business executives' (p175). • Through the practices of message collecting, sending chain messages, and the collective reading and composing of messages, teens 'collectively engage each other and sustain their teenage relations' (p182). • The popularity of text messaging is not, they report, hampered by the physical constraints of the medium. Kasesniemi and Rautianen note occasions when friends would help compress what the sender wanted to say, and how this serves to heighten group solidarity. • An example of a shift in social practice caused by texting is its use in initiating or even maintaining relationships at times which allows shy teens to act two different lives. An extreme case involved a couple who met daily at school but only acknowledged their deeper relationship through text messages. Furthermore, a text message 'finds its way to times and places where a call would be impossible' (p171).
Ling and Yttri 2002	Norway	Different age groups with heavy focus on teens	Interviews	The use of mobiles phones and text messaging in coordinating activities as well as more expressive functions. • Evidence of an 'adoption cycle at a social level' (p142), with the initial use of the phone in solving the specific problem of security developing into various types of coordination (or micro-coordination) as its use becomes routine and then into a more expressive interaction (hypercoordination).

				• Micro-coordination involves the mundane coordination of everyday activities, seen for example in messages asking a partner to stop at a shop on the way home, as well as the progressively exact arrangement or adjustment of appointments and meetings as the need arises (p139), such as texting from the bus stop to let a friend know you'll be late. • Hypercoordination, used by teenagers, is a more expressive function including 'chain text messages, and the use of texts to initiate friendships and romances and which serve to quantify popularity and strengthen group ties. • The use of slang (homophones, cognates and abbreviations), encouraged partly by the constraints of the character allowance, serves to heighten group solidarity and give teens a separate identity from their parents. • Texting plays a crucial part in 'the culture of concealed use': that is, the understanding of when it was socially acceptable to use a mobile. The authors use Goffman's (1959) concept of 'front-stage' activities to suggest that, with text messaging, teens could simultaneously operate on parallel front stages: in class, for example. • The role of the phone in marking boundaries between child and parent: The mobile phone, Ling and Yttri suggest, highlights the importance for teens of being accessible to their social network, and being able to bypass their parents in a way in which a fixed home phone cannot. • The authors also note the popularity of text messaging over voice calls, due in part to the reduced financial cost, but also to the careful and controllable context of asynchronous communication which allows for reflection, and limits loss of face.
Harkin 2003 (backed by independent	Britain	Four focus groups of teenagers	Discussion and interview with the four focus	Study of mobilisation, 'the process by which mobile technologies are folding themselves into the fabric of our economies, social lives and communities' (p9).

think tank, Demos)		and adults.	groups which each contained 6-8 participants.	• The intimacy of texting and the controlled, carefully considered language allows greater connection between shy or less expressive people. • The phone plays a role in reaffirming the notion of 'locality' is reflected in the fact that people text individuals whom they see regularly and use unconventional language which is incomprehensible to outsiders.
Taylor and Harper 2003	Britain	Teenagers (16-19 year olds at sixth form college)	Field notes Interviews (collected over 10-week period) No text message data	Social and anthropological concepts of gift-giving • The success of mobile phones can be explained by their role in young people's existing exchange ceremonies, or gift-giving, in which phones, credit and text messages act as currency, as well as objects of attention and topics of conversation (see also Grinter and Eldridge 2001; 2003 who note a similar practice with teens' address books which contain more numbers than they actually use). • As such, phones and texting can confirm and strengthen alliances and friendship. • The value placed on phones can depend on the form of the language: not sticking to punctuation rules or abbreviating the language so as to avoid sending more than one message can lower the value assigned to the message. • The future design of mobiles and mobile applications needs to acknowledge the fact that phones operate within existing social practices. For example, the fact that text messages are easily assigned meaning and therefore value throws a new light on memory cards being developed on which messages can be stored; while voice messages need to become more interactive and user-friendly in order to be considered special.
Madell and Muncer 2004	Britain Teeside	1340 children	Questionnaires	Focus on Internet use • Comparison of children's Internet and mobile phone use found the Internet to be 'less important' as the mobile phone becomes more

				important and begins to supplement functions fulfilled online.
Oksman and Turtianen 2004	Finland	Teenagers	168 thematic individual and group interviews Observation (2000-2002) Corpus of nearly 8000 messages (see Kasesniemi and Rautianen 2002)	Symbolic interactional approach (e.g. Goffman 1959): 'if interaction is guided by a 'frame' which decrees what is acceptable in any communicative situation, what does that established by mobile communication constitute?'; 'Are new forms of social interaction being produced?' • Finnish teenagers use mobile communication to maintain social networks through rapid interaction with friends, and to form new relationships or to initiate 'courtships', as well as in construction of their social identity through articulation of their personal space (away from parents), in defining their relationships with others (or establishing their membership of a community) and their presentation of self. • The intimacy of a text message over a phone call is reflected in the language used, which incorporates plays on language, spoken expressions ('written orality'—Fortunati 2001: 314) and newly-coined phrases. Despite the restrictions of the character allowance and screen size, text messages feature a range of emotions and expression.
Ling 2005	Norway	Teenagers	Questionnaire administered to 11,928 pupils in February 2002	Explores the role of the mobile phone in teens' 'emancipation' process, and in particular the distancing from parents, peer-group integration, and deviant activities such as drinking, truancy and narcotic use. Finds that SMS is more 'normal' than voice telephony, which correlates more strongly with extreme boundary-testing behaviours.
Ling 2008	Norway	Teenagers	'Observation method' (following Goffman 1959)	Uses work into social ritual by Goffman (1959) and Collins (2004) to explore how rituals are mediated, and social ties strengthened, by mobile telephony, exploring the impact which mobile communication has on interactions with those who are physically co-present.

The above overview of sociocultural research highlights the observation made by social scientists that mobile phones in general and texting in particular are used to mediate *existing* social practices within localised communities, such as gift-giving, and thus that existing social practices shape technology use rather than the reverse. Oksman and Turtianen (2003: 332), for example, describe 'its natural, almost organic niche in everyday life'. This means that texting is used for informal and everyday purposes: as Kasesniemi and Rautianen (2002: 170) put it, texting has quickly become a 'means to deal with everyday life, rather than a service for special user groups'. As a further illustration of its integration into everyday lives, texters tend to text while engaged in other tasks or interactions (Grinter and Eldridge 2001; 2003; Ling 2008). At the same time, due to texters' intimate relationships, a social value is placed on otherwise ephemeral, quickly written messages. The implication of the above is that, alongside the technology, social practices and interpersonal factors are likely to shape Txt.

The integration of texting with other forms of communication, including electronic and face-to-face interaction, is another recurrent theme in the literature (Madell and Muncer 2004; Oksman and Turtianen 2004; Faulkner and Culwin 2005; Ling 2008). For example, in their study of Internet use, Madell and Muncer (2004: 366) suggest that the internet is becoming less important to young people as the mobile phone increases in significance and supplements functions previously fulfilled online. Texting has all the advantages of online communication (controlling self-presentation, diminishing geographical distance, creating an arena for personal space) whilst avoiding any sense of disembodiment because participants are intimates, who already share close, personal relations established through other communication channels (Madell and Muncer 2004). According to Oksman and Turtianen (2004: 332), differences between texted and computer-mediated communication lie in the fact that mobile communication is 'more grounded socially than the virtual encounters of the internet'. The choice as to which means of communication to use is of course made according to which is deemed appropriate for the context and purpose of the interaction, and which is available (Grinter and Eldridge 2001; 2003; Oksman and Turtianen 2004: 328). Different forms are also integrated: mobile and face-to-face interactions may be conducted simultaneously (Ling 2008), or text messages used to arrange face-to-face or phone conversations (Grinter and Eldridge 2001; 2003). Grinter and Eldridge suggest that following up a text message conversation with an arranged meeting constitutes a way of 'finishing' the exchange (see also Oksman and Turtianen 2004: 328).

These findings emphasise how texting is integrated into other forms of communication, electronic as well as face-to-face, and embedded more deeply into existing practices than computer-mediated communication (CMC) which is less well grounded in everyday lives.

Observations made in the above studies regarding the nature of Txt are relevant in that they begin to link linguistic features to characteristics of the situation in which text messages are produced, with the result that the highly abbreviated, informal language often criticised in the press[1] emerges as an appropriate response to the medium. Firstly, the informal and intimate nature of texting encourages the use of speech-like language.[2] Oksman and Turtianen (2004), for example, note how the intimacy of texting encourages spoken expressions or the creation of new phrases; while Kasesniemi and Rautianen (2002: 187) describe meandering 'natural conversations' which are 'contemplative, impulsive and meandering' and with certain turn-taking expectations to reflect those in spoken language.[3] Secondly, the abbreviation of words, including the omission of inflections in certain languages such as Finnish (Kasesniemi and Rautianen 2002: 183) and the omission of spaces as well as homophones and cognates, is described as being encouraged by the intimacy of texting, as well as by the speed at which messages are sent and responded to, and by teenagers' use of texting in defining group boundaries. The fact that participants are intimate, and interact through other means of communication including face-to-face, ensures their understanding of the code, while those outside the group do not (Kasesniemi and Rautianen 2002: 183; Ling and Yttri 2002: 162; Harkin 2003). What is important about these observations is that the focus of most studies is on the influence of these social and interactive factors, rather than on physical constraints of the medium, in shaping Txt.

The implication is that interpersonal factors can override technological constraints. For example, although physical limitations of the medium are seen as necessitating abbreviated forms, this is not felt to restrict the emotions and expressions conveyed. Instead, messages are described as creative, showing a variety of styles and an awareness of audience (Kasesniemi and Rautianen 2002; Oksman and Turtianen 2004). Despite anxieties from Finnish teachers as to the effect of texting on teenagers' writing skills, Kasesniemi and

[1] See Chapter 1 (and Thurlow 2006; Crystal 2008a)
[2] And vice versa, according to Harkin 2003, in that the limited space and need to abbreviate allows for a more controlled personal expression.
[3] Although Grinter and Eldridge's (2001; 2003) data-based study (outlined below) found that most of their data comprised 'single' messages, rather than those that were part of longer 'threads'.

Rautianen highlight the fact that texting provides opportunities for 'purposeful and emotional' writing, particularly for boys who are more resistant to writing than girls, with teens reflecting on how to encapsulate what they want to say and the possible interpretations of each message.[1] These observations challenge the presumed prevalence of abbreviation. For example, text messages can be more or less abbreviated depending on who the intended recipient is: teens write more formal text messages to teachers and parents, but use 'slang, plays on local dialects, puns and insider vocabulary' with close friends (Kasesniemi and Rautianen 2002: 185). Kasesniemi and Rautianen (2002: 185) also quote girls as saying they must compose text messages to boys in 'plain language', that is, devoid of abbreviations, 'references and suggestions'. In other words, we begin to see in these studies the realisation that texters have choices in how they text, dependent not only on physical constraints but interpersonal considerations.

Limitations of these studies for present purposes are, however (as well as the fact that most cited above are not concerned with English), that they do not go much beyond a relatively limited focus on noticeable features of Txt: largely, abbreviations, omitted spaces and phonetic spellings. Furthermore, due to the sociological purposes of the studies, their methodology involves interview and observation, rather than analysis of actual data. Data-based studies of texting are outlined below.

2.3 Linguistics of Texting

2.3.1 Overview of the research

Several studies adopt (to varying extents) a data-based approach in the sense that descriptions of communicative functions and linguistic features of texting are based on analysis of text messages themselves (Doring 2002; Hard af Segersteg 2002; Kesseler and Bergs 2003; Grinter and Eldridge 2001; 2003; Thurlow and Brown 2003; Faulkner and Culwin 2005; Fairon and Paumier 2006; Ling and Baron 2007; Crystal 2008a).

[1] The freedom to use abbreviation and spelling variation in text messaging is discussed further in Chapter 5, where it emerges as a resource for creating identity rather than a constraint on personal expression (Shortis 2007a,b).

Table 2.2 Research into the Linguistics of Text Messaging

Study	Country	Subjects	Collection method	Data	Study focus
Doring 2002	Germany	200 participants 1000 messages 13,357 words	Not stated.	1000 messages (13,357 words)	Abbreviations and communicative functions
Hard af Segersteg 2002	Sweden		Various sources: an anonymous webpage, from messages forwarded from volunteers and from family and friends	1152 messages (17,024 words)	Abbreviation and grammatical ellipsis
Grinter and Eldridge 2001, 2003	Britain	10 participants (5 boys and 5 girls) aged 15-16	Record logs in which participants recorded messages sent and received over 7 consecutive days	477 messages (236 sent and 241 received)	Communicative functions, message singles and threads, abbreviations
Thurlow and Brown 2003	Britain	159 British university students, aged around 19	Five text messages transcribed and submitted by 135 students at end of a university lecture	Corpus (544 text messages)	Abbreviations, communicative functions (plus message length)
Faulkner and Culwin 2005	Britain	Adults in 20s	Recruitment by students of acquaintances and of customers in mobile phone shop	Questionnaire (565 respondents), and diary study including text messages (24 computer studies students)	Communicative functions, gender differences, integration and comparison with other technologies
Fairon and Paumier 2006	Belgium	Up to 3,200 participants aged 12 to 65.	Project was broadcast nationally (October and December 2004) and participants requested to send texts to a free mobile number.	30,000 messages (from an initial 75,000) (sociolinguistic information including 'ability to decrypt SMS' requested)	Compiling and preprocessing the corpus, and translating the language into 'standard' French, for future study.

2.3.2 Spelling variation and abbreviation in texting

Most data-based investigations of text messages focus primarily on spelling variation, and for this reason findings are reviewed at greater length in Chapter 5 of this thesis. As that discussion will show, as well as identifying and categorising types of abbreviation found in their data, the main findings of linguistic explorations of text messages is that the use of abbreviation is in fact limited (Doring 2002; Grinter and Eldridge 2001; 2003; Thurlow and Brown 2003; Faulkner and Culwin 2005); and that where it occurs it is, as found in the sociocultural studies, appropriate and creative (Hard af Segersteg 2002; Thurlow and Brown 2003). However, the failure in most cases to go beyond this concern for unconventional spelling creates a gap which the present thesis aims to fill, as does their relatively limited corpus size.[1]

2.3.3 Communicative functions

Some of the above studies, however, also explore the communicative functions of texting, as summarised below. These studies show that, despite great variation in purposes fulfilled through texting, most are highly interpersonal.

Table 2.3 Communicative functions across the literature

Study	**Communicative functions**
Doring 2002	Communicative functions based on Brinker's (2001) text type theory include: • Contact function (54%), incl. appointments and greetings and emotional support. • Information • Appeal • Obligation • Declaration Average message length: 78 characters (2-160)
Grinter and Eldridge 2001; 2003	Four types of text message conversation identified: • Communication coordinations (24%) • Planning activities (25%) • Chatting (39%) • Other (jokes, greetings) (12%)

[1] See Table 2.2 above and Chapter 4 on corpus compilation.

Thurlow and Brown 2003	Communicative function determined using Content Analysis (Krippendorf 1980),[1] showing that messages are primarily relational or interactively oriented. 1. Informational-Practical Orientation: *Put money in ur account* (14%); 2. Informational-Relational Orientation: *I'm not feeling v well can you get the lecture notes for me please* (8%); 3. Practical Arrangement Orientation: *Wanna come to tesco?* (15%); 4. Social Arrangement Orientation: *R WE DOIN LUNCH THIS WK?CHE* (9%); 5. Salutory Orientation: *Yo man whats de goss* (17%); 6. Friendship Maintenance Orientation: *Don't worry bout exam!Just had hair cut & look like a ginger medussa!Arrgh!* (23%); 7. Romantic Orientation: *Each time ur name appears on my phone i smile like this :)* (9%); 8. Sexual Orientation: *Read ur email-thought waz gonna burst so horny xxxxxx* (3%); 9. Chain messages: *I believe friends are like quiet angels who lift our feet when our wings forget how to fly!send to 4 friends and sont send back and see what happens in 4 days* (2%).
Faulkner and Culwin 2005	Messages classified into following groups, with questions, sign offs and dates the most frequent (and rendezvous the least popular): • adverts • questions • rendezvous immediate and ongoing • rendezvous near future • events • instructions • reminders • jokes • sign on • sign off • gossip • dates • information—personal • information—commercial • information—operational

The predominance of intimate and interpersonal functions is shown, for example, by the fact that only 31% of Thurlow and Brown's (2003) messages were considered to be of 'low

[1] Thurlow and Brown (2003) point out that although the messages were classified where possible into one of the nine categories, 22% fell into more than one category, which were anyway 'not strictly exclusive of each other'.

intimacy, high transactional orientation' (informational-practical, practical arrangement, and chain messages) but 61% of 'high intimacy, high relational orientation' (social arrangements, salutary, friendship maintenance, romantic and sexual messages), with informational-relational messages embodying elements of both. In other words, around two-thirds were *explicitly* relational or interactional, with others perhaps more implicitly so.[1] Faulkner and Culwin (2005), however, who focus not on adolescents but on young adults, found transactional functions such as questions, sign-offs and arranging dates to be the most frequent, in contrast to the studies of teenage use. They suggest that this older age group showed more restraint and followed different rules which constrained the variety of uses to which they put their text messages. Although Faulkner and Culwin's study is limited to adults in their 20s, it serves to complement the generally teen-oriented body of other research and highlights the varied purposes that texting fulfils, according perhaps to the age and other social factors of participants.

One study which seeks to overcome both the limitation of size and study focus is Fairon and Paumier's (2006). Fairon and Paumier's (2006) compilation of a corpus of 30,000 French text messages is motivated by the lack of text message data necessary to explore the phenomenon of 'new habits' in text messaging in order to develop text processing software able to parse non-standard texts. The goal of the project is therefore to provide a reference corpus of text messages representing texting in its diversity. As the authors point out, this remains the largest number of authentic text messages so far collected, and from the widest range of contributors (Fairon and Paumier 2006: 4). As yet, however, although the fact that the spellings are translated into 'standard' French indicates an interest in other linguistic features, use of the corpus as the basis of a thorough linguistic attempt to define the nature of this speech-like informal language variety is just beginning and, with relevance to the present study, limited to messages in French.

Age constitutes a final gap or limitation in the literature, in that participants across studies are predominantly young (Grinter and Eldridge 2001; 2003; Kasesniemi and Rautianen 2002; Ling and Baron 2007). As well as creating a large database of text messages for linguistic investigation, the current study addresses the age imbalance by recruiting

[1] Chain messages, for example, if thought of as 'gifts', can also be considered to have a predominantly relational function in that they 'communicate some desired aspect of identity, but also [act] as means of social bonding through (potentially) shared humour and taboo breaking' (Thurlow and Brown 2003: 11; see also Ling and Yttri 2002: 159; Taylor and Harper 2003).

participants aged 19 to 68. This is discussed further along with other corpus compilation issues in Chapter 4.

2.4 Forensic Research

Linguistic investigation into the authorship of text messages, which has provided evidence in several court cases, rests on the assumption that authors of any text can be identified through distinct linguistic features of their *idiolect*: the individual use of language which emerges from speakers' preference for particular lexical items (Halliday et al 1964: 75; Abercrombie 1969 in Coulthard 2004: 432). An example of a criminal case in which the author's co-selection of certain items allowed a written text to be attributed to him was that of the *Unabomber* (so-named because he sent postal bombs to *un*iversities and *a*irlines), whose family recognised his idiolectal preference for certain words and phrases in the thesis he published in return for a cessation of attacks (Coulthard 2004; 2007). Text messages constitute particularly fertile ground for determining authorship because, according to Coulthard (2007), the fact that texting is subject to 'fewer controls'[1] means that less data is needed to ascertain idiolect than in other forms of speech or writing.

The first case in which linguistic analysis of Txt was used as evidence (in the UK at least) involved the murder of Danielle Jones in 2002 (Coulthard 2007). Her uncle, eventually convicted for her murder, told police that Danielle had sent him two text messages, after the time she was in fact found to have died, and Coulthard's linguistic analysis revealed substantial differences between them and seventy other text messages written by Danielle to which Coulthard was given access. The first of the two 'forged' text messages was as follows.

Figure 2.1 Message allegedly sent by Danielle Jones (Coulthard 2007)

HIYA STU WOT U UP 2.IM IN SO MUCH TRUBLE AT HOME AT MOMENT EVONE HATES ME EVEN U! WOT THE HELL AV I DONE NOW? Y WONT U JUST TELL ME TEXT BCK PLEASE LUV DAN XXX

[1] By which, Coulthard means that there is less pressure to follow prescriptive conventions in texting and so more room for individual variation in orthography (Coulthard pers comm; see also Shortis 2007)

The most striking difference between this text message and Danielle's preferred style, notes Coulthard (2007), is its case: sixty-six of the seventy text messages sent by Danielle were in sentence case rather than capitals. Nor did the abbreviations match those generally favoured by Danielle, who tended to use *wat* rather than *wot*; *at da mo* rather than *at moment*; *1* rather than *one*; *hve* rather than *av*; and who signed off using various spellings of *text back please* but never the one above. As it turned out, other evidence proved vital in convicting Danielle's uncle: the text messages between their phones were sent from his house, and he sent his second text message twice, presumably sending it for the second time because he could *see* the first had not arrived but, unfortunately for him, it did arrive much later. However, Coulthard's linguistic analysis shows that Danielle's text messages displayed a distinct orthographic idiolect which differed from the style her uncle used.

Interesting questions emerge from this case as to whether it can be said that texters in general have identifiable orthographic idiolects and whether these can be used to attribute authorship of any given text message. Although studies of Txt certainly point to a substantial amount of variation in spelling (see Chapter 5), and although these have been explained as resulting from active choices made by individual texters (Shortis 2007a,b), it is not yet certain whether the choices texters make are consistent enough, in the face of other variables, to constitute orthographic idiolects.

The University of Leicester's (2006) Forensic Linguistic study recognises these issues, in that it explains orthographic choices as *texting strategies* (defined below), adopted to suit particular circumstances, and explores other choices as well as spelling variation.[1] However, there remains an apparent assumption that individuals are consistent in their orthographic choices. Participants in the study were asked to submit 10 of their own messages to a website (http://www.le.ac.uk/pc/aa/ked6/index.html) and the researchers hoped to recruit at least one hundred texters in order to analyse 'linguistic consistency and variation in individuals' texting style' and, by means of a sampling technique known as *snowballing* in which participants are encouraged to urge friends to contribute, to also explore 'the influence of peer groups upon writing style and texting language'. This in itself acknowledges the fact that texters' style will vary and is dependent on external factors such as intended audience, which the researchers address by looking at what were introduced

[1] The main researcher in this study is Dr Tim Grant, now head of the Forensic Linguistics Centre, Aston University, Birmingham. No publications have yet been produced from the study (Grant pers comm).

above as texting strategies: that is, texters' responses to technological constraints or other external factors resulting in, for example, structural ellipsis, which they suggest is encouraged by predictive texting; or vowel deletions and homophones which constitute part of 'more abbreviated text language' (University of Leicester 2006). Although it seems reasonable to suggest that analysis of which strategies individuals prefer 'may help to identify the writer of a text', the assumption that these 'texting strategies are likely to remain relatively constant within individuals' is not yet proven. Nonetheless, this innovative field highlights individual choice and style, and the distinctive nature of Txt, as well as a practical need for greater understanding of it.

2.5 Enhancing the efficiency of mobile technology

Text message research emerging from the fields of computer science and, in particular, of human and computer-interface interaction, raises again the issue of physical and time constraints on the otherwise informal, speech-like language. These studies are driven largely by attempts to improve the speed and efficiency of the mobile keypad, and highlight the need for a greater understanding of Txt in doing so.

2.5.1 Existing methods of text-entry

Existing features of text-entry form the starting-point for attempts to facilitate the process. The problem is that there are generally only eight buttons on a mobile keypad onto which 26 letters of the alphabet are stored, [1] which means that more than one letter must be stored on each button. One method for entering text is *multi-tap* where more than one key tap is required to distinguish certain letters: for example, one tap produces *a*, two taps produce *b* and three produce *c*: the word *you* requires 8 taps, and *tomorrow* 18. The problem with this method, as illustrated above, is the high keystroke per character, or KSPC (MacKenzie 2002). *Predictive text-entry devices* address this by using an in-built dictionary to predict the most likely letter required with every tap, based on the sequence so far and on frequency lists: in other words, the phone is effectively trying to guess the word as it is entered (Dunlop and Crossan 2000; Haestrup 2001).[2] However, these do not generally suggest words longer than the sequence tapped so far and so cannot achieve a KSPC of less

[1] That is, until the introduction of technologies such as the iphone, which use a QWERTY keyboard, and other slimline phones which can be opened up to access larger keypads.
[2] Commerical examples include T9 (Tegic—Grover et al 1998), eZiTAP (ZiCorp 2002) and iTAP, used by Motorola.

than 1. A further problem is ambiguity, as multiple words generally correspond to any one letter sequence. If the word selected by the device is not that required by the user, they must scroll down a list of words arranged in order of their frequency in the corpora used, and this increases KSPC. The other problem is that, as linguistic and forensic investigation shows, Txt is sufficiently distinct to warrant the use of text message corpora in the compiling of such lists, and this is only beginning to be explored by predictive text device designers. LetterWise, also produced by Eatoni Ergonomics (MacKenzie et al 2001), uses letter digram possibilities to predict sequences rather than words, which avoids the problems of dealing with words not in the dictionary. However, whether or not users select predictive texting, and features of the device, will not necessarily shape Txt. Hard af Segersteg (2002) finds predictive texting responsible only for certain typos and the splitting of compound words, while Ling (2007) suggests that 'chattiness' and message length were only slightly increased through its use. Similarly, the drive documented below to increase speed and efficiency will not necessarily entail alternation to the nature of Txt in the face of user variables.

2.5.2 Attempts to enhance text-entry methods

Ongoing attempts to enhance text-entry methods include dictionary-based disambiguation algorithms and keyboard remapping (Hasselgren et al 2002; Gong and Tarasewich 2005; How and Kan 2005). These studies, despite their focus on increasing speed and efficiency rather than their impact on the nature of Txt, highlight the practical need for a greater understanding of Txt, and the use of text message corpora in this growing field.

Table 2.4 Attempts to enhance text-entry methods

	Overview of study	Limitations
Dictionary-based disambiguation algorithms		
Hasselgren et al 2003	The authors report on a method for improving text-entry which uses bigram probabilities to enable accurate prediction of subsequent words (known as HMS). The idea is to supplement dictionaries with word and bigram probabilities. The bigram frequencies were extracted from a Swedish news corpus, Stockholm-Umea and stored in the phone's memory, with greater weight given to shorter words. The method was implemented on a software-simulated mobile keyboard. When compared with T9 on Nokia 3410 in a trial involving participants typing invented messages and news stories, HMS was found to use fewer keystrokes, attributed not only to enhanced prediction but the fact that it predicted words in advance.	As pointed out by the researchers, the method would show greater advantage if trained on a corpus representative of Txt, thus highlighting again the need for understanding how language is used through the medium in order to enhance the interface. In the absence of a suitable corpus, however, Hasselgren et al (2003) planned to collect data from web-based chat sites.
Keyboard remapping		
Lesher et al 1998 Hirotaka 2003 Saied 2003	Various attempts have been made to remap the keyboard in order to reduce the likelihood of ambiguous tapped sequences, based on word dictionaries and statistics.	Raises the issue that remapped keyboards damage novice usability and user performance.
Gong and Tarasewich 2005	The authors explore the possibility that an optimized but alphabetized design can improve user performance yet avoid damaging ease of learning. Their usability test involved three possible designs: the standard one currently in use, the unconstrained proposals (above) and the alphabetised. The optimal unconstrained design was found through a Genetic Algorithm-based heuristic technique and considered in part to be one with a high *disambiguation*	The wordlist used by Gong and Tarasewich (2005) was derived from spoken discourse in the BNC and, as this thesis will suggest, may not prove appropriate for texting. The authors acknowledge this, flagging the creation of

	accuracy (an increased probability that, if a sequence is tapped, the correct word appears first time - in effect, meaning that the most frequently-occurring word must be the first choice offered by the dictionary). Although the researchers report 'encouraging' results, and increases in speed suggest that the constrained keypad was easy to learn and remember, and quicker than the unconstrained, there was little difference in terms of speed between their remapped keypad and the standard.	a 'common SMS word list' for more realistic future trials.[1]
How and Kan 2005	This research is based on the authors' 10,000-word corpus of text messages (University of Singapore). Measurements first made of the efficiency of the existing methods: the standard text predicting device was found to require 74.004 taps and take 59.7 seconds, compared to 118.925 (79.3 seconds) with the multi-tap. Attempts then made to optimize text-entry by remapping the keyboard and by predictive word completion. The latter is based on predictions from previous words (so that *home* can be predicted to follow *at* + *in*) and the choices given then ordered by probability according to the text message corpus. It was found that both methods increased text input efficiency, and that the greatest improvement came with a combination of the two methods, although the authors admit that that experts who looked less at the screen were likely to benefit less, and that their findings rest on the assumption that users make no mistakes.	Limitations include the fact that only five messages were taken from the corpus to be used by a small number of subjects; that they only go as far as word frequency; and (for the present purposes) the fact that their corpus would be representative of Singaporean English. Nonetheless, this work draws a little at least on naturally-occurring texted data and highlights another practical use for text message corpus analysis.

[1] The experiment involved eight students who were deemed novice users of text messaging with a median age of 29 and who were asked to enter six testing phrases using the different interfaces. The phrases, obtained from large databases of sentences compiled by MacKenzie and Soukoreff (2003), all contained the same number of ambiguous words for each keypad design, but were not necessarily similar to messages normally sent. They included: 'the generation gap gets wider'; 'sad to hear that news'; 'never mix religion and politics'.

2.6 Chapter Summary

Research into texting is chiefly relevant for its identification of features of the medium which shape the language used. Despite the emphasis on technological constraints in attempts to improve text entry, other research areas focus attention on user-related variables. Two areas of text message research, sociocultural studies carried out by Taylor and Harper (2003), Kasesniemi and Rautianen (2002) and other social scientists, and the data-based linguistic research of Hard af Segersteg (2002) and Thurlow and Brown (2003) among others, show many overlaps in terms of conclusions regarding text message interaction and language. They highlight the adaptation of texting into existing social practices and the effects of this on the way the technology is used and on the language produced, its varied role as an informal and intimate means of communication, the mundane and expressive functions fulfilled, message length and speed of production, variation across text messages in terms of both function and forms, and (as discussed in later chapters) the appropriacy and creativity of unconventional spellings. Forensic research into texting idiolect complements this, by highlighting the role of individual style, together with varying situational factors, in shaping Txt. At the same time, however, the four areas of research reveal similar gaps: the focus (where language is concerned) with non-conventional orthography; lack of any thorough linguistic analysis of texting as a language variety based on a large corpus; and a relative shortage of research into English-language text messages (because studies emanate from other European countries as well as a general lack of data). The need to address these gaps is due not only to the practical applications outlined in this chapter, but the necessity of understanding texting as part of language use. The present study fills the gaps, firstly through the compilation and documentation of a corpus of over 11,000 messages written in English and, secondly, by looking beyond the unconventional orthography to determine which other features define texting as a language variety.

CHAPTER THREE
LANGUAGE VARIATION RESEARCH AND THE ROLE OF CORPORA

3.1 Introduction

This study is situated within the field of language variation according to use; that is, of variation between language produced not by different *people* but by the same people in different situations. Someone conversing with friends, for example, produces language which is different from that which they use when addressing senior colleagues at work, or when writing reports, or composing text messages. In each situation, the language is shaped by non-linguistic features often grouped into three categories: topic and purpose of communication; relationships between interlocutors and their roles in the interaction; and the *mode* of communication: spoken, for example, written or electronic. Specifically, this study of text messaging is placed within the tradition of investigating text types along a speech-writing spectrum but, by describing texting as a variety in its own right and in isolation from other electronically-mediated technologies, it also serves as a reminder of the limitations of defining varieties simply according to mode, as evidenced in Biber's (1988) multi-dimensional framework. Such an approach also serves to highlight the linguistic impact of user agency alongside that of technological features. As a consequence of this perspective, and as suggested by North (2006: 212), technological *constraints* can be reconceptualised as *affordances* (Gibson 1986) exploited according to individuals' *effectivities* (Zaff 1995). A description, based on differences in mode, of texting on a spectrum 'somewhere between speech and writing' must account for interactional and user-related factors alongside those of the physical medium.

The insights offered to variation studies through access to large electronic corpora have added a 'new dimension' to the established tradition of language variation study (Hunston 2002: 157). While early work into variation offered descriptive, in-depth linguistic analyses of certain varieties, based on limited numbers of extracts generally taken from written discourse (Crystal and Davy 1969), later studies begin to explore differences between written and spoken varieties and how production, transmission and social function

encourage distinct linguistic features (Chafe 1982; Tannen 1982; Halliday 1985/1989). Corpora make possible more representative sampling of language varieties while facilitating comparative analysis (Conrad and Biber 2001: 4-6), and earlier observations can be confirmed or challenged through quantitative analysis of word and feature distribution across large numbers of texts. This study adds to previous commentaries on texting with investigation of CorTxt, the largest corpus of English language text messages to date, in order to go beyond the use of orthographic features highlighted in the press and explore other patterns and features that typify the corpus.

The aim of the first half of this chapter is to position the study in the field of language variation, showing how an approach which prioritises variation can inform investigation of the language of texting, Txt, and, in the second half, to explore potential insights of corpus-based and corpus-driven approaches to corpus study. In the first half, after defining terminology and outlining areas of existing research, the possible impact on Txt of technological constraints or mode is explored, tracing a development in the literature from a *dichotomous* representation of speech and writing to its reconceptualisation as a *spectrum* or spectra of spoken and written texts which differ according to user- and usage-related factors and which redefine technological constraints as *affordances*. The chapter then looks at the impact of corpora, and raises the question as to whether variation studies can be driven by, rather than based on, data. It concludes by outlining the approach adopted in the present study.

3.2 Spoken and Written Varieties

3.2.1 Language variation

The field into which this study is placed is large and diverse. What brings together studies of language variation into the one field of enquiry is their focus on how linguistic features are shaped by situational features, and reliance on *comparison* in order to identify variety-specific features. Studies differ, however, not only in terms of the level of generality of the variety but also the varying scope of their research (Biber 1995: 10). Studies of language variation across domains involve contrasts and comparisons between varieties (Baron 2000; Carter and Cornbleet 2001; Biber et al 1999). The potential drawback, as Biber (1995)

suggests, is the focus on one parameter of difference, often that of mode[1] (although Biber's criticism is directed towards studies which focus overly on any one parameter) and Biber's (1988) multi-dimensional analysis (MDA)[2] seeks to address this balance, focusing on several dimensions along which varieties differ. In contrast, Biber (1995: 11) points out, in-depth studies of single varieties (as in the current study) are more likely to focus on various linguistic features but again along one parameter of difference, while tending to be synchronic[3] and to focus on specialised or professional domains (Crystal and Davy 1969; Ferguson 1983; Bhatia 1993; Halliday 1993; Yates 1996; Beard 1998, 1999; McLoughlin 2000; Shortis 2000; Gledhill 2000; Groom 2003).[4] They also generally require, as a point of comparison, a notion of what general language is like, against which features departing from this norm are identified.[5]

Different analytical frameworks also emerge. Crystal and Davy's (1969) intention, for example, is to 'establish a theory of textual comparison' (Quirk 1969: v).[6] Other studies apply Halliday's SFL model of field-tenor-mode to analysis of language varieties (Eggins and Slade 1997; Yates 1996). Biber's MDA also provides a comprehensive if unwieldy framework for application (Conrad 2001; Collot and Belmore (1996). The starting-point for this study of texting, however, is the influential speech-writing dichotomy, and spectrum: the use of mode as the defining parameter against which various spoken and written varieties can be compared. The explicit focus on mode in this study recognises that differences between spoken and written language constitute the most widely-explored variable in situationally-defined language variation (Biber 1995: 12), but also seeks to highlight limitations of the tendency to focus overly on the channel of communication at the expense of other variables and, ultimately to argue that texting be explored as a language variety in its own right but one that draws on resources from spoken and written discourse.

[1] Mode here refers to variation in terms of the medium of communication, and does not include other aspects implied by Halliday's definition such as role of language or rhetorical thrust (Butt et al 2000).
[2] MDA is the abbreviation used by Xiao and McEnery (2005).
[3] Although diachronic studies include Halliday's (1988) tracing of the historical development of scientific language; and see Biber (1995: 13-15) for more diachronic studies.
[4] See Biber (1995: 12-13) for further discussion of such studies.
[5] As outlined in section 3.4 below, the advent of language corpora has enhanced such studies, providing representative samples of specialised language domains, reference corpora of general English, and sophisticated tools for exploring patterns typical of the domain.
[6] This is asserted by Quirk in his foreword to Crystal and Davy (1969) and cited by Hendricks (1971: 990).

A final difference across studies concerns terminology. This thesis uses the relatively uncontroversial term *language variety*, adopted by Crystal (2001: 6), rather than *register*, *genre* or *style*, all used variously and sometimes contrastingly with theoretically-specific and 'pre-theoretical' meanings (Biber 1995: 8-10; Lee 2001; Nunan 2008). *Variety* in this thesis refers to:

> texts produced in situations sharing one or more features of production, transmission or function which can be linked to one or more linguistic features occurring across the texts.

Varieties, defined initially through shared situational features, thus include general varieties aligned along one parameter and sharing fewer linguistic features, and specific varieties emerging from the intersection of other parameters and linguistically more similar.[1] This definition acknowledges the impact of other variables alongside mode: texting is by definition defined by its physical channel as a short message service (SMS) sent between mobile phones, yet what makes Txt distinctly recognisable is dictated not only by technological constraints but its exploitation by participants.

3.2.2 Writing–Speech Dichotomy

The dichotomous relationship between situations in which speech and writing are characteristically produced is well-documented (Baron 2000; Biber 1988; Crystal 2001: 26-8; Halliday 1985/1989; Shortis 2001). The list below summarises the main differences identified.

Table 3.1 The speech-writing dichotomy

Speech	**Writing**
Aural	Visual
Informal	Formal
Low social evaluation	High social evaluation
Shared knowledge between participants	Shared knowledge not assumed
Shared time and space	Separated in time and space
Dialogic	Monologic

[1] As pointed out in relation to registers by Biber (1995: 9-10).

Ephemeral	Durable
Real time constraints	Fewer time constraints
Spontaneous	Planned
Unedited	Edited
Interactional purpose	Ideational purpose

Bearing in mind that the dichotomy represents perceptions of writing as academic and speech as comprising everyday conversation,[1] what is important is the significance of the above features in shaping typically-perceived written and spoken language, the main features of which are summarised below (Chafe 1982; Halliday 1985/1989; 1993; Brazil 1995; Baron 2000; Crystal 2001: 26-28).

Table 3.2 Linguistic features of spoken and written language

Speech	**Writing**
Non-fluency features such as hesitations, false starts, self-corrections, repetitions and fillers	Lack of errors or visible self-corrections. Writing is generally more organised and structured than speech.
Clausal complexity and paratactic relations (Halliday 1985/1989; 1993); that is length co-ordinated sentences with a frequent use of *and*.	Lexically complexity or density (Halliday 1985/1989; 1993), with a high number of content words (noun phrases) and hypertactic relations between clauses (that is, use of subordinating conjunctions)
Response to feedback from listener (i.e., meaning may be negotiated) and the use of discourse markers such as 'you know'	Essentially monologic.
A lack of clear sentence boundaries and structures – instead utterances are divided into manageable chunks (by intonation	Sentences, as defined in orthographic terms (demarcated by a capital letter and full stop).

[1] The idea of conversation as a basic form of language against which others can be measured is also an enduring one, found in Crystal and Davy (1969), for example, who posit their investigation of conversation as a 'benchmark' for the analysis of other varieties, as well as in Halliday (1993) who perceives the everyday use of speech to be 'prior', not only in that it comes first in human history and each individual's life, but that it is less removed than other (written) varieties from our direct experience of life; and in Carter and McCarthy's (2006) grammar of spoken and written language, in which 'spoken language' in fact refers specifically, albeit with little acknowledgement, to 'everyday informal spoken texts' (Carter and McCarthy 2006: 9).

and pause)	
Use of deictic expressions	Explicitness: lack of deictic words.
Vague language, as speakers pursue communicative goals 'with respect to a second party' (Brazil 1995: 29).	A different kind of 'purposefulness', as writing is often expression of 'content' rather than maintenance of social relations
High frequency of personal pronouns	The use of the passive construction
A high number of verbs	Nominalisation
Everyday vocabulary	More formal and topic specific vocabulary
Prosodic features such as intonation, stress, loudness	Graphic features such as punctuation, capitalisation and paragraphing

The linguistic features in the table above are described as being shaped or encouraged by different constraints on language production within situations in which speech and writing are characteristically produced, so that writing is detached and integrated; speech involved yet fragmented (Chafe 1982). Halliday (1985/1989), however, by focusing from a systemic functional standpoint on the impact of the different social functions of written and spoken language, begins to highlight the significance of user-related variables alongside those of mode. According to Halliday, the use of writing to store, convey and develop knowledge produces dense, explicit text and encourages nominalisation and other forms of grammatical metaphor; while the clausal intricacies of spoken language lend it the flexibility to be adapted to fulfil various communicative functions.[1] Too narrow a focus on constraints imposed by the mode of communication, this suggests, is insufficient in explaining the linguistic properties of either domain.

The validity of distinctions between spoken and written grammar is seen in its application to language teaching (in, for example, *Natural English* by Gairns and Redman 2003), its increasing recognition in general grammars (Biber et al 1999; Carter and McCarthy 2006) and the teaching of oral English within the British education system (Carter 1999). It also

[1] See Chapter 8 for further development and investigation of the differences between written and spoken grammar.

provides a conceptual starting-point for investigation, of Tannen's (1982) and Chafe's (1982) spectral approach, of registers in Biber's (1988) MDA; and, on the basis that it merges features associated with both speech and writing, of computer-mediated communication (CMC) (Baron 2000; Collot and Belmore 1996; Crystal 2001; Shortis 2001). However, by recognising that, although a useful starting-point, the dichotomy must be reconfigured as a simplistic generalisation (Halliday 1985/1989: 46), we can see that it not only fails to consider the effect of situational features other than mode but neglects the role of technology in blurring distinctions between face-to-face synchronous spoken language and distanced asynchronous written language (Baron 1998a). The impact of technology, such as SMS, on blurring distinctions between speech and writing and the role of other user- and usage-related features are better encapsulated within spectral, or multi-spectral, descriptions.

3.2.3 Spectral and Multi-Dimensional Models of Speech and Writing

The idea that varieties can be placed along a spectrum of spoken and written varieties according to other situational variables, which draws on Malinowski (1922) and disciples Firth (1935) and Halliday, is found in descriptions by usage-oriented linguists of features commonly associated with writing in the 'oral literature' of an Iroquois nation (Chafe 1982), and differences between oral narratives told by Greeks and Americans (Tannen 1982). In both case studies, how distinct communities perceive the situation, relationships between them and the purposes they wish to pursue shapes the language as well as, if not overriding, features of the physical channel. Practice also varies within communities, as it cannot be assumed that particular text types are always presented in a particular modality: that all books are read silently, and plays performed (Baron 1998a: 138). Meanwhile, Baron continues, as individual writers react against learned writing conventions, certain speakers develop strategies for avoiding unwelcome questions and, of current relevance, although email (and texting) is a written mode, participants exploit it for typically spoken purposes. This observation, however, also highlights the limitation of a spectrum which remains primarily organised around one parameter of difference: generally that of mode.[1]

[1] Although Biber's (1994: 9/18) criticism of the spectral approach is on any study which focuses, in his words, on comparing 'registers varying along a single situational parameter, such as formality, attention paid to speech, extent of planning, or physical mode'; that is, along just one spectrum of difference.

The definition of varieties along several parameters, alongside mode, finds an initial basis in work described above. According to Tannen (1982), the continuum of *fixedness*, for example, varies according to social situation and the value granted to fixed expressions across cultures. Tannen found that while Greeks like to fix ways of expressing themselves in particular social situations, Americans avoid 'cliches'. A further continuum of language varieties explored by Tannen reflects speakers' relative focus on *involvement* or *content*: the Greek storytellers mirrored everyday conversation by focusing on themes, making judgments and calling on personal experience, while those told by Americans prioritised the story's message, including substantial amounts of detail and focusing on temporal sequences. The idea that spoken and written varieties can be distinguished in different ways by situational features and thus by clusters of various linguistic characteristics is elaborated on by Biber (1988), whose corpus-based analysis builds on his initial framework of differences between typical writing and speech situations to identify various 'dimensions' along which to characterise language varieties.

The scope of Biber's MDA is impressive, drawing on nearly 500 texts from the written Lancaster-Oslo/Bergen (LOB) and the spoken London-Lund (LL) corpora, and spanning 23 registers and 50 linguistic features (Biber 1988: 66-7). The strengths of his approach lie not only in its recognition that language varieties may not differ primarily in terms of mode, nor in its scope and use of corpora, but in his framework for defining varieties empirically through several co-occurring functional dimensions and linguistic features itself (Hunston 2002: 166; Conrad and Biber 2001: 13-42). His identification of dimensions allows varieties including academic prose, face-to-face conversation, newspaper reporting and fictional prose to be compared along clines such as the extent to which they are involved or informational; abstract; persuasive or non-persuasive; narrative or non-narrative. This is done through interpreting the co-occurrence of frequently-occurring linguistic features in terms of their underlying function (or dimension).[1] The role of frequency is important, in that it is not that features do *not* occur in certain texts but that they are more or less frequent. Furthermore, as stated above, the framework recognises the *parallel* importance of distinctions other than that of mode.

[1] So, for example, the co-occurrence of high frequencies of past-tense verbs, third-person pronouns and adverbs is taken as indicating a narrative function; and corresponding low frequencies, a non-narrative function.

Despite its insights, however, limitations of MDA preclude its use in this study. The selection of both register types and linguistic features from existing corpora and previous research (Chafe and Tannen 1987; Quirk et al 1985) raises questions as to its objectivity for linguists who advocate a corpus-driven approach (Tognini-Bonelli 2001). Furthermore, despite the empirical methodology, conclusions ultimately rest on subjective interpretation of findings as reflecting shared functions or dimensions such as 'Involved vs Informational', not all of which prove easily substantiated (Conrad and Biber 2001: 39). Other criticisms include the model's failure to fully explain links made between situational features of each register and their linguistic character (Collot and Belmore 1996). The validity of the broad registers used by Biber can also be questioned (Hunston 2002: 161). Finally, the scale of Biber's framework and the sophisticated software used means that it is not always possible to replicate, although it has provided a framework for smaller-scale studies into particular varieties (Collot and Belmore 1996; Conrad and Biber 2001). In conclusion, however, although its methodology and underlying assumptions can be challenged,[1] MDA is groundbreaking in theoretical significance and in scope, and its significance to the present study lies in its foregrounding use- and user-related factors alongside mode.

One relevant consequence of the focus on mode in the spectral approach is that technologically-mediated varieties tend to be grouped together regardless of differences in how they are used. The effect of technology on production and dissemination of spoken and written language is frequently accommodated on a speech-writing spectrum (Baron 1998a; Crystal 2001), in studies which conceptualise technology-mediated channels as blurring distinctions between speech and writing.[2] Online communication tends to be seen as moving writing closer to the features more typical of speech outlined above, and the prevailing view is still, to an extent, that electronic communication is a 'hybrid' form comprising features typical of writing and speech (Baron 1998a; Crystal 2001). The problem with this view is the assumption that electronic communication can be described

[1] Its limitations in comparison with corpus-driven investigation are explored in section 3.4 below.

[2] The effects of the printing press on language, for example, are widely documented (Baron 2000; Eisenstein 1979): limitations of space led to a heavy use of abbreviation by early printers, for example, who were also responsible for fixing spelling and helping to prescribe grammatically-defined rules of punctuation. The impact of the telegraph is another striking example of the effect of technology in shaping language, in that the severe limitations of space led to heavy grammatical ellipsis and abbreviated spelling (Baron 1998b, Carey 1989; Crystal 2003). Similarly, speech can also be mediated, through radio, the landline telephone, voice mail and mobile telephony, with corresponding effects on the language caused by factors such as the separation of interlocutors in time and space (Baron 1998b).

as one homogeneous variety and therefore, as suggested above, that mode is the predominant variable in defining variety.

Differences between the constraints and potential functions of electronic modes of communication are increasingly recognised and include varying synchronicity, message length, relationships between interlocutors, their motivations and expectations (Collot and Belmore 1996; Gains 1999; Crystal 2001; Herring 2001; Shortis 2001; Hard af Segersteg 2002; Baron and Ling 2003; Thurlow and Brown 2003). Such differences challenge assertions that electronic communication is collectively different from paper-based or other varieties, and suggest each variety be considered on its own merits or in relation to varieties similar in terms of features other than mode. Gains (1999), for example, finds business and informal emails do not share communicative purpose, and questions whether they differ from paper-based predecessors, with the implication that business emails be considered alongside business letters rather than informal online communication. Collot and Belmore (1996) describe computer bulletin boards as involved, persuasive and linguistically similar to public interviews, due to the audience presence, and highlight the significance of situational variables such as topic[1] (see Moran and Hawisher 1998). Texting emerges from the existing literature as a text-based domain characterised by varying degrees of asynchronicity, mundane purposes, interpersonality, informality, relatively short length and a concentration of non-standard orthographic features (Hard af Segersteg 2002; Thurlow and Brown 2003). The conclusion is that texting be considered a language variety in its own right, regardless of 'formal similarities' (Thurlow and Brown 2003) with electronic communication.[2] Despite its starting-point in comparison with written and spoken texts, this is the view adopted in this study.

The term *affordances* is used throughout this study, as it allows greater consideration of the impact of user agency over new technologies (North 2006), including texting (Thurlow and Brown 2003). Affordances are defined by North (2006: 212) as 'properties of the environment, arising from its material characteristics'. As North (2006) explains, the term comes from psychology and was first used by Gibson (1986) in relation to possibilities

[1] Collot and Belmore apply Biber's MDA (outlined below) to analysis of computer bulletin boards.
[2] What also emerges, of course, is variation within electronic situations: Gains (1999), for example, distinguishing between business and private emails, and the suggestion could be that 'texting' not be considered as a homogenous variety. This is of course the logical conclusion of this line of thought, but not one considered within the constraints of the current thesis.

presented to animals by their environment: trees afford perching' to eagles, for example, but shade to lions. Not only can affordances be contrasting: water affords different things to humans and fish; but they can also be negative (drowning) or positive (drinking). Affordances exist even where participants are unaware of them: sticks always afford tools and fire, even before humans used them for such and for animals which do not (North 2006). In such cases, however, actors' *effectivities* determine whether affordances are taken advantage of. The term effectivities emerges from research into human-machine interaction and can be defined as 'dynamic capabilities of that individual taken with reference to a set of action-relevant properties of the environment' (Zaff 1995: 240). In other words, how technology (or any object) is used depends not only on what it affords, but on how it is interpreted by users, as well as their cultural or individual knowledge or expertise (North 2006: 212). The implication is that technology limits but does not control behaviour (Baron 2002) and users use the same technology in different ways (North 2006: 212).

The suggestion that Txt is shaped by affordances is made by Thurlow and Brown (2003: 13). Texters can determine how synchronous exchanges are and, for example, create speech-like dialogues (Kasesniemi and Rautianen 2002, cited in Thurlow and Brown 2003: 13) or exploit the relative asynchronicity of texting to control and reflect on what they write (Ling and Yttri 2002: 159, cited in Thurlow and Brown 2003: 13). Other 'affordances' are the ability to screen calls, or to turn the sound off to make texting exchanges more private, all of which fulfil needs of intimacy and social intercourse, while the fact that average message length is 65 characters[1] suggests that abbreviation is not primarily motivated by the need for space, but by time pressures and other interactional demands (Thurlow and Brown 2003). Investigation of CorTxt suggests that, rather than being constrained by the physical limitations of the phone, texters are primarily concerned with performing identity through Txt.

3.2.4 Conclusion

The above limitations of a spectral model of spoken and written varieties do not preclude its usefulness, which can be seen in recent grammars (Biber et al 1999; Carter and McCarthy 2006), in conceptualising electronic communication (Baron 1998a), and as the starting-point for Biber (1988) and for the current study. However, Biber's multi-

[1] In Thurlow and Brown's data, at least (see Chapter 2: Research into Text Messaging).

dimensional model warns against an overriding focus on mode as the defining feature by which varieties are compared and contrasted, while the inclusion on the spectrum of technology-mediated communication highlights the view that technology does not dictate linguistic practice but affords possibilities to users who exploit them according to their effectivities and functions they perceive the technology to fulfil. A vivid illustration of this is provided by CorTxt which shows that texting, although highly constrained in terms of physical qualities, is nonetheless exploited in the performance of personal identity and that texters actively draw on their awareness of spoken and written discourse in doing so. This, along with differences in the technologies themselves, also underlines the observation that texting cannot be included alongside electronic communication as one homogenous variety, but must be considered as a variety in its own right, defined by features other than that of mode.

3.3 Corpus-based and Corpus-driven Approaches

3.3.1 Introduction

This linguistic investigation of texting fills a gap left by other studies in drawing on over 11,000 text messages to reach its conclusions. As mentioned above, electronically-stored corpora transformed research into situationally-defined language variation (Hunston 2002: 157), just as they did other fields of linguistic analysis. This study draws on two impacts of corpora: firstly, the size and breadth of the data; and, secondly, the new insights into language descriptions which, although often escaping intuitive analysis, emerge through corpus investigation and challenge traditional grammar models.[1] However, although corpus investigation into language variation exploits the huge amounts of data at corpus linguists' disposal to explore and extend observations based largely on pre-corpus research (Biber 1988), there have been fewer attempts to establish and exploit more inductive approaches using un-annotated data. In other words, most corpus studies of variation have been *corpus-based*, rather than *corpus-driven*. The argument made below is that, despite ground-breaking revelations of corpus-based research and increasingly large-scale, sophisticated analyses now conducted across varieties, the field of language variation could be extended further through the inclusion, alongside this work, of inductive, in-depth, corpus-*driven* investigation of single varieties, exploring new insights into language which looking at

[1] See, for example, Hunston and Francis' (2000) pattern grammar.

large amounts of data reveals. The inductive approaches followed in this study, given the dearth of existing corpus-driven studies of language variation and of corpus analyses of text messaging, are therefore necessarily both exploratory and pragmatic.

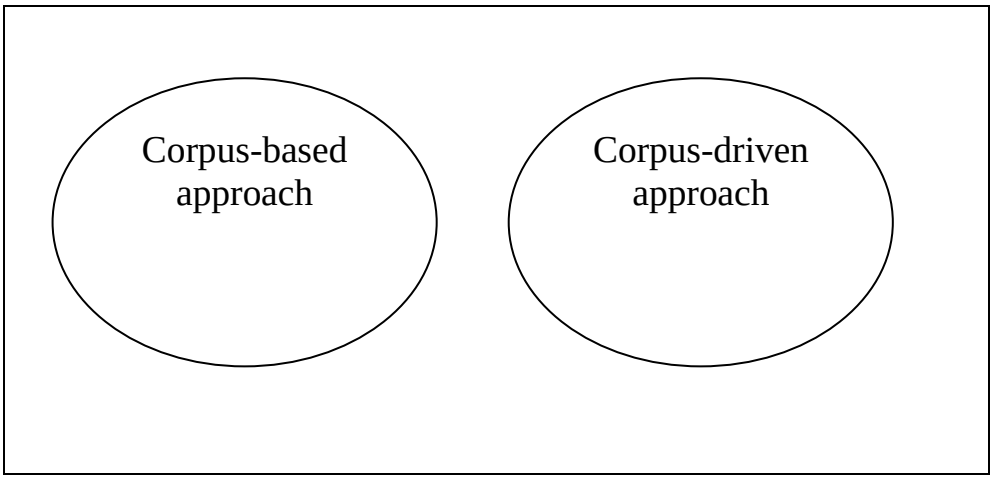
The remainder of this chapter posits the case for in-depth and inductive analyses of single varieties alongside quantitative, comparative corpus-based studies which currently comprise the greater part of corpus variation research. It starts by exploring what corpus-based and corpus-driven approaches involve and questions the accuracy and usefulness of presenting them in a dichotomous relationship (as with speech and writing), suggesting instead that corpus-driven studies be seen as a subset of the corpus-based approach. Although differences between the two can be explained with reference to factors such as researchers' purpose which cut across the approaches, the main difference (and of relevance to the current study) will be seen to lie in their identification and retrieval of linguistic features, with the corpus-driven approach adopting an inductive approach which delays human intervention. The field of language variation, it will be argued, can benefit from inductive, in-depth study of single language varieties, regardless as to whether findings corroborate the lexical grammar model also associated with the corpus-driven approach. Finally, potential frameworks for the inductive investigation of CorTxt are explored.

3.3.2 Defining the approaches

3.3.2.1 Corpus-based/corpus-driven dichotomy

Conceptualisation of corpus-based and corpus-driven approaches as existing within a dichotomous relationship, proposed by Tognini-Bonelli (2001), is sufficiently influential that general discussions of corpus linguistics may devote space to it, even where the aim is to dismiss or bypass it (e.g. McEnery et al 2006: 8-11). The relationship could be represented as below, with the two approaches separate from and in confrontation with the other.

Figure 3.1 Dichotomous relationship between the corpus approaches



Other less widely used terms abound, including corpus-informed (Carter and McCarthy 2006: 11-12), corpus-assisted (Stubbs 1998; Partington 2006), corpus-infused (Bang 2009), which testify to debate over the two main terms above. However, the intention here is not to add to the plethora of terms but, after elaborating on both approaches, to make suggestions as to how the relationship between the corpus-driven and corpus-based could better be construed, what the main distinctions are and, at the end of the chapter, where within all this the current study of texting can be positioned.

3.3.2.2 Corpus-based approach

Corpus-driven and corpus-based approaches are strongly associated with their respective proponents, and emerge from strongly-held beliefs as to the purpose of corpora. Corpus-based research, as mentioned above, uses corpora as a resource from which to provide evidence of pre-corpus descriptions of language or, in Tognini-Bonelli's (2001: 74) words, to 'expound on, or exemplify, existing theories, that is theories which were not necessarily derived with initial reference to a corpus'. In practice, the procedure varies according to the theory being expounded on or exemplified, but generally the approach requires corpora tagged at least for part of speech (POS), as well as information regarding word and grammatical feature frequency, and interpretative analysis of feature co-occurrence. The approach lends itself more readily than its corpus-driven counterpart to application in other fields, so that corpus-based study can be seen as a method or set of techniques rather than a field of study in itself. Research based on corpora, for example, is undertaken by systemic functionalists such as Matthiessen (2006) and those investigating diachronic language

change (Mair et al 2002; Leech et al 2009), while several language variation studies can claim to be corpus-based (Conrad and Biber 2001; Biber et al 1999). As these studies suggest, corpus-based study is a valuable tool which allows researchers to validate and elaborate on theoretical claims. However, it can be argued that by presupposing (and imposing) certain ways of describing language, the corpus-based approach potentially discourages *new* insights. Instead, the scope and nature of findings are to some extent predetermined by preconceptions held by researchers, who can in theory always find what they are looking for.

3.3.2.3 Corpus-driven approach

Corpus-driven statements, according to Tognini-Bonelli (2001: 74),[1] differ from those *based* on corpora in being 'fully consistent with, and reflect[ing] directly, the evidence provided by the corpus' which is 'not adjusted in any way to fit the predefined categories of the analyst'. The underlying argument lies in the observation that traditional grammatical categories, such as POS and clause components, do not always fit naturally-occurring data and that, rather than adjusting data to fit theory,[2] new descriptions of language are required. Corpus-driven linguists attempt to approach data with no explicit preconceptions as to what they will find and allow patterns to emerge from data itself, producing descriptions wholly consistent with corpus data in so much as they are informed wholly *by* it (Tognini-Bonelli 2001). In practice, this means a reliance on frequency in indicating significance, as well as a focus on lexis, recognition of multi-word phrases as the basic unit of lexical description, and adherence to the descriptive power of lexicogrammatical patterning,[3] with insights emerging through word-frequency lists, keywords, concordancing, and clusters (Sinclair 1991; 2004). As this suggests, the corpus-driven approach is associated not only with particular methods of data extraction, themselves grounded in certain beliefs regarding language, but also with a set of observations emerging from corpus-driven study.[4] Corpus-driven methods and observations have led to the development of pattern grammars (Hunston and Francis 2000), language-learning syllabuses based on lexis and phrases

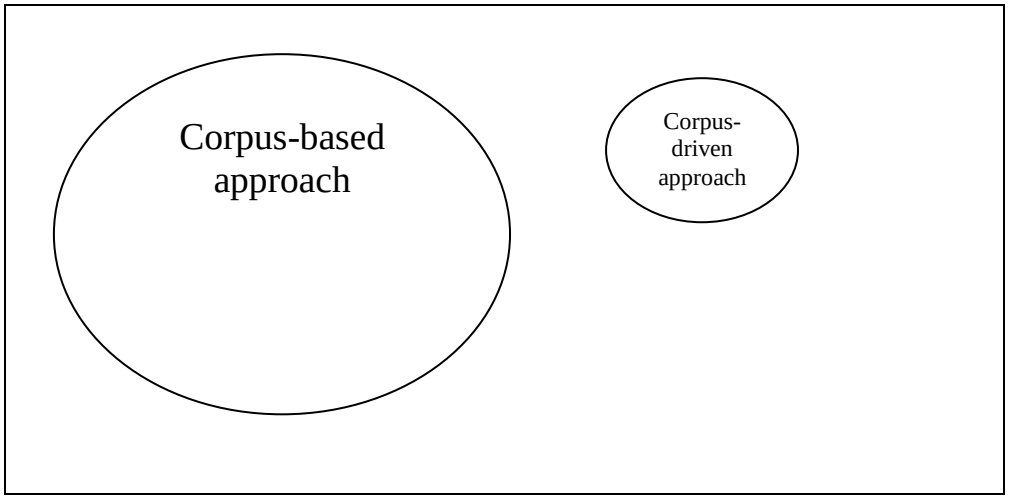
[1] It is worth bearing in mind during the following discussion of her terms that Tognini-Bonelli is herself a proponent of the corpus-driven approach.
[2] As corpus-driven linguists may accuse corpus-based linguists of doing.
[3] That words generate their own grammatical patterns rather than being slotted into grammatical structures.
[4] This distinction between methodology and language model, it will be argued later, can be made explicit through recognition of the *corpus-driven approach* as a theoretical position subsuming an *inductive* methodology (Groom 2006) and an emerging description of language known as the *lexical grammar model* (Sinclair 2000). This distinction is important for current purposes and discussed later.

(Willis 1990; Lewis 1993), and the COBUILD dictionaries (Sinclair et al 1987), as well as work into scientific and academic disciplines (Gledhill 2000; Groom 2006).[1]

3.3.2.4 Reconceptualising the dichotomy

The way in which corpus-driven and corpus-based approaches are reconceptualised in this chapter entails the former positioned as a subset within the wider field of corpus-based investigation. The dichotomous distinction outlined by Tognini-Bonelli (2001), although useful and influential, is challenged here on four counts.[2] Firstly, it overlooks the comparative scale of the approaches, as corpus-based research greatly outweighs corpus-driven work both in terms of the amount of work produced and its variety. A more accurate representation of their relationship involves adjusting their relative scope and size.

Figure 3.2 The relationship adjusted to reflect size and scope of approaches



Secondly, the dichotomy overlooks similarities and shared priorities of the two. Proponents of both share beliefs in attested data when making statements regarding language and agree that this data should take the form of corpora: collections of authentic texts designed for linguistic purposes, planned accordingly, stored and accessed electronically, and analysed non-linearly and quantitatively (as well as qualitatively). Another similarity is that insights granted by both approaches over non-corpus or intuition-based work rest largely and in different ways on *frequency*; the corpus-based approach through investigating feature

[1] As discussed below, the approach followed in these latter studies provide a model for implementing a greater focus on corpus-driven research to complement or balance trends towards larger and more differentiated studies based not only on corpora but on pre-corpus language models.

[2] For similar reasons, some researchers, such as McEnery et al (2006) and Hunston (2002), decide to avoid the terms altogether.

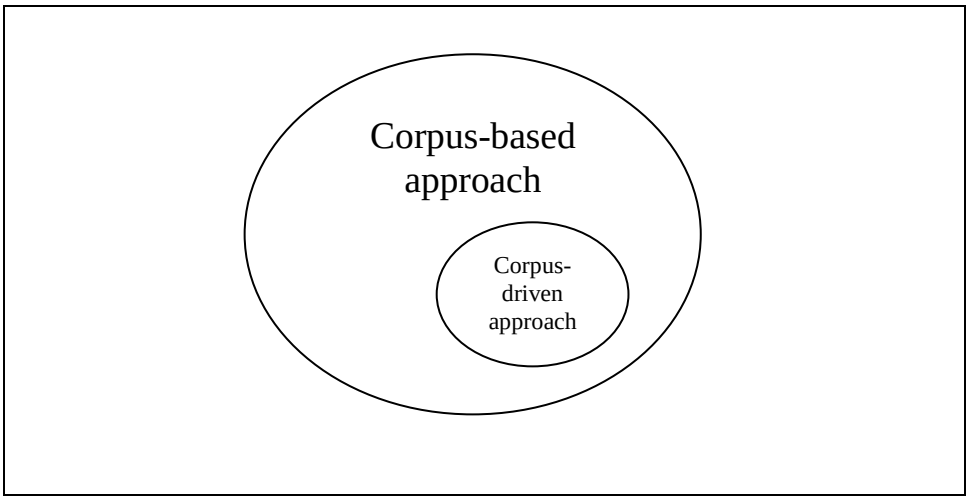
frequency and distribution within and across corpora; and the corpus-driven approach by using frequency (rather than existing theoretical categories) as initial criteria through which to identify significantly-occurring words, patterns, and units of meaning. In overlooking similarities and focusing on differences, the concept of dichotomy only enforces confrontational relationships between corpus researchers.

Thirdly, the boundary between the two is, using McEnery et al's (2006: 11) words, 'fuzzy' or 'overstated' and it is suggested that their relationship is best represented as a cline (McEnery and Gabrielatos 2006: 36). Although proponents of both approaches may hold strong underlying convictions, in reality researchers working within the corpus-based tradition with heavily annotated corpora can adopt inductive approaches, just as 'corpus-driven' studies can depart in various ways from the 'prototypical' approach outlined by Sinclair among others. Rayson (2008), for example, describes his approach as *data-driven* in order to capture the fact that features he studies are selected through consideration of their keyness or salience in the corpus while distancing his work from the corpus-driven perspective within which his use of (key) semantic categories and POS-tags could be questioned. Meanwhile, although Gledhill's (2000) and Groom's (2003; 2006) work is most consistently driven by their corpora, in allowing frequency rather than existing theory or explicit preconceptions regarding language description drive findings, Carter and McCarthy's work into spoken language can be placed further towards the middle of the cline.[1] Their work involves the re-evaluation of descriptive categories and creation of new linguistic terms (for spoken language) within an otherwise traditional grammar model. As discussed below, the distinction within the corpus-driven approach between the lexical grammar model and an inductive approach can be significant in describing differences between such studies.

The alternative way of conceptualising the two approaches shown below captures their similarities and differing scope. Prototypically corpus-driven studies exemplified by Sinclair (2004) could occupy the centre of the corpus-driven circle, and become progressively less prototypically corpus-driven towards its outer ring.

[1] This is not to say that Carter and McCarthy would describe themselves using the term 'corpus-driven' and, in fact, these authors avoid any mention of the dichotomy (Carter and McCarthy 2006; O'Keeffe et al 2007).

Figure 3.3 An alternative conceptualisation of the relationship



The need to avoid a dichotomous model is supported by discussion of two of the 'four basic differences' between the approaches (attitudes towards existing theories and towards intuition, and research focus)[1] which suggests that in some cases differences have not only been overstated (McEnery et al 2006: 8) but can be put down to factors such as purpose, data type, and how general or specialist the corpus which typify, but are not inevitably linked to, one or other approach. In relation to the current focus on language variation studies, the two approaches differ most strikingly when determining which linguistic features should be studied in describing language varieties[2] and how these are identified and retrieved (deductively or inductively). This distinction is significant in the current study for it underlies the argument made for the inclusion, alongside quantitative corpus-based studies, of corpus-driven studies: inductive, in-depth, data-driven analyses of single varieties.

A further advantage of avoiding the dichotomous model, however, lies in the assertion that the two approaches can complement rather than confront each other and that they have much to learn from the other: it is here, Sinclair (1991: 36) suggests, that progress lies. Correspondingly, although the current study is shaped primarily by an intention to let the data drive findings, in exploring what is a new domain in corpus linguistics, it does not confine itself to any presumed distinction between corpus-based and corpus-driven

[1] The other differences include their paradigmatic claims and the suggestion that corpus driven approaches are more 'radical' (largely due to their attitude towards existing theories and their research focus); and types of corpora (raising issues of size and representativeness) used by the two approaches. These issues are less relevant to the current chapter and are explored in Chapter 4: Considerations and Challenges in Corpus Compilation.

[2] Given that, unlike a dictionary, a description cannot take account of all the words in a language variety.

procedures but draws, as we shall see, on both a word-frequency approach and on the application of an existing (inductive) model (Carter and McCarthy 2006).

3.3.3 Attitudes towards annotation and intuition

The decision made in this study not to annotate CorTxt is determined in part by practical considerations. Spelling variation, abbreviation and omitted spacing, as well as the speech-like nature of Txt, complicate attempts at automatic annotation and there is as yet no tagger trained on Txt.[1] However, the decision also recognises that different attitudes towards annotation reflect theoretical preferences which in turn encourage different methodologies: corpus-driven linguists start from wordforms; corpus-based researchers from grammatical categories (Hunston 2002: 67), and this inevitably determines the features selected and retrieved for investigation.

Sinclair's (1991: 21-22) 'clean-text policy' recognises the limitations of annotation (Hardie 2007; McEnery et al 2006: 27; Hunston 2000: 80; Scott 2001: 50; Hunston and Francis 2000: 92-195; Hunston and Sinclair 2000: 76; Sinclair 1991: 81-98), and enables breakthroughs in language description which would not be possible had all linguists followed a corpus-based approach.[2] Rather than starting with corpora tagged for grammatical, semantic or pragmatic categories (Mair et al 2002; Archer et al 2008; Rayson 2008), the software used by corpus-driven linguists takes *wordforms* as the basic unit of analysis,[3] and identifies units for investigation through frequency and statistical significance (Sinclair 1991). Tools used include node-word searches, concordance lines sorted alphabetically to the left or right of the node word to highlight recurring patterns, lists of statistically-significant collocates and, in Wordsmith Tools (Scott 1996), cluster analyses which identify frequently-occurring three-word clusters within concordances. This does not deny interrelation between method and theory: lexicogrammatical patterning described by corpus-driven linguists may emerge as much from tools listed above as it does from data. However, the reliance on wordforms and frequency attempts at least to allow

[1] My initial attempts to train the spellchecker, VARD2 (Baron and Rayson 2008), on CorTxt serve to highlight the differences between the historical texts on which VARD2 was initially trained but suggests a future possibility for the annotation of text messaging: that is, to 'normalise' the spelling. Normalisation is discussed further in Chapter 4.

[2] Descriptions of language emerging from corpus-driven research is described in the following section (section 3.3.4)

[3] Which, defined as characters delimited by spaces, is basic and pre-theoretical. See Chapter 7 for more discussion of wordforms.

findings to directly reflect previously unanalysed data. A distinction is thus assumed between a methodology which, by delaying human intervention through automatic frequency profiles, can be described as *inductive* (Groom 2006), and the *lexical grammar model* emerging from inductive research by, for example, Sinclair (2000).

Criticisms are directed at the need to constantly refine models in the light of current datasets. McEnery et al (2006: 10) identify a paradox in the assertion by corpus-driven linguists that findings are fully consistent with corpus evidence, and yet require further refinement when faced with further evidence, suggesting that their original corpus need be more representative, and that continual category adjustment leads to inconsistency with which analysts and language learners cannot cope. Problems with this charge are, firstly, that few linguists or scientists would presumably agree that the search for accuracy or knowledge refinement be compromised due to the 'burden' it places on those involved. Secondly, dismissing the idea that accurate, final language descriptions inevitably emerge once corpora are sufficiently representative,[1] the refusal to refine 'in the light of corpus data' and the acceptance of descriptions unsupported by data amounts to an argument against corpus data, reflecting charges levelled at corpus linguistics by, among others, Chomsky. McEnery et al surely do not intend to take this stance. The belief that linguists must be open to the continual development of how texts are analysed and language described underlies WordSmith (Scott 2001: 50), for many corpus linguists the main tool in carrying out corpus-driven studies. In Scott's (2001: 50) words, a programme which involves no indexed corpora, tagging, mark up or dictionary knowledge, able to handle standard texts and 'on the fly' processing, 'implies freedom for the user to seek patterns which the designer could never imagine' and he anticipates future changes in how data is viewed. Again, we see that the link between an inductive approach and the lexical grammar model is far from inevitable.

Another criticism is levelled at the notion that a 'truly corpus-driven' approach is one with no recourse to existing language theory and requiring analysts without backgrounds in linguistics (McEnery et al 2006: 10). This somewhat overstates and even misconstrues what corpus-driven linguists do. Firstly, it is not true that corpus-driven linguists make no use of existing categories. Instead, they recognise their value because of their familiarity and usefulness in describing word categories. In pattern grammar, for example, Hunston and

[1] See Chapter 4 for a discussion of representativeness.

Francis (2000) use traditional word-class terminology despite defining classes primarily in terms of their 'patterns of behaviour' rather than form or abstracted meaning, and despite highlighting words which do not fit existing categories (mentioned above) and identifying new word classes.[1] The problem for corpus-driven linguists is where, in Sinclair's (2004: 191) words, computers 'work with the tags and ignore the language' (also cited in Groom 2006: 68). As shown in the following section, corpus-driven investigation of language highlights different roles for traditional categories of lexical and grammatical words.

Secondly, it is ridiculous to assert that linguistic investigation be conducted by people with no knowledge of existing language models or categories. Instead, professional dissatisfaction with traditional categories not seen to match corpus data leads to their considered rejection. In contrast to McEnery et al's claim, for example, Hunston and Francis's eventual decision not to include structural categories (object, complements, adjuncts) within pattern grammar came about after informed attempts to fit categories to their data descriptions, and their realisation that traditional structural categories were unnecessary and insufficient in describing attested language use (Hunston and Francis 2000: 160/176). Corpus-driven linguists are able to let corpora drive their findings *because* of their linguistic expertise and knowledge of existing language models (combined with their approach to corpus data).

In summary, corpus-based and corpus-driven linguists share beliefs in the value of the data and existing categories: in practice, corpus-driven studies may re-interpret or incorporate existing categories, alongside new descriptive labels, provided they are consistent with the data. Consequently, the main difference lies in the timing of their intervention in initially retrieving linguistic features.

3.3.4 Research focus

Differences in the prototypical research focus of the two approaches must be acknowledged but again downplayed: corpus linguists within both approaches recognise the significance of language usage, multi-word units and collocation. Nor is it possible to draw a definite line between corpus-based and corpus-driven language descriptions. Keeping in mind the need to be inductive as much as possible, the current study not only follows one

[1] New word classes include variable nouns (N-VAR), mass nouns (N-MASS); and COLOUR for colours (Hunston and Francis 2000: 182-3).

prototypically corpus-driven approach to data retrieval but, as we shall see, also draws on a existing data-driven frameworks (namely Carter and McCarthy's 2006 grammatical model) in determining effective ways into the new dataset.

Couched in general terms, however, corpus-based linguists use tagged and annotated corpora as a resource from which to provide attested evidence for language theories and models, thereby applying data to investigation of various levels of analysis: lexis, syntax, pragmatics, semantics or discourse; to various fields of study, including translation (Olohan 2004), stylistics (Semino 2004), language teaching (O'Keeffe et al 2007) and forensic linguistics (Woolls and Coulthard 1999); and to grammar theories including systemic functionalism (Matthiewson 2006; Yates 1996). Although the corpus-based approach tends to constrain itself to analysis of POS-tag distribution or grammatical parsing, analysts are increasingly able to tag and explore other aspects of language description, including semantic, discourse and pragmatic features. Using WMatrix to compare key semantic domains of UK 2001 General Election manifestos, Rayson (2008: 541) highlighted, for example, Labour's focus on employment or the priority given by Liberal Democrats to environmental issues. The sociopragmatic annotation scheme devised by Archer and Culpeper (2003) and applied to historical corpora (Archer et al 2008) annotates data with sociolinguistic information such as sex, status and age, as well as interaction roles. As the above suggests, corpus-based linguistics cannot be aligned with one language model but is best described as methodology for exploring language from various usage-based perspectives.

The prototypical corpus-driven approach is, however (as previously seen), associated with a distinct language model. Starting with frequent wordforms, corpus-driven linguists explore how they combine into larger units: the *lexical grammar model* (Sinclair 2001), criticised by McEnery et al (2006: 11) as simplistically conflating syntax, pragmatics, semantics and discourse into a lexical model. It is true that Sinclair's work with the COBUILD team proceeded from a concern with lexis and as such unsurprising that their work foregrounds lexis at the expense of other levels of language description.[1] However, McEnery et al's claim downplays the significance of the 'holistic' corpus-driven approach, and the complex way in which elements combine to form units of meaning.

[1] Except that, according to Sinclair (1991: 2), the initial emergence of new lexical patterns from their database was not anticipated by the researchers, who thought they were using the computer simply as a labour saving device.

Corpus-driven research readjusts the relationship between lexis and grammar [1] and highlights the role of phraseology in meaning-making (Sinclair 2004: 24-48/131-148). By focusing on how closed-class words constrain, and are constrained by, choice of lexical item, Hunston and Francis (2000) show how meaning emerges from co-selection, so that *afraid to* differs from *afraid that*, and how meaning attaches itself to shared patterns: so that words belonging to the pattern ADJ *that* form categories including 'having a reaction to a situation' and 'being certain or uncertain'. Sinclair's (2004) *lexical unit* shows how meaning can build up around collocations, as well as wordforms, and that units of meaning are often highly variable. Illustrated by his well-known treatment of *the naked eye*, his lexical unit contains two obligatory categories, the core and semantic prosody, and three optional, semantic preference, colligation and collocation. [2] The model operationalises Sinclair's idiom principle, which suggests lexical units constitute single choices interpreted as such before readers or hearers switch to the open-choice principle,[3] by showing how users' selection processes are governed by consideration of semantic prosody and shaped by typical semantic, grammatical and lexical co-occurrences. That is, the attitudinal meaning users wish to express shapes choices made regarding other elements of the lexical unit.[4] According to Sinclair (2004: 34), from users' perspectives the categories are thus realised in descending order of *abstraction*: semantic prosody, less easily identifiable, is applied to semantic preference which in turn determines more tangible colligational and collocation patterns.

Although Sinclair's lexical unit is used to investigate spoken data (Cheng 2006) and tourist pamphlets (Lam 2007), one limitation is that linguistically-variable lexical units determined by meaning are not appreciated by computers (Groom 2006: 42) and thus cannot automatically be identified and retrieved: a limitation also of pattern grammar (Hunston and

[1] Although the *wordform* is 'the lowest level of abstraction', the research focus is on the different ways in which wordforms interact with their environment (the cotext) according to various categories of co-occurrence or co-selection which combine word forms into meaningful lexical units.

[2] The core comprises 'invariable' wordforms which constitute 'the evidence of the occurrence of the item as a whole' Sinclair 2004: 141); and *semantic prosody* is the evaluative meaning which is not part of any inherent built-in part of a wordform but is more covertly formed through use of the word in pleasant or unpleasant contexts so that it only becomes apparent through analysis of patterns of use. It determines 'how the rest of the item is to be interpreted functionally' (Sinclair 2004: 141). Of the three optional categories, *semantic preference* is the tendency for wordforms to co-occur with words of certain semantic sets; *colligation*, the frequent co-occurrence of lexical with grammatical words; and *collocation*, a similar co-occurrence between lexical items.

[3] A view shared by those researching the holistic processing of multi-word units (Wray 2002; 2008).

[4] The choice, for example, of *budge* over *move* lies in the fact that 'the user wishes to express or report frustration (or a similar emotion) at the refusal or inability of some obstacle to move, despite pressure being applied' (Sinclair 2004: 145).

Francis 2000). For example, the fact that semantic prosody is not realised in consistent surface forms enabling automatic (or reliable human) retrieval (Sinclair 2004: 144-145) means that the starting-point must be with observable collocational and colligational choices.[1] Nonetheless, in rebuttal to McEnery et al's criticism, Sinclair's lexical item has implications for semantic and pragmatic description. In relation to the current study, however, application of Sinclair's analysis is not so much the priority as his emphasis on delayed human intervention. Other studies which adopt largely inductive approaches and do not necessarily result in the same descriptions (Carter and McCarthy 2006; Groom 2006; O'Keeffe et al 2007)[2] highlights the advantages of, as in the current study, both delaying intervention and exploring various ways into the data. Similarly, the fact that linguists described as corpus-based have also been pivotal in describing multi-word units such as lexical bundles (Biber et al 1999), means that it is therefore difficult to draw a firm line between corpus-driven and corpus-based findings. This is recognised in the current study which, in attempting to adhere as much as possible to determine an inductive approach, draws on both corpus-driven and corpus-based work.

3.3.5 Summary

The above discussion proceeded from the assumption that corpus-based and corpus-driven approaches are fundamentally alike: both believe in corpus data and the significance of frequently-occurring patterns. Another assumption was that studies cannot be divided neatly into one or other camp, but vary according to purpose and data. Corpus-driven linguists' rejection of grammatical categories was described as overstated as many exploit well-established labels, where consistent with data. In fact, it could be argued that although inductive methods tend to result in lexicogrammatical descriptions, this should not be inevitable, while corpus-based work could also result in descriptions which look beyond the word at multi-word units or keyness. The difference, it was argued, lies mainly with how linguistic items are initially retrieved: through tags (EHI) or frequency counts (DHI). The implications include the need to recognise that a corpus-driven approach subsumes an inductive method and a lexical grammar model which tend to be, but are not inevitably,

[1] For example, in strong colligation to the immediate left of the core *true feelings* are possessive adjectives or structures; and further to the left verb phrases such as *communicate*, *give vent to*, and *make public* which show semantic preference on the part of *true feelings* for expression. Still further, there is a semantic prosody of reluctance: *will never reveal, prevents me from expressing, less open about showing*; or of inability: *try to communicate, incapable of experiencing*; *unable to share* (Sinclair 2004: 35-47).

[2] Pedagogic grammars by Carter and McCarthy (2006) and O'Keeffe et al (2007) incorporate findings emerging from data within fairly traditional grammar models.

complementary; and that the corpus-driven approach should be positioned as a subset of corpus-based approaches, including prototypical corpus-driven work at the core but also other studies drawing on inductive methods or lexical grammar. Finally, the above discussion focused on the benefits of inductive investigation. The case is made below for inductive studies of language varieties, such as the current investigation of CorTxt, alongside large, quantitative studies which dominate the field of language variation. The approaches consequently adopted in the current study are also outlined.

3.4 Corpora and language variation

3.4.1 Corpus-based approach: corpus as a tool in variation studies

The main question dividing corpus-based and corpus-driven linguistics, that of how to initially select features for investigation, is of significance to the study of language variation and to the current study, as the features chosen inevitably determine the way in which varieties are described and perceived. The transforming impact of corpus-based research on language variation studies has been possible not only because balanced corpora such as BNC cover many situationally-defined language varieties,[1] but because tagged data is ideal for determining word and grammatical feature distribution and co-occurrence (Hunston 2002: 161-164). Corpus-based investigation is thus able to determine that word frequency is not even across varieties: for example, the twelve overall most frequent verbs, including *say*, *get*, *go*, *know* and *think*, comprise 40% of lexical verbs in conversation and only 20% in academic prose (Biber et al 1999: 373). Grammatical features similarly vary. Negative forms, for example, are more frequent in conversation than in academic prose (Biber et al 1999: 159), perhaps because of the high frequency of verbs, repetition and tag questions, as well as the possibility for disagreement arising from spoken interaction. Furthermore, using WMatrix (Rayson 2008), researchers can determine how key a semantic domain is in different varieties. As these examples show, application of corpus-based tools to language variation allows for large-scale, comparative, quantitative analysis across varieties (Biber 1988; Biber et al 1999).[2]

[1] As a general language database, the BNC can be exploited either for the varieties represented in one or more of its subcorpora or in its role as a reference corpus (Hunston 2002: 15), as a point of comparison for the investigation of other specialised corpora.

[2] The *Longman Grammar of Spoken and Written English* (LGSWE) (Biber et al 1999) is based on a 40-million word corpus (the Longman Spoken and Written English Corpus (LSWE)), the strengths of which lie in the fact that it takes into account variation between four registers: conversation, fiction, news writing, and academic prose. It therefore improves on grammars such as Quirk et al (1985) in its use of authentic examples and its description of conversation as an interactive spoken register. Furthermore, while many of these

Such analyses, in placing emphasis on categories derived from pre-corpus findings, tend to confirm earlier observations regarding the nature of writing and speaking. For example, as Hunston points out, the higher frequency of nouns in news and academic prose was noted by Chafe (1982) and Halliday (1985), as was the greater frequency of pronouns in conversation (Hunston 2002: 162). The drawback is that existing categories preclude unanticipated findings. Collot and Belmore's (1996) study of computer bulletin boards, which adopts linguistic categories used by Biber, omits features which descriptive studies find to be particular to electronic media (Werry 1996; Petrie 1999; Hard af Segersteg 2002).[1] The same can be said of spoken language, Carter and McCarthy's (2006) in-depth investigations giving prominence to features such as headers and tails in speech which are misinterpreted within traditional grammar and corpus-based linguistics alike, where the term 'left-dislocation' implies deviance from the norm.

The corpus-based approach is also less effective in considering how words differ not only in terms of frequency across varieties but their functions across language situations (Gledhill 2000: 33/102; Tribble 2000; Hunston 2002; Baker 2004). Gledhill's (1995) discovery, for example, that numbers function as pronouns to refer to chemicals in research articles has implications for their pronominal system as well as for any corpus-based study which concludes from the absence of traditional pronominal items a lack of cohesion (Gledhill 2000: 34). In this sense, discourse or genre analysis by researchers such as Swales (1990) may prove more insightful than large-scale corpus-based counting. As described below, however, corpus-driven work can also compensate for shortcomings of the corpus-based approach. Gledhill's (2000) and Groom's (2003; 2006) inductive approach, for example, shows how patterns or phraseologies associated even with common function words can differ between article sections. In other words, benefits of large-scale corpus-based work are limited by pre-determined categories which not only hinder emergence of how meaning may be constructed in other ways and across lengthier units of meanings, but risk overlooking distinct ways in which features function across language varieties. Despite the benefits that the corpus-based approach brings to the field of language variation, the

features appear elsewhere, new features highlighted include that of lexical bundles, 'recurring sequence[s] of three or more words' (Biber et al 1999: 990) such as 'going to be a' and 'I said to him'. These differ between conversation and written registers in that the former tend to comprise the beginning of finite clauses and the latter, parts of noun or prepositional phrases.

[1] Petrie's (1999) study of 40,000 tagged emails, for example, isolates features 'rare in handwriting and/or typed communication', including non-standard spelling and email abbreviations.

case is therefore made below for inductive studies which allow features to emerge from the data.

3.4.2 Corpus-driven approach to describing variety

3.4.2.1 What is a corpus-driven approach to language variation?

The argument in this section is therefore not so much that variation studies, including the present one, can benefit from the lexical grammar model but from an inductive approach. This rests on the distinction, posited in this chapter, between the *corpus-driven approach* as a theoretical position subsuming methodology which is *inductive* (Groom 2006) and, on the other hand, as an emerging description of language called *the lexical grammar model* (Sinclair 2000).[1] The argument adopted in this chapter is that, while the lexical grammar model appears consistent with data explored so far, the priority in exploring language variation must be towards an inductive approach which does not seek to impose language models without scrutiny of the variety being described. From a starting-point determined by frequency (and computer-generated keyness calculations) the direction which inductive investigation takes should be determined by the nature of each language variety, including text messaging, and findings which frequency counts produce.

Corpus-driven study of language variation is therefore defined as that in which the linguistic profiles of language varieties are not adjusted to fit existing models, but in which the linguist approaches each variety with no explicit preconceptions as to what will emerge. While not denying that corpus-driven methods determine to an extent what is found, this potentially precludes reference to previous investigation of other varieties, and the imposition of previous observations regarding lexical grammar. Although, for example, the attempt to automate pattern recognition (Mason and Hunston 2004; Mason 2008) would allow frequencies of meanings to be calculated across language varieties, the risk is that imposing pattern grammar across varieties creates a corpus-based approach, albeit one

[1] This distinction was emphasised throughout this chapter, with the term *corpus-driven* describing studies which use (or claim to use) inductive analysis which results in various observations regarding lexical grammar. By acknowledging these two ways of interpreting 'corpus-driven' work, however, we potentially include in our definition studies of the lexical grammar model which, while not inductive, build on observations emerging from inductive work of researchers such as Sinclair (1991; 2004) rather than from pre-corpus language models. Despite the fact that inductive analysis so far of corpus data has followed a lexical grammar model, the distinction also and necessarily allows for inductive work which does *not*. Rayson (2003; 2008), as mentioned above, follows what he calls a data-driven approach in which he identities significant features through frequency counts and keyness calculations although these features include traditional grammatical and semantic categories.

based on corpus-driven findings.[1] Unfortunately, however, the implication of the need to be truly inductive is that it excludes comparative analyses across language varieties, in favour of in-depth explorations of single varieties. Furthermore, if existing studies and linguistic categories are not to form the starting-point, the questions remain as to how to select and retrieve linguistic features for investigation (Gledhill 2000: 83). Possible answers include the keyword approach illustrated by Tribble (2000), and extended in Gledhill (2000) and Groom (2003; 2006) to provide more systematic investigation of phraseologies and collocation characterising disciplines. A final possibility explored here is Carter and McCarthy's (2006) in-depth investigation of spoken structures lying outside traditional grammar models.

3.4.2.2 Keywords and variety

The advantage of keywords (Scott 1996), probably the most oft-cited corpus-driven approach to corpus description (Rayson 2008), is not simply that it can successfully approximate MDA analyses, as Xiao and McEnery (2005) conclude. Nor does the only benefit lie in the ease with which computers calculate frequency in comparison, for example, to Biber's 'time-consuming and computationally/statistically demanding' model (Xiao and McEnery 2005: 68) rendering keyword investigation suitable for experts, other researchers and teachers alike (Tribble 2000: 78). Instead, analyses based on keyness yield distinct insights which aim to be consistent with the data to hand. For example, Xiao and McEnery (2005) found that Wordsmith revealed a high frequency of interjections in the spoken corpus overlooked by MDA because they are not included in Biber's list of features drawn from grammars such as Quirk et al (1985). As mentioned above, similar drawbacks are identified for investigation of electronic language, due to its observed 'novelty' or 'uniqueness' (Werry 1996).

Tribble (2000) illustrates how key wordforms in particular corpora can highlight characteristics of the variety represented: in his case, a corpus of European Union project proposals (the PP corpus), as well as the romantic fiction subcorpus in LOB (RomFict).[2] The list of positive keywords generated for RomFict indicates its speech-like nature,

[1] As well as other practical difficulties involved in identifying patterns over that of, say, lexical bundles or n-grams, and in determining where a co-occurrence is meaningful.

[2] Tribble's approach is explored in McEnery et al (2006: 308-320), in which they apply it to the comparison of conversation and speech in American English.O'Keeffe et al (2007: 208-210) also use keywords to define their corpus of spoken academic and business English.

confirming Biber's findings, while the negative keyness of *the* and *of* highlight the low frequency of nouns and relatively low need to elaborate on meaning. In comparison, keywords in the PP corpus showed the variety to be informationally-dense with substantial elaboration of meaning. Tribble's exploration of the contextual associations of one positive keyword, *experience*, reveals semantic prosody (as he calls it) which contrasts with the wordform's use in general English: *experience* in PP is 'professional capital' (Tribble 2000: 87). This, arguably, would not emerge from Biber's (1988) corpus-based approach.

Limitations to the keyword approach highlight the fact that it is one step in a flexible process involving careful consideration of the data and used in conjunction with other methods (Baker 2004: 357). One relevant drawback is that lexical forms which tend to emerge from keyness calculations do not lend themselves to investigation of clausal relations, for example, or transitivity.[1] The neglect of grammatical distinctions is addressed in this study through application of Carter and McCarthy's (2006) descriptive model of spoken grammar, alongside a frequency-based approach. Another drawback is the arbitrary nature of decisions as to which keywords to explore. Even where attempts are made to be comprehensive, researchers may find themselves ignoring proper nouns, grammatical words, or those not felt to reap interesting findings. Keyword lists also require cut-off points, below which items are not deemed to be significant. As Baker (2004) explains, low cut-off points can result in overwhelming amounts of material while, with too high a threshold, researchers risk being selective and overlooking significant features. These limitations of keyword investigation highlight the need for systematic procedures which extend initial keyword lists in ways that facilitate description of language varieties. One such procedure is outlined by Gledhill (2000) and explored below.[2]

[1] This point relates to McEnery et al's concern, outlined above, that corpus-driven linguists focus on lexis, at the expense of other levels of analysis. Another level which keyword analysis arguably overlooks is semantic meaning. However, semantic annotation (Rayson 2003; 2008), which enables identification of *key categories*, can never be wholly objective (Baker 2004: 353) and as such contradicts corpus-driven principles.

[2] Another, less immediately relevant drawback is that keyword lists preclude frequently-occurring clusters around words not on their own flagged up as key (Baker 2004: 355-6). However, although cluster analysis would further delay human intervention into explorations of 'how words combine, multi-word units have yet to be satisfactorily defined (Wray 2002; 2008) and it is far from clear whether cluster-frequency lists must be manually assessed for *meaningful* units, which inevitably involves subjective decision-making. (Furthermore, initial focus on clusters risks overlooking non-contiguous phrases and other ways in which the same words combine.) This issue is explored in depth in Chapter 7 (and see Wray 2002; 2008).

3.4.2.3 Frequency-driven corpus study of phraseologies

The inductive approach devised by Gledhill (2000) in investigation of cancer research articles is applied by Groom (2003; 2006) to US stem-cell patent specifications and to exploration of links between epistemology and phraseology in two disciplines, history and literary criticism. To briefly summarise their methodology, investigation begins with the generation of keywords not only in comparison with a reference corpus but (in Groom 2006) between the two disciplines and (in Gledhill 2000 and Groom 2003) between rhetorical sections.[1] For this purpose, corpora are divided into subcorpora containing, in Gledhill (2000), title, abstract, introduction, methods, results, and discussion. The ten most frequent key grammatical or closed-class items form the basis of investigation, and the phraseologies into which these closed-class items enter explored using concordance lines. In Groom (2006) this is extended to make claims about the link between phraseologies and each discipline's epistemological stance.

Gledhill's and Groom's defence of their methodology as corpus-driven is essentially that it is inductive, in that wordform frequency and keyness direct the course of investigation, guided by previous corpus-driven findings concerning the significance of phraseology (Sinclair 1991; 2004) and closed-class words which contribute to the meaning of phrases as much as the lexical items which control them (Hunston and Francis 2000). Phraseologies which emerge are not those evident within traditional approaches.

Figure 3.4 Semantic sequence identified by Groom (2006: 86) as part of his model investigation

process	**of**	**object**
the building	**of**	a church
the attribution	**of**	insensitivity
the rescue	**of**	Jews
A change	**of**	mind
the elaboration	**of**	native histories

As with corpus-based studies, there is still a degree of selection in determining features to study. The assumption made is not only that keywords play a role in significantly frequent

[1] In other words, words which are significantly more frequent in one section compared to the rest of the corpus thus form the basis of much of their analysis, regardless of their significance in comparison to general language use.

phraseological patterns (Gledhill 2000: 102), but that these phraseologies are of greater significance than other features traditionally explored in academic or scientific texts such as nominalisation or modality, discussion of which is precluded by focusing on phraseologies. However, the research draws attention to the existence of phraseological patterns which vary according to language functions and so challenge corpus-based findings.

Far from providing a framework for application to similar phraseologies across varieties, these corpus-driven studies highlight the fact that what serves to describe one kind of language may be inappropriate for another. Groom (2006), for example, is concerned with describing the way in which knowledge is perceived and constructed linguistically in literary criticism and history.[1] In other words, he focuses on links between phraseology and epistemology: 'epistemology is manifested in phraseology' while 'phraseology produces and reproduces epistemology' (Groom 2006: 25). In fact, he suggests, each distinct disciplinary discourse emerges from fusing its epistemological stance to knowledge[2] and to the 'conventionalised forms of expression', or phraseologies, used to capture them (Groom 2006: 24). Although this epistemological perspective could prove insightful in investigating other academic or science texts, it would hardly be appropriate in application to more generalised varieties whose users share no epistemological perspective: such as texting.

Practical drawbacks include the potential unwieldiness of Gledhill and Groom's approach with larger datasets or less specialised varieties. The corpora used in both studies are small, around 500,000 words (150 articles) and 5,500,000 (405 patents) respectively.[3] As Groom points out, investigation of the larger section in his patent specifications, 'Detailed descriptions of the invention', is difficult to summarise within the limits of his thesis. Furthermore, academic writing is 'specialised' in the sense that participation is limited to individuals who learn and accept the discourse community's meanings, values and goals, as well as the language and specialised terminology used to express them (Groom 2006: 24). The formulaic nature of sections such as 'Titles' suggests that less specialised language may not be amenable to this approach.

[1] Rather than how language is used within different genres.
[2] For example, whether the discipline is soft or hard, pure or applied (Becher e.g. 1987).
[3] Incidentally and as briefly alluded to in an earlier section, both can claim, due to the limited generality of the varieties studied, a degree of representativeness: the texts in Gledhill's (2000: 51-53/90-98) corpus reflect the discourse community at one Pharmaceutical Science Department (PSD), while Groom's texts comprise *all* US stem-cell patents issued at the time (Groom 2003: 23).

The extent, therefore, to which this model can inform other corpus-driven studies of language varieties is by definition limited. The danger is that a corpus-driven approach quickly becomes corpus-based, in the sense that observations driven by one dataset are applied (perhaps inappropriately) to another, and subsequent datasets adjusted to enable comparisons. To ensure that future investigations are truly corpus-driven, researchers can similarly start with frequency and computer-generated keyness calculations, but the direction their investigation takes should be determined by the frequency and significance of features in their data. For example, Gledhill's and Groom's focus on closed-class rather than open-class words is useful in warning against dismissing the former, but should not preclude investigation of open-class words; similarly, the focus on positive keywords should not preclude consideration of negative keywords. Gledhill's and Groom's work is important and the principles they uphold guide the current investigation of texting, although the exact form the investigation takes necessarily differs. The implication, however, is that corpus-driven study can never be truly comparative nor generate models for application across varieties.

The final limitation to the work described above is that, by focusing on lexical patterns in which significantly frequent function words operate, the researchers risk missing other features or 'levels' of language description. These features include discourse markers, as well as features without searchable surface forms, such as ellipsis. The work of Carter and McCarthy (2006), inductive but based on in-depth reading, potentially fills gaps left by the corpus-driven approach.

3.4.2.4 Text-driven approach to spoken language

Work by Carter and McCarthy (2006) using CANCODE is corpus-driven in the sense that it reveals features of spoken language misinterpreted or neglected by traditional models of grammar. As Carter (2004a) explains, some reveal marked differences in the way certain grammatical features (such as discourse markers and modal expressions) are used, which prove problematic for existing word-class categories and description of grammatical functions, while *headers* and *tails* and the positioning of adverbials pose problems for existing descriptions of clause structure. The way in which clauses are connected differs in spoken language from descriptions assumed in traditional grammars, while situational ellipsis challenges the relationship between ellipted and 'full' sentences (Carter 2004a). However, although Carter and McCarthy's findings challenge and extend pre-corpus

perceptions, their concern is not to posit a new grammar in the way that Brazil (1995) does. Their approach is best described as incorporating data-driven observations into existing descriptions, to raise awareness of spoken features for language teaching and for challenging ideas of correctness. In their words, the *Cambridge Grammar of English* is 'informed' rather than driven by corpus data, and by pedagogical considerations which limit the use of authentic language (Carter and McCarthy 2006: 11-12).

Their methodology, furthermore, could be described as *text*- rather than *corpus*-driven, to use Bednarek's (2006) term. It is driven by corpora in that existing grammatical categories are not imposed on data through tagging, but instead words (if not wordforms) and phrases form the basis of corpus investigation techniques such as frequency (McCarthy and Carter 2006: 12-16). Frequency is an important feature for inclusion in *CGE*, as is recognition of 'the patterned relationship between vocabulary and grammar' which emerges from CANCODE.[1] However, many observed features are structural differences not realised in consistent, searchable wordforms which can be identified in word frequency counts and concordances: *headers* and *tails* for example have no easy search-term other than, perhaps, pronouns; situational ellipsis involves searching for something not there; while other observations involve examination of features across turns rather than concordances. As the extract of conversation which opens Carter's (2004a: 28) investigation suggests, these features can only be identified through thorough, qualitative investigation of stretches of text. For the purposes of the present discussion, this limits description of Carter and McCarthy's approach as *corpus*-driven. However, as mentioned above, there are also implications for corpus-driven work and the features that it may overlook,[2] and the need for qualitative, descriptive accounts alongside frequency-driven approaches. In this investigation of texting, Carter and McCarthy's (2006) description of spoken grammar will therefore be applied, alongside other *corpus*-driven, frequency-based approaches, to CorTxt.

3.4.2.5 Summary of inductive approaches

The relevance of the approaches outlined above is firstly that they propose models, for use in the current study, by which linguistic features or patterns can be retrieved from un-annotated data; and secondly that findings challenge corpus-based approaches whose pre-

[1] See, also, the chapter on 'Grammar and Lexis and Patterns 'in O'Keeffe et al (2007: 100-119) which recognises the relationship between grammar and lexis.
[2] As well as having similar and obvious implications for the corpus-based approach.

determined categories preclude investigation of these features. It is not so much that the studies above reject traditional categories per se, but they do not form a filter through which language features are selected and studied: instead, features emerge directly through consideration of their frequency in the current dataset. These studies also highlight, however, that care must be taken in attempts to use them (as in the current study) as a framework to guide investigation of vastly different varieties. As with other corpus-driven findings, although they inform future work into language variation, the need for inductive approaches which stay true to the data to hand is paramount.

3.4.3 Summary and conclusion

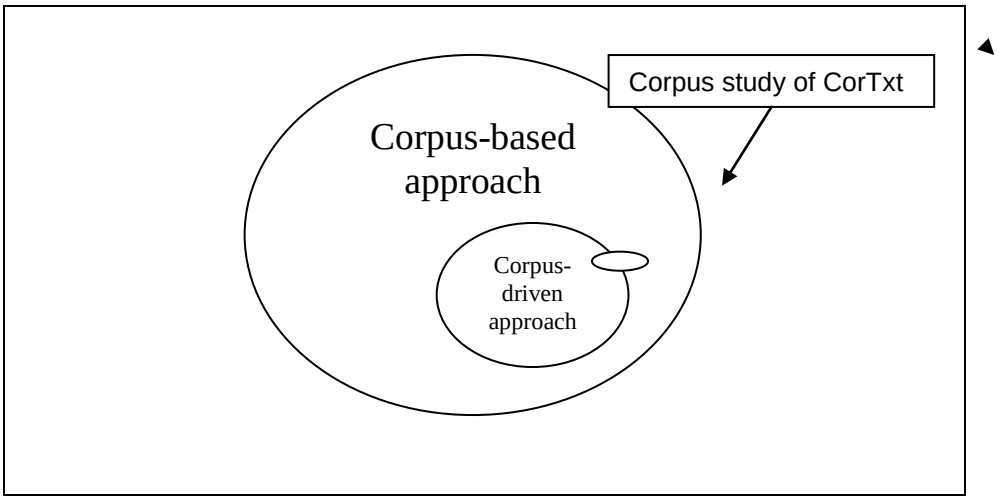
Whether language variation studies are intended as corpus-based or corpus-driven to some extent depends on the adherence of researchers to one or other approach. For researchers seeking to work with existing grammatical categories and wary of prominence granted to lexis by corpus-driven linguists, corpus-based approaches provide more evidence for differences in feature distribution across language varieties than that accessed by pre-corpus analysts and in this way corpora can substantiate claims. For those who feel pre-corpus language descriptions obscure other ways of describing language, the danger that corpus-based approaches pose is that combining pre-corpus categories with attested data can confirm inaccurate or misleading interpretations of data with as much authority as accurate or relevant ones. For example, as suggested above, not only can calculation of feature frequency preclude differences in how features function in different varieties and the distinct meanings conveyed, but it overlooks investigation of patterns in which words occur and the extent to which these differ across language situations.

Alternatively, however, the two approaches can be reconceptualised in terms of their similarities, with the distinction between them recognised as that of initial retrieval: that is, whether the approach is inductive or deductive. Firstly, recognition that the corpus-driven approach is distinguished primarily by its identification of significant features through their frequency and significance in the data being explored rather than pre-corpus constructs allows us to focus on the need to be inductive in studies of language variation and, more specifically, in the current study. Such inductive investigation is by definition limited to single varieties, precluding (re-)application of models of description and therefore excluding automatic, quantitative, comparative analysis across language varieties. Far from

upholding this as a limitation of corpus-driven research, the conclusion drawn here is that the field of language variation would benefit from more in-depth inductive studies of single varieties.

Secondly, however, the reconceptualisation avoids some of the constraints of an (as we have seen) unjustified dichotomy in selecting appropriate methods of investigation. Given the novelty of text messaging data in corpus linguistics and the need to be exploratory, this allows exploration not only of a word-frequency or keyness approach, but also the application of other (inductive) models (such as Carter and McCarthy's 2006 text-driven framework of spoken grammar) to the data. As such, the current study can be positioned at the intersection of studies described as corpus-driven and corpus-based in our diagram.

Figure 3.5 Positioning the current corpus study of CorTxt



3.5 Chapter Summary

As well as highlighting complexities inherent both in existing conceptualisations of written and spoken varieties and approaches to corpus analysis, this chapter highlights, in various ways, the need to ensure that texting is described consistently with the data rather than in comparison to other varieties or datasets. The first half outlined drawbacks of defining varieties solely according to the extent to which they resemble speech or writing, and highlighted the need to consider other factors such as users' agency in exploiting technological affordances to fulfil interpersonal functions. In short, texting must be regarded as a variety in its own right, albeit one drawing on resources of written and spoken texts: and this emerges from investigation of CorTxt. The second half discussed two

approaches to corpus investigation and argued that language description must be consistent with the data: the approach must be inductive. When applied to the study of language variation, although existing findings based on inductive study can be used to guide further investigation, a focus on in-depth, inductive studies of single varieties (such as text messaging) based on wordform frequency is required. At the same time, given the relative novelty of the current dataset, the approach presently required must be pragmatic and exploratory. In the following chapters, CorTxt is explored not only through a corpus-driven approach starting with frequently-occurring wordforms and a corpus-based application of Carter and McCarthy's description of spoken grammar, but also other approaches (namely, investigation of spoken creativity) emerging from the data.

CHAPTER FOUR
CORPUS COMPILATION: CHALLENGES AND CONSIDERATIONS

4.1 Introduction

The purpose of this chapter is to outline and explain considerations and challenges behind the creation of the CorTxt corpus. Designing and compiling a text message corpus for linguistic analysis raises considerations addressed by compilers of other corpora: corpus size and representativeness, sampling, and ethical considerations. However, the distinct nature of text messaging also posits new challenges, including the need to select appropriate data-collection methods, the ethics of collecting electronic data, and problems of *message selection* (whereby participants are selective in messages they submit for the corpus) as well as *transcription error* (where transcription mistakes are made by participants). Such issues must be considered not only with reference to the general literature on corpus design (Sinclair 1991; Hunston 2002; McEnery et al 2006) but to previous attempts to compile text message corpora (Hard af Segersteg 2002; Kasesniemi and Rautiainen 2002; Fairon and Paumier 2006). Discussion of each consideration emphasises the need to be pragmatic when faced with obstacles to obtaining and handling data and the importance of documenting decisions made.

The chapter starts by outlining current corpus specifications, and describing data collection procedures followed. It then addresses issues raised by, and explains procedures followed in, the compilation, anonymisation, storage and exploitation of CorTxt. Methods and procedures are justified throughout with reference to previous research and the need to be pragmatic in obtaining data. The chapter, in framing the current study, thus also provides a framework for consultation by future researchers of texted data.

4.2 Corpus Specifications

4.2.1 Data

Specifications for CorTxt are tabulated on the following page.

Table 4.1 Corpus specifications

	CorTxt
No of messages	11,067
No of words	190,516
Average no of words per message	17.2 words
Collection period	March 2004 – May 2007 (3 years 2 months)
Collection methods and procedures	From friends and family (10,626 messages) Recruited informally through by email (Appendix 4.1),[1] in which the purpose of 'examining how language is used' was explained and some areas of interest named, including linguistic creativity. Participants[2] requested to submit all received and sent[3] messages with dates, times, and senders' names or initials[4]. Number submitted was up to participants. It was made clear messages would be anonymised, and senders required to complete consent forms and personal profiles (Appendix 4.2). Finally, interest in 'determining how people text on a day-to-day basis' was highlighted and participants asked not to select interesting or amusing texts[5]. 'Templates' were attached as Excel files into which participants could enter messages.
	AOL anonymous online public forum (441) A further 441 messages came from an anonymous public forum provided by AOL for forwarding text messages. After it was discontinued, people kept using it: that is, messages used in this study were those which were sent to the service but not forwarded.

[1] Although most were people with whom I also had face-to-face or telephone communication and some had previously heard from me about my research.
[2] This description of recruitment procedures distinguishes, for reasons of convenience, between *participants* (those I approached and who directly contributed messages from their phone) and *senders* (all those who sent messages which were contributed, including participants and people who sent messages to the participants who then contributed them).
[3] Some phones do not automatically save sent messages.
[4] Or some anonymous way of distinguishing between senders.
[5] Although it was made clear that participants could omit private messages which they (or their senders) did not feel comfortable contributing.

Composition of texters	Mainly British English speakers, aged 19 – 68, professionals and students
Language of communication	English (mainly British English)
Type of communication	Mainly personal communication, although some business text messages evident, including: Hi Are you interested in working taster day at Cornwall College, Camborne 27.4.06 4hrs@£6 per hour? Ring if interested. NAME173

4.2.2 Bibliographic composition

Although some senders were not identified or distinguished between, a total of 9023 messages were sent by known senders.[1] Their composition is tabulated below. It should be noted that the largest contributor supplied only 14.1% of the total number and many contributed only one message. There is no reason why unidentified senders would not also fit the bibliographic profile below, but no conclusions are drawn.

Table 4.2 Bibliographic composition

	CorTxt
Total number of participants (those directly recruited who submitted messages)	16
Total number of named senders (including participants plus those whose messages were sent to and submitted by participants)	235 (sending 9023 messages)
Age range	- 21 3% 22-25 25% 26-29 23% 30-35 30% 36-39 5% 40-45 3% 46-49 4% 50-55 3% 55-59 6% 60+ 3%
Gender	M 28% F 62%

[1] 2044 messages were from unknown senders, either from the public AOL site (401 messages) or sent to participants who could neither remember who had sent messages nor recognise the content or style as a particular interlocutor (see section 4.4 for the ethics of this)

Profession	Education	48%	Students, teachers, student support and information management
	Health and community care	9%	Social services management, medicine, nursery officer, counselling, speech and language therapy
	Engineering and construction	7%	Construction management, architectural technology, technology
	Office and administrative	6%	Office assistance and administration
	Arts, craft and entertainment	4%	Publishing, photography, design
	Other (no one category over 3%) Hospitality IT Legal and Financial Management and Executive Manual work Public Service Retired Unemployed	26%	
Native language	English	96%	
Use of predictive texting[1]	Yes No	83% 17%	
Average length of time texters have used text messaging	Under 5 years 5-7 years Over 7 years	28% 58% 14%	
Frequency of texting	Infrequent user (one or fewer messages a day) 35% Daily user (around one or two messages a day) 20% Frequent user (more than one or two a day) 45%		

[1] *Predictive texting* is a device now found in most if not all phones which allows the phone to predict the word you are typing, generally based on an in-built dictionary which users can add to. The relevance of predictive texting for the language used in text messaging is not only that it has been shown to make typing faster (How 2004), but it affects how words are abbreviated. *Tomorrow*, for example, appears variously in CorTxt as *tomora*, *tomoz*, *tomo*, and *tmw*, and it seems predictive texting would favour the shortening *tomo*, as phones would recognise the word being typed and produce *tomo* after just four key presses.

Although ages range from 19 to 68, most senders are 22-35,[1] a little older than other studies. The sample was predominantly female (62%), and first language overwhelmingly English (presumed, in most cases, to be British English). Jobs varied, but included many in education, health care, engineering and office administration. Most senders used predictive texting and had been texting for 5 years or over, although the frequency with which they texted varied. The general profile comprises young, mainly female, often professional British adults, accustomed to interacting by text.

4.3 Considerations behind Corpus Design

4.3.1 Data collection

The question of how to collect text messages parallels that faced by compilers of written or spoken corpora, who must similarly identify the nature of data required, locate it, and access it (Hunston 2002: 26-27),[2] and their methods similarly differ according to the nature of the data, research questions and practicalities. The decision made in this study to recruit family and friends emerges from evaluation of drawbacks and benefits of other more quantitative and qualitative approaches used in previous texting studies, in the light of current aims and the need to be pragmatic to collect sufficient data.

Data collection methods used in previous studies of text messaging are tabulated below. Broadly categorised, as well as the recruitment of family and friends (Hard af Segersteg 2002), they include qualitative approaches which probe the texting practices of a few individuals (Grinter and Eldridge 2002; Kasesniemi and Rautianen 2002) and quantitative, often anonymous and web-based, methods (Fairon and Paumier 2005; How and Kan 2005; University of Leicester 2006).

[1] At the time of completing the personal profile.
[2] The issue of defining the target population and sampling from it are also looked at in section 4.3.3 below.

Table 4.3 Data collection procedures used across the text messaging literature

Study	Corpus specifications	Data collection methods	Research focus
Grinter and Eldridge 2001; 2003	477 messages 5 participants aged 15-16.	Record 'logs', in which participants recorded messages sent and received during 7 consecutive days, followed up by video-taped group discussions and questionnaires.	Communicative functions, analysis of conversation 'threads' and abbreviations
Hard af Segersteg 2002	1152 messages 17,024 words	Various: 112 from an anonymous webpage, 252 messages forwarded from volunteers (two males and two females aged 12-25) and 788 from family and friends	Abbreviation and structural ellipsis, to determine how written communication adapts to technology.
Kasesniemi and Rautianen 2002	Nearly 8000 messages from nearly 1000 teenagers, accompanied by interview, field notes, and observation journals (completed by participants)	Teenagers recruited through various channels, including television, the Internet, teachers at schools and the 'snowball' technique. Messages submitted by teenagers, with 'cover letters', along with the observation journals.	Social-scientific study into the communicative practices of mobile technology.
Oksman and Turtianen 2004			Exploration of texted interaction through a symbolic interactional approach to determine whether new forms of social interaction are being produced.
Thurlow 2003	544 messages 135 British university students, aged around 19	5 text messages were transcribed and submitted by students at end of university lecture, during which they were recruited.	Linguistic study, of abbreviations, communicative functions and message length
How 2004; How and Kan 2005	10,000 messages	3,348 messages from a 'website collection programme' with the 146 undergraduate participants financially rewarded for their contributions 6,167 messages from a 'small pool' of 20 participants aged 18-22; and 602 from a Yahoo SMS chat website.	Studies aimed at improving predicted text entry.

Fairon and Paumier 2006	30,000 messages (from an initial 75,000) 3,200 participants aged 12 to 65.	Project was broadcast nationally (October and December 2004) and participants requested to send texts to a free mobile number. (sociolinguistic information including 'ability to decrypt SMS' and 'writing habits'	Compiling and preprocessing the corpus to serve as a reference corpus, and translating the language into 'standard' French for future study.
University of Leicester 2006 Main researcher: Dr Tim Grant, now at Centre for Forensic Linguistics, Aston	[final corpus specifications pending, Grant 2009 pers comm.]	Participants in the study were asked to submit 10 messages to a website and the researchers hoped to recruit at least one hundred texters Sampling technique snowballing used, whereby participants were encouraged to urge friends to contribute.	Forensic linguistic research which aims to analyse 'linguistic consistency and variation in individuals' texting style; and also 'the influence of peer groups upon writing style and texting language'.
Ling and Baron 2007	191 text messages (1473 words) collected.	Paper diaries distributed to university students at an American university.	Linguistic comparison with IM communication

Advantages emerge, from the above studies, of recruiting family and friends over quantitative methods often involving anonymous participants. These include the greater likelihood of ensuring authenticity (Hard af Segersteg 2002: 209-210; How and Kan 2005; Fairon and Paumier 2006; University of Leicester 2006),[1] greater familiarity with participants' backgrounds, ability to acquire personal information (Fairon and Paumier 2006)[2] and the fact that recruiting friends and family achieves *depth* (allowing greater understanding of individuals' behaviour) as well as *breadth*. Advantages over more in-depth studies are time and cost (Kasesniemi and Rautianen 2002),[3] as well as a greater focus on breadth due to the enhanced possibility of recruiting willing participants (Hard af Segersteg 2002).

Choice of method, however, also depends upon the current research focus. The advantage of the snowballing technique,[4] for example, is that it enables researchers to explore not only how individuals vary in linguistic practices, but how *networks* of texters differ: ultimately the goal of the University of Leicester (2006). Others that aim to describe linguistic features typical of texting require breadth and variety of textual data rather than (or as well as) in-depth knowledge of participants' backgrounds, and so adopt quantitative procedures for at least part of their collection process (Fairon and Paumier 2006; Hard af Segersteg 2002). Similarly, How and Kan's (2005) purpose is to increase typing speeds through rearrangements of the phone keyboard layout, for which they need textual rather than sociological data. In contrast, social-scientific researchers such as Kasesniemi and Rautiainen (2002: 171) who aim to document 'text messaging as it relates to the life of teenagers' require text messages to support insights gleaned through qualitative fieldwork, rather than a wide range of participants. Recruiting family and friends enables a substantial number of participants and text messages, whilst facilitating the collection of some

[1] The 75,000 messages received by Fairon and Paumier (2006) were filtered down to 30,000 in part by removing messages written for their team's attention. Hard af Segersteg (2002: 209) notes that some messages 'seemed to actually be filled in by someone … making things up for the fun of it' and that these were 'ignored'. How and Kan (2005) also report removing 'non-genuine' messages, although they do not explain how they were identified.

[2] Of Fairon and Paumier's (2006) 3200 participants, only 2500 returned sociolinguistic data, while other studies did not request it (How and Kan 2005; Hard af Segersteg 2002).

[3] Even in a relatively small-scale study such as Hard af Segersteg's (2002), the four participants were paid for messages submitted, by having their pay-as-you-go cards topped up (p210).

[4] The snowballing technique adopted by the University of Leicester involves initial participants being urged to invite friends to join the study, who then invite their friends and so on. They were also asked to devise ID codes for themselves and the recipients of the submitted texts, presumably to allow them to identify whether their own recipients later join the study as a participant in their own right.

bibliographic information, and is appropriate for a study which prioritises textual data rather than sociological information.

The decision in the current study to supplement messages from friends and family with 441 messages from an AOL website reflects those made by other researchers intent on ensuring breadth and depth. Also evident in these decisions, as with all data collection decisions, is the need to be pragmatic to obtain sufficient data or, in Hunston's (2002: 26) words, 'make use of as much data as is available', not only because of time or cost and practical difficulties involved in transferring text messages into computer-readable formats, but people's reluctance to part with messages (Hard af Segersteg 2002: 207) due to their well-documented private nature (Kasesniemi and Rautianen 2002: 181-2; Harkin 2003; Taylor and Harper 2003). Attempts to overcome this include obtaining data from various available sources, including online text message services (Hard af Segersteg 2002; How and Kan 2005). The drawbacks to using online text messages, such as those from the AOL website, are that message content and language may reflect the fact that they were sent through a third-party service, that it is not possible to know who sent the messages, nor to ensure that messages constitute genuine communication, and impossible to gain consent from senders. At the same time, in this case, there is no reason to believe messages were made up, whilst the latter objection (potentially more serious) is offset by the fact that the AOL service was never particularly private.[1] Accordingly, for pragmatic reasons and to increase breadth of participants, the online text messages were included.

4.3.2 Corpus Size

Corpora such as CorTxt comprising less than 200,000 words are relatively small. However, although the size of CorTxt is limited by practical constraints such as obtaining data, the corpus is sufficiently large for present purposes. In fact, the size of any corpus depends on these two factors: practical considerations and research focus (McEnery et al 2006: 72-3). *Specialised* corpora, for example, can be smaller than general *reference* corpora (Hunston 2002: 15), for they aim to represent smaller subsections of language than general corpora. There is growing recognition of the value of smaller, specialised corpora alongside large reference corpora (Ghadessy et al 2001; O'Keeffe 2007), for practical reasons as well as a shift towards analysis of particular language domains. As a specialised corpus, CorTxt need

[1] The ethical considerations behind this issue are covered below in section 4.4.

not compare to the BNC or CANCODE, which covers several spoken varieties. Corpus size is similarly dependent, if not on the frequency of particular features studied,[1] then on researchers' purpose in studying them (Sinclair 2001; Scott 2001): language textbook compilers anxious that learners frequently encounter new vocabulary require large corpora, while studies of authors' stylistic use of obscure or creative hapaxes (e.g. Marcinkeviciene 1998) need only smaller corpora (Scott 2001; Sinclair 2001).[2] Finally, size is never absolute, given growing dimensions and changing perceptions as to what constitutes large and small corpora (Sinclair 2001: ix), and the fact that size can be determined by number of texts rather than words: an *isotextual* not *isolexical* approach, to use Oakey's (2009) terms; and perception of the size of CorTxt shifts if calculated according to the number of text messages (11,067) rather than the words.[3] In short, the dimensions of CorTxt are not considered problematic: as Sinclair (1991: 18) argues, in respect to any corpus, '[i]n the long run, they matter very little'.

Practical considerations, furthermore, in practice often outweigh principles (Hunston 2002: 26) as researchers struggle to secure data of limited availability or inaccessible formats. Difficulties in obtaining text messages, due to their private nature, requires researchers to take data from different sources (How and Kan 2005) and settle for relatively small corpora (Hard af Segersteg 2002; Ling and Baron 2007) as in the current study. Accumulating larger corpora may be possible only for wider, well-funded projects (How and Kan 2005; Kasesniemi and Rautiainen 2002).[4]

The concept of *stability* may therefore be more important than corpus size. The methodological approach to determining corpus stability outlined by Dahlmann (2007) builds on the concept of *closure* (McEnery and Wilson 2001: 166) or *saturation* (Belica 1996: 61-74). Corpora have been said to be lexically saturated when, divided into segments, the number of new lexical items does not change with the addition of new segments

[1] Frequently occurring features, although of course occurring more frequently than less frequent items, arguably still require large corpora to capture the various ways in which they are used.
[2] Groom's (2003; 2006) inductive approach to the investigation of phraseologies circumvents the size problem by looking at more frequent key grammatical, rather than lexical, words and their patterns.
[3] Firstly, the notion of the 'word' is now a debated construct in linguistics (Sinclair 1991; 2004) and determining the number of tokens in CorTxt is particularly problematic, given alphanumeric strings such as *We'l av2go out4drink soon* and other omitted spaces, while the degree of structural ellipsis challenges the validity of word counts.
[4] Kasesniemi and Rautianen's (2002) study is part of a larger research project run by the Information Society Research Center (INSOC) at the University of Tampere, Finland which in 2002 had already been running for five years. The same data is also used by Oksman and Tautianen (2004).

(McEnery et al 2006: 15-16). As McEnery et al (2006: 16) point out, this concept of saturation relates only to what the rate of lexical growth says about how representative corpora are of a lexicon, but Dahlmann (2007) extends this to measure corpus stability in terms of the linguistic feature under investigation rather than new words. The linguistic focus of Dahlmann's research is frequent multi-word units (MWUs) retrieved through automatic extraction from two spoken corpora,[1] and her aim to assess effects on corpus stability of two different extraction methods: n-grams and Wmatrix (Piao et al 2003). The assumption is that the larger the dataset the less movement among the top ten MWUs and that a cut-off point can be determined in terms of corpus size, above which the top ten remain stable (Dahlmann 2007). The segments used by Dahlmann were of 25,000 words, and Dahlmann concludes that 200,000 words is 'big enough' for her analysis of MWUs. Her work highlights the need to judge each corpus not only according to its composition but also to its particular purpose (Dahlmann 2007).

The order of the most frequent words in CorTxt is therefore used to determine the degree to which the corpus can be considered stable. The total corpus, comprising 190,516 words was divided into 8 sections of nearly 25,000 words each. Accumulating wordlists were calculated in Wordsmith, for the first segment then for the first and second segments together, then the first, second and third segments, and so on. The table below shows that, despite some unexpected discrepancies (and similarities) and despite its relatively small dimensions, the corpus is very stable in terms of the most frequent wordforms.

[1] The two include a corpus of 400,000 words of native speaker English language (the ENSIC) and one of English learner language at 200,000 words (the NICLE).

Table 4.4 CorTxt Corpus Stability (method adopted from Dahlmann 2007)

	Segment 1 Top 10 words from 23,814 words		Segment 2* Top 10 words from 23,814 words x 2		Segment 3 Top 10 words from 23,814 words x 3		Segment 4 Top 10 words from 23,814 words x 4		Segment 5 Top 10 words from 23,814 words x 5		Segment 6 Top 10 words from 23,814 words x 6		Segment 7 Top 10 words from 23,814 words x 7		Segment 8 Top 10 words from 23,814 words x 8
1	W1	1	**W1****	1	**W1**	1	**W1**	1	**W1**	1	**W1**	1	**W1**	1	**W1**
2	W2	2	W3	2	**W3**	2	**W3**	2	**W3**	2	W2	2	**W2**	2	**W2**
3	W3	3	W2	3	**W2**	3	**W2**	3	**W2**	3	W3	3	**W3**	3	**W3**
4	W4	4	W6	4	**W6**	4	**W6**	4	**W6**	4	**W6**	4	W5	4	**W5**
5	W5	5	**W5**	5	**W5**	5	**W5**	5	**W5**	5	**W5**	5	W6	5	**W6**
6	W6	6	W8	6	**W8**	6	**W8**	6	**W8**	6	W7	6	**W7**	6	**W7**
7	W7	7	W4	7	W7	7	**W7**	7	**W7**	7	W8	7	**W8**	7	**W8**
8	W8	8	W7	8	W9	8	**W9**	8	**W9**	8	**W9**	8	**W9**	8	**W9**
9	W9	9	**W9**	9	W4	9	**W4**	9	**W4**	9	**W4**	9	**W4**	9	**W4**
10	W10	10	W12	10	W11	10	**W11**	10	**W11**	10	**W11**	10	**W11**	10	**W11**

*Segment 2 is in fact an accumulation of segments 1 and 2; and so on, until segment 8 is in fact an accumulation of segments 1, 2, 2, 4, 5, 6, 7, and 8.
**Items in bold are those which show no difference in position from the preceding wordlist. W1 is the only item not to change across the 8. Looking at segment 8, the fact that all items are in bold shows that there are no changes between segments 7 and 8. Interestingly, there are also no changes between segments 4 and 5. Most changes occur, as we might expect, in segment 2, followed unsurprisingly by segment 3 and more unexpectedly in segment 6.

4.3.3 Representativeness

The extent to which it can be argued that CorTxt is 'representative' of texting as a homogenous language variety is limited, not only because of its size and the non-random way in which participants were selected (as family and friends), but in light of the assertion that representativeness is anyway a questionable notion in corpus compilation (McEnery et al 2006: 73; Hunston 2002: 28-30). Obtaining representative samples (such as of text messaging), for example, requires accurate knowledge of the texting population and this, as with other varieties, is 'unknowable' (Hunston 2002: 28; McEnery et al 2006: 20-21).[1]

The solution adopted in the present study is pragmatic, recognising difficulties in ensuring representativeness under any circumstances, as well as practical challenges in obtaining data. CorTxt was compiled without excessive concern for balance but with careful recording of the demographic composition of message senders, and the parameters of the texting population then adjusted according to the nature of the sample.[2] In other words, the assumption is that a statement about corpus evidence is a statement about that corpus, not about the language or register of which the corpus is a sample (Hunston 2002: 22-3), and recognition given to the fact that generalisations must therefore be cautious and hedged (McEnery et al 2006: 73). What this requires is explicit 'bibliographic' information and design criteria (Sinclair 1991: 21; McEnery et al 2006: 18) which must be acknowledged when data is interpreted (Hunston 2002: 30) and which, in the case of texting, favours qualitative approaches over anonymous web-based collection-procedures unable to reliably

[1] Figures cited in text message studies suggest that, as reflected in CorTxt, females text more than males (Kasesniemi and Rautianen 2002; Ling 2004: 146-9; Faulkner and Culwin 2005) but that, unlike CorTxt, the texting population is predominantly young (Ling 2004; Faulkner and Culwin; Harkin 2002; Kasesniemi and Rautianen 2002; Ling and Yttri 2002). Older age-groups, however, cannot be ignored. According to the Mobile Data Association, the average age of mobile users in 2003 was already rising above 30 (Jellinek 2003). Puro (2002, in Faulkner and Culwin (2005), notes that phone ownership in Finland is 'very even' across all age groups except those aged over 60 who live alone. Gender and other sociological variables such as education and employment are equally difficult to ascertain.

[2] Another possibility for population sampling, given the widespread use of texting, is to aim for a sufficiently large and random to inevitably contain a diversity of texters, without worrying about the weighting to give each variable (Fairon and Paumier 2006). Certain variables will anyway be over-represented in relation to the number of messages actually sent.

Another solution is to study the texting practices of a particular group (Kasesniemi and Rautianen 2002; University of Leicester 2006). The advantages are that participants are likely to be homogeneous with respect to certain features (such as geographical location) so that boundaries can be clearly demarcated and the study made more manageable. However, although texters in CorTxt are described as comprising a 'texting network', they are neither as self-contained nor as homogenous as, for example, in Kasesniemi and Rautianen's (2002) research.

obtain personal data.[1] Conclusions drawn from CorTxt are made, and should be interpreted, in the light of the circumstances under which data was collected and the nature of the texting 'network' (see section 4.2.2 above),[2] and neither assumed to apply to texting as a 'whole' nor to, for example, teenage texting practices.

4.3.4 Transcription Error and Message Selection

Previous text message studies face similar issues of transcription error and message selection. Transcription error, whereby messages are copied incorrectly, arises because messages must be transcribed: in this study, either by hand (as in Kasesniemi and Rautainian 2002), or typed into an Excel template and emailed. Both methods risk transcription error. It is easy to make 'typos' when copying from tiny screens, to correct misspellings or add punctuation, observes Hard af Segersteg (2002: 208-9), while keyboarding itself, Sinclair (1991: 14) notes, is laborious. Although Kasesniemi and Rautainian (2002: 174) assert that no alternatives to manual transcription exist, messages could be forwarded to researchers' phones, despite incurring cost (Hard af Segersteg 2002: 210); or, if participants have adequate phones and equipment, data transferred directly to computers. Only one or two of my participants were able to transfer data directly, and issues included messages being split by the software, some problems in extracting messages from the software into a useable format, and restrictions on participants' opportunity to screen messages before submission. Until the software becomes less problematic and more widely available, the other transcription methods must be accepted, as they are for spoken language (Sinclair 1991: 14).

Message selection also remains problematic in this and other studies. Participants can select messages they wish to submit, because they are private, shocking, special or neutral (Hard af Segersteg 2002: 208-9); or invent them (How 2004: 6). Qualitative fieldwork involving personal contact with participants may reduce problems, by encouraging genuine participation (Kasesniemi and Rautianien 2002) and revealing through interview how messages were selected (Hard af Segersteg 2002: 209). In the present study, the close

[1] See the section on data collection above. Another limitation of collecting text messages through web-based questionnaires, for example, is that it favours texters with Internet access and know-how (Hard af Segersteg 2002: 207); while family and friends are neither randomly selected nor necessarily representative of texting populations.

[2] Participants in the current study can be said to comprise a texting *network*, not only in that they share certain social characteristics but that text messages are collected from a group who to some extent text each other.

researcher-participant relationships facilitated explanations of the need to send text messages which reflected 'day-to-day' texting practices. At the same time, however, participants were told not to submit messages which they did not feel comfortable sending,[1] and it is difficult to get around this when dealing with a private medium.[2] As with the issue of error transcription, solutions must be pragmatic and, to the extent that both are unavoidable, potential bias and possible errors in transcription acknowledged.

4.4 Ethical Considerations

4.4.1 Ethical and legal frameworks

The documentation of ethical decisions not only creates transparency but serves, given the lack of explicit legal guidelines and information regarding procedures followed in previous studies of texting, as a much-needed framework for future research into texted data. Corpus compilation involves consideration of legal and ethical implications of collecting and processing data comparable to researchers obtaining spoken (rather than published, written) data: participant consent, anonymising data, and ensuring participants' bibliographic information is safely stored. That is, the main ethical considerations in the current study are to protect the rights and interests of those who wrote messages used and those named in them. In ensuring that participants' interests are protected, procedures adopted in other studies can be drawn upon, as well as existing legal and ethical guidelines. In the UK, ethical guidelines established by disciplines such as BAAL,[3] institutions such as the University of Birmingham[4] or AoIR, the Association of Internet Researchers (Ess 2002), need to ensure researchers work within principles outlined in the Copyright, Designs and Patents Act 1988 (AHDS 2003)[5] and the Data Protection Act 1998[6] (Rock 2001: 8).[7] Working within ethical guidelines is, however, far from straightforward: the legal

[1] See section 4.2.1 where the above points are also made.
[2] In fact, as explored below, message selection forms a central part of the ethical procedure by which data was collected, in that it allowed participants to control which data should or should not be used in the analysis.
[3] See BAAL's 'Recommendations for Good Practice' for applied linguistic research: http://www.baal.org.uk/goodprac.htm, accessed 20th October 2007.
[4] See University of Birmingham's Code of Conduct, http://www.ppd.bham.ac.uk/policy/cop/code8.htm, accessed 20th October 2007.
[5] See http://www.opsi.gov.uk/acts/acts1988/UKpga_19880048_en_1.htm
[6] See Office of Public Sector Information, http://www.opsi.gov.uk/acts/acts1998/19980029.htm.
[7] Rock (2001: 8) provides an overview of legal frameworks, particularly as they relate to the anonymisation of data.

documents are somewhat complex and ambiguous,[1] whilst ethical frameworks themselves (necessarily) tend to state general principles allowing flexibility rather than concrete procedures or requirements (Ess 2002; BAAL 2006: 2).[2] Uncertainty increases with postgraduate research: although since September 2007 researchers at the University of Birmingham have been required to submit self-assessment forms for staff-led research projects[3] for ethical review by University Ethical Review Committees, postgraduate research cannot yet be reviewed,[4] and responsibility lies primarily with supervisors. Ultimate responsibility for making ethical decisions rests, however, as evident in legal and ethical frameworks, with specific consideration of the data and project in question. This means that even in research fields or institutions with established codes of practice, decisions and procedures regarding ethical issues are shaped by the nature of particular studies and by issues such as research purpose, the vulnerability of certain participants with respect to age, social status and powerlessness, and perceived degree of 'risk' to participants (Milligan 1999: 13; Ess 2002: 2).

Ethical guidelines for handling texted data cannot, unsurprisingly, be found. For example, ethical guidelines drafted by AoIR for electronic domains focus on regulating research into *public* online forums (Mann and Stewart 2000; Ess 2002) including Bulletin-Boards (Collot and Belmore 1996) and city-council emails (Hard af Segersteg 2002). Texting, in contrast, is highly private[5] and we could assume expectations among users that communication will not pass into public domains and therefore greater obligations on the part of the researcher.[6] However, message receivers are thus trusted to retain the privacy of messages, or at least to choose when messages can be distributed: Kasesniemi and Rautianen, for example, find teenagers often read (and compose) messages together. Close relationships noted between

[1] Rock (2001: 8) comments on how attempts to interpret legal guidelines can result in confusion: 'This EU legislation [the EU Parliament Directive 1995] has been enforced in Italy, resulting in such a degree of confusion that a group of academics are contacting legal authorities about the status of linguistic data'.
[2] Stringent legal constraints are imposed only in certain research areas, such as those involving patient data (e.g. Lee 2006), whilst projects covered by the NRES or the Animals Scientific Procedures Act have long been subject to institutional review
[3] Available: http://www.research.bham.ac.uk/ethics.shtml
[4] See: http://www.rcs.bham.ac.uk/documents/Clarification_note.doc)
[5] Although, in the case of text messages obtained from public online texting services, the public nature of the site reduces users' expectations of privacy. Messages collected by How and Kan (2005: 4) from the Yahoo SMS website were publicly accessible, and obtained neither with the consent nor identification of senders who thus remained anonymous. This is also the case with the messages obtained for the current study from the AOL website.
[6] Mobile phones are 'emblems of intimacy' used to promote trust and solidarity in teenage groups (Harkin 2002: 20-1) and by lovers who, according to Kasesniemi and Rautianinen (2002), send text messages as a sign of confidence in their relationship.

texters suggests receivers can gauge the level of privacy or sensitivity of each message. Researchers' responsibilities are also reduced by the relatively trivial content of messages often used for co-ordinating activities (as described in Ling and Yttri 2002) and the focus not on content but linguistic form (Ess 2002: 7-8). In fact, many of my participants expressed interest in the research (as in Kasesniemi and Rautianen 2002), suggesting benefits for those involved (Ess 2002). The reduced 'risk' to participants in studies of text messaging is perhaps reflected by the fact that researchers tend not to be very explicit in their treatment of ethical issues: Fairon and Paumier (2006: 3), for example, simply note that at the stage of translating their messages into standardised French, they were 'already anonymised'.

The following overview of ethical considerations for the current doctoral study of text messaging, chiefly consent and anonymisation, therefore serves not only to support the current study but proposes a possible framework for adaptation to future studies of text messaging; bearing in mind that ultimate responsibility for making ethical decisions will rest, as evident in legal and ethical frameworks, with specific consideration of the project in question.

4.4.2 Obtaining consent

Consent to text messages in the present study was obtained at two stages: firstly, through initial informed consent from participants who agreed, either verbally or through the act of submitting them, to submit messages;[1] and, secondly, through written consent obtained not only from initial participants but from all message senders where possible (see Appendix 4.2).[2] The consent form produced for the current study[3] meets specifications outlined by the Data Protection Act and AoIR's recommendations.

[1] And to select those which they did not want to submit, as explained in section 4.3.4 above.
[2] Kan (pers.comm) points out that in their study (How and Kan 2005), only messages sent by participants were used, thus ameliorating the need to gain consent from other senders.
[3] A draft of the consent form was drawn up by Malcolm Kendall of the University of Birmingham's Information Services team.

Figure 4.1 The consent form used in the current study

I consent to the contents of my SMS text messages between myself and others participating in the research project organised by Ms Caroline Tagg, sent during the period November 2005 until October 2007, being used by Ms Tagg for the purpose of linguistic analysis as part of her research activities at the University of Birmingham. I also consent to the use of any of my personal data necessary for the purpose of analysis in connection with this research. I understand that linguistic findings pertaining to the text messages may be used for a commercial purpose, but that individual messages or personal data will not.

I give my consent on the understanding that all such personal data relating to myself shall be processed in accordance with the Data Protection Act 1998 by Ms Tagg, that the personal data shall be destroyed by the end of 2010, and that, if the contents of the messages are ever published in a dissertation, thesis or other format for any purpose, all such content will be anonymised by Ms Tagg such that neither myself or my correspondents can be identified from that content.

SIGNATURE DATE

NAME

In the majority of cases, it was possible to obtain written consent, but exceptions included messages taken from the AOL website, and those from senders with large circles of fluctuating interlocutors who could not recall who had sent which messages. Hard af Segersteg (2002: 79) similarly notes that consent was obtained in 'as many instances as possible'. The argument that message receivers can be deemed responsible for obtaining consent from those who have sent them messages (who are likely to be, according to the literature, their intimates), or of judging whether messages be submitted, can be justified with reference to the guidelines outlined by the AoIR (as mentioned in section 4.4.3 above).[1] However, in the present study, personal information regarding participants was not obtained without consent from the individuals themselves. This was ensured by including the questionnaire used to elicit personal information in the same document as the declaration giving explicit consent (Appendix 4.2). In other words, where written consent was not obtained from someone who had sent messages (for reasons explained above), this was justified in two ways: firstly, that consent had either been obtained from the message receiver or that the messages had emanated from a public website; and secondly, that the messages were anonymous in that different senders were not distinguished and no personal

[1] This is also the stance adopted by Hard af Segersteg (2002: 210) in the second stage of her study, when participants forwarded all messages, sent and received, to her phone for analysis and where responsibility for obtaining the senders consent was left firmly with the participants: 'To forward messages that they had received, they [the participants in her study] had to have the original senders' permission'. The intimate relationship assumed between participants and senders also meant that the latter could be briefed by the participants and could raise any doubts with them.

information obtained. With consent obtained from those receiving messages and without accompanying bibliographic information (or even names), it is believed that participants' ethical rights are protected.

According to the European Union Data Protection Directive (1995),[1] not only must data subjects unambiguously give consent for personal information to be gathered and processed, but they must be informed as to why material is collected. My research goals were explained to the initial group of participants sufficiently for subjects to understand that linguistic form, rather than content, is the focus, as reiterated in the consent form. Documentation of the text message data with personal information 'necessary for the purpose of analysis in connection with this research' was also explicitly consented to, and it was made clear to whom the data will be made available: that while linguistic findings may be used for commercial purposes, neither the text messages themselves nor personal information would be. Senders were also informed that the personal data will only be held until 'the end of 2010'. In other words, consent was informed and participants were aware of the research purpose and boundaries.

Participants were also informed that messages would be anonymised so that no individuals could be identified (see section 4.4.3 below). Subjects need to be informed of procedures such as anonymisation not only because it forms part of what they are are consenting to but also, as Rock (2001: 7) points out, because consent is vital to the facilitation of the anonymisation process in that it allows researchers to gauge informants' perceptions of what anonymisation entails (Rock 2001: 7).[2] As discussed below in section 4.4.3, however, the wording of the consent form masks certain assumptions regarding the ease with which data can ever truly be anonymised.

The European Union directive also states that data subjects must be able to correct erroneous data and opt-out of data collection. The AoIR similarly frames the obtaining of consent as an ongoing issue, suggesting that, in order to protect their rights to privacy, confidentiality and autonomy as well as informed consent, subjects should be approached at the beginning of research projects and informed if research goals change significantly over time (see Rock 2002: 6). In the present study, to enable participants to correct data or opt

[1] Official Journal of the European Communities of 23 November 1995 No L. 281 p. 31.
[2] In her own study, for example, Rock (2001: 7), 'uses consent forms which detail items for anonymisation and invite participants to nominate additional items for anonymisation as they see fit'.

out, message transcripts were sent to receivers for correction once the corpus had been compiled, and research drafts made available. Although senders were not directly approached, a channel of communication was available through the participants.

4.4.3 Anonymisation

4.4.3.1 Principles and practicalities of anonymisation

The procedure followed in anonymising CorTxt emerges, as with other corpora, in part from the practicalities of what can feasibly be anonymised whilst maintaining data as much as possible in its original form (Rock 2001: 2). For example, personal information is intricately embedded into the text message below and anonymisation would significantly alter the data.

(4.1) Also knackered. Worked late every night - have become marking machine! And sleeping badly at mo. Observation nxt wk - don't know when yet as c not back from hols. Heard today have interview in telford weds. So buy wk ahead! Wld suggest meeting sat eve but not sure have time as lots of work n out all day sun

Given the nature of texting, it is unsurprising that much content is highly personal, albeit mundane, and therefore it is not feasible to remove all personal information without compromising the data.

The principle of protecting participants in corpus projects must, however, ultimately guide anonymisation. This principle is interpreted here, firstly, as requiring the replacement of all names and contact details relating to specific individuals who could subsequently be identified and/or located by people who do not otherwise closely know them.[1] For example, the embedded personal information in the text messages below relates to specific individuals and was therefore replaced.[2]

(4.2) NAME238 what number STREET do u live at? Is it 11?

[1] In other words, although I do not assume messages can be anonymised so that participants are identifiable to *nobody*, not even their texting partners, they are protected against, for example, work colleagues or fellow academics and acquaintances.

[2] In the first, the code STREET is used (in line with the general procedure adopted in this study), whilst in the second, the idiosyncratic use of nationality and initial are simply altered.

(4.3) | Bloody hell, cant believe you forgot my surname Mr NAME242. Ill give u a clue, its spanish and begins with m... X

The protection principle is also interpreted as requiring that participants themselves are comfortable with the personal information retained. As Rock (2001: 5) points out, researchers cannot necessarily determine what is potentially harmful without consulting participants, who may be uncomfortable with information that researchers find harmless. However, where it was not possible to contact participants, I exploited my position as close acquaintance and fellow participant to use my judgement as to what I would be comfortable with.

As well as juggling principles and practicalities, the seemingly straightforward process of anonymisation is the result of decisions made at every turn as to what anonymisation means and how this understanding can be applied to text message data. With as yet no 'standard' by which linguists can anonymise data, this section draws extensively on Rock's (2001) attempt to investigate the possibility of formulating standard policies which can be adapted to different linguistic research projects.

4.4.3.2 What to anonymise

What should be anonymised remains a matter of debate among researchers (Rock 2001: 9). One issue is whether anonymisation should be formal or 'true, complete': the problem being that 'formal' anonymisation, involving the automatic replacement of formal indicators of identity (such as names) may be insufficient in achieving 'true, complete' anonymisation based on qualitative analysis of data content. A 'true, complete' anonymisation relies on researchers' case-by-case judgements as to which items to replace, based on consultation with participants and close reading of data,[1] although complicated by the need to consider future corpus uses. The strengths are that, firstly, the risk of 'over-anonymisation' is avoided. Automatically replacing personal names in CorTxt, for example, meant that place names such as *Victoria Square* and *St Paul's Square* were distorted, and *June*, *bill* and *mark* were replaced with name codes in the following.

(4.4) | Are you available for soiree on June 3rd?

[1] As well as consideration of the research question and the items needed for the analysis (Todd 1999).

(4.5) Your bill at 3 is £33.65 so thats not bad!

(4.6) I have a rather prominent bite mark on my right cheek

Secondly, items which do not *formally* reveal identity can still be removed (Rock 2001: 10). Where a participant is known to the research community, for example, corpus-users may recognise participants from job titles or descriptions (Rock 2001: 9) and statements such as 'we all know that the town clerk fancies little boys', suggests Sampson (1999,[1] cited by Rock 2001: 9), could prove 'highly damaging'. In the following text message, for example, although the name is removed, the individual is potentially identifiable through the (fairly harmless) description of their duties.

(4.7) I think your mentor is NAME172 SURNAME13, but not 100 percent sure.

Rock (2001: 9) also cites Krishmanurthy's (1999) observation that project workers who contributed to the BoE could be identified by colleagues from unusual topics or 'turns of phrase'. Difficulties in anonymising aural recordings when people can be recognised by their voices (cf Du Bois 1999) can be compared to the possibility that texters can be identified through individual texting styles. However, although forensic linguists show that authorship can be attributed on the basis of texting idiolects (Coulthard 2007), it seems likely that only those in regular texting contact could recognise each other's orthographic preferences.

Disadvantages of a case-by-case approach include labour intensity and the fact that decisions are made subjectively and according to criteria not always obvious to other researchers using the same data (Rock 2001), as well as possible data distortion. For instance, as Rock suggests, information regarding dates and locations of conversations in the BNC may be crucial for studies of regional variation, and removing it would seriously diminish the data's value, while changing 'turns of phrase' would potentially invalidate, for dialectal studies, the linguistic data. It was therefore decided in this research to follow a formal approach. There is neither the time nor sufficient risk to participants to justify thorough 'case-by-case' consideration of each message. Over-anonymisation was alleviated

[1] Researchers cited by Rock include correspondence collected in 1999 from a corpora mailing list and a forensic linguistics mailing list. They are referenced as such in this thesis' bibliography.

as much as possible through manual checks, and the risk of retaining items other than formal indicators which may prove to reveal identity addressed by post-hoc consultation with participants.

The decision as to which formal indicators of identity to remove is not straightforward. Standard items include names and postal addresses but, Rock notes, bank details, telephone numbers, registration numbers, and dates of birth may also need to be replaced. The criteria used to guide anonymisation in CorTxt were that details relating to specific individuals, allowing them either to be identified or contacted, should be removed.

Items not anonymised included public places, firstly because they did not relate specifically to one individual: streetnames, for example, were retained but streetnames with house-numbers was assumed to be private addresses and anonymised. A more practical reason concerned difficulties in determining where to draw the line regarding which place names could be retained: *France*, presumably, would not be a risk; nor *London* or even *Birmingham*, the city where this research was conducted and read; but what about *Moseley* (a district of Birmingham), or the *Selly Soak* and *The Bristol Pear* (pubs in Selly Oak, Birmingham), or *Raglan Rd*? The other factor was the extent to which data would have to be manipulated for these changes to be effected, as well as decisions as to whether anonymised items could all be relabelled <PLACE>, or distinctions made between <CITY>, <VILLAGE>, <PUB>.

One exception is the following mention of a pub which, although public, relates specifically to the sender, the new landlord.

(4.8) Hi there. We have now moved in2 our pub, PUBNAME, POSTALADDRESS. PHONENUMBER. Would be great 2 c u if u cud come up.[1]

One final exception arose when one participant requested the change. The place name in this case was replaced by another, more general, description rather than by a code. The fact

[1] The example given previously ("NAME238 what number STREET do u live at? Is it 11? X"), in which a street name is replaced by STREET1, was another exception, made because the information being conveyed was a private address (specific to a particular person) rather than a public place and so was treated as much as possible as an address.

that other participants did not request changes supports the assertion that the inclusion of place names need not be considered a risk to participants.[1]

4.4.3.3 Replacement names or codes

What to replace indicators of identity with is no straightforward process. In the light of complications involved in randomly selecting replacement fictitious *names*, however, the more effective option is to replace names with codes (Rock 2001: 11-13). The problem with many possible codes is that, while they may include useful information, the choice of what to include (such as gender, rather than profession, age, role, or name parts) is arbitrary and can influence readers' sense of the data. A further issue is the complexity of coding, as well as the risk of opacity, for future researchers as well as the same researcher returning to the data (Rock 2001: 13). A basic code avoids these disadvantages, however, and is relevant for studies with less focus on individuals' sociological background and more on the data itself. The decision was therefore made, in the current study, to replace personal names with NAME followed by a number. The approach was formal, in that the same nameform or string of letters was given the same number throughout CorTxt, regardless of social criteria: that is, whether it belonged to a different individual or was an abbreviation or nickname of another named individual. To take invented examples, *Caroline* throughout CorTxt would be replaced by NAME1 and *Carol* by NAME2. Surnames were tagged separately, in similar fashion: SURNAME1, SURNAME2 and so on. In practical terms, basic codes are easier to select and keep track of than fictitious names, and they avoid questions concerning the extent to which names should carry social information.

Other contact details were replaced using the following codes.

[1] The same participant expressed general discomfort with some of the messages, and asked that 10 be removed. I went back into both versions of the anonymised corpus (the .txt file and the Access database) and removed the messages, replacing them with other messages from my inbox. This exercise proved interesting, not only because of the discomfort felt by one participant, but by the fact that the overwhelming majority appeared to be happy for their messages to be used.

Table 4.5 Contact details and replacement codes

Detail	Code	Freq.	Examples
Phone numbers	PHONENUMBER	69	Hey gorgeous man. My work mobile number is PHONENUMBER. Have a good one babe. Squishy Mwahs. Xx
Extension numbers	EXTNUMBER	1	Message from NAME214 SURNAME32 . I am at Truro Hospital on PHONENUMBER ext EXTENSION. You can phone me here. as I have a phone by my side
Postal addresses	POSTALADDRESS	38	Come round, it's POSTALADDRESS. Xxx
Postcodes	POSTCODE	5	Oops. 4 got that bit. POSTCODE
Email addresses	EMAILADDRESS	13	Hi my email address has changed now it is EMAILADDRESS
Car registration numbers	REGNUMBER	3	Yes, my reg is REGNUMBER. Ciao!
Sort codes and account numbers	SORTCODE ACCOUNTNUMBER	3 3	My sort code is SORTCODE and acc no is ACCOUNTNUMBER. The bank is natwest. Can you reply to confirm i've sent this to the right person!
Dates and times of birth	DATEOFBIRTH TIMEOFBIRTH	1 1	Hi NAME256 here. NAME11 have birth at TIMEOFBIRTH on the DATEOFBIRTH to NAME485 at 8lb 7oz. Mother and baby doing brilliantly.

Drawbacks to basic codes include the homophones throughout CorTxt. Confusing alphanumeric sequences were created, such as:

(4.9) say thanks2NAME1154me. Xx

where *NAME115* runs on into the homophone *4*. This, however, was infrequent, and other alphanumeric sequences were clearer. In the following, the homophone *4* occurs twice, but in both cases before the NAME-tag, so that the numbers are separate.

(4.10) NAME281's gone2the railway club4NAME156's b-day. He went at9for a couple of hours and stil isn't bak. The reason that i didn't go was so i cud stay here&finish essay, but NAME79 stayed here2wait4NAME281, so couldn't do anything! I'm a bit hurt that he's gone4nearly4hrs and its my last friday here. Am bit close2tears at mo, cos

don't know if i'm over reactin and don't know what2do.

Another disadvantage of codes is that the depersonalisation can distort readers' sense of the data. Given the issues concerned with using replacement names, however, these were felt to be minor and inevitable drawbacks.

4.4.3.4 Method of anonymisation

Methods for anonymising data can be manual, automatic or a combination of both (Rock 2001). Given the labour intensity involved in manual anonymisation, and the problems faced by fully automated methods whose failure to recognise, for example, film titles (*When Harry met Sally*) and institutions (the Queen Elizabeth hospital) (Rock 2001: 18) would only be further complicated by the unconventional spelling throughout CorTxt, a semi-automatic process involving thorough manual checks was adopted and is explored below.

The first step was to generate a case-sensitive word-frequency list from which words beginning with a capital were extracted. The capital-letter list was processed manually, and all sentence-initial words or place names deleted. These tended to be the more frequent, and included greetings (*Hello*, *Hey*) and first-person pronoun *I* (1788 occurrences). Deleting these left 526 apparent personal names, of which the most frequent had only 41 occurrences. Most names occurred just once (364), with 70 occurring twice, and 92 occurring more than twice. At this stage, however, it should be noted that the low frequency of each name was caused in part by inclusion in the list of the following as separate entries.

Figure 4.2 Variants of names considered as separate entries at this stage

- names followed by punctuation; so that *Louise* was listed separately from *Louise!* and *Louise's* (or *LOUISE*), as well as one example of *Annnnnaaaaaa!*;
- names as part of alpha or alphanumeric strings, with or without punctuation: *Hi.Ruth*; *Robbie!I'm*; *Maryxxx*; including some without capitals that would otherwise not have been picked up: *Come2dale's*; *U&nikki*; *Me+stephanie*; *Lunch?steve*;
- probable misspellings of certain names: *Caroline* occurred also as *Caoline*; *Amanda* as *Amanada*; *Chris* as *Chrid*; *George* as *Georpge*;

- names that may be abbreviations of other names: *Rache*, for example, occurred as well as *Rachel*; *Ads* as well as *Adam*.

Other issues to be resolved included the following.

Figure 4.3 Issues arising from the anonymisation process

- Names had yet to be checked as to whether they referred to private individuals or public figures (*Robbie*, it was suspected, referred to the singer *Robbie Williams*; *Paula* to the runner *Paula Radcliffe*; and *Jon* to *Jon Stewart*, the American comedian). Public figures would not be anonymised on the basis that their names were in the public domain and the reference was to their public personae, rather than to them as private individuals;
- Some words judged to be unusual names or nicknames but not yet checked: *Chuckles*; *Danno*; *Ieri*; *Tinalunen* (common pet names such as *Honey* or *Bird* were not included in the list);
- Some postcodes had been picked up and were retained at this stage as these would later be anonymised;
- Names occurring only in lower case were not identified.

The next step was to remove duplicates of the same nameform: items differing only through punctuation (*Louise!* and *Louise*, for example) or because of their inclusion in strings involving punctuation (such as *U&nikki*); and to generally clean up the list (by taking out postcodes, for example) so that only names remained. This left around 350 names, bearing in mind that, at this stage, names which differed because of added letters rather than punctuation were retained (eg, *Nikkis* or *Helenx*).[1]

It was necessary to separate nameforms that could also be other words or occur within other words (the *alsowords* list) from those likely to occur only as names (the *onlynames* list), to facilitate automatic replacement: searching for names surrounded by spaces and punctuation was deemed inefficient (as those embedded in alphanumeric and other sequences would be missed) yet necessary for names which occurred also as parts of words. The following text message illustrates what happened when all names were replaced automatically.

[1] Where *x* is a symbol representing a kiss at the end of a message.

(4.11) Hey mate. Spoke to the mag people. We're on. NeNAME211 the isNAME349 deliverNAME211 by the end of the month. DeliverNAME211 on the 24th sept. Talk later. NAME182

NAME211 is *Ed*, which occurs in *Need* (*NeNAME211*) and twice in *delivered* (*deliverNAME211*); NAME349 is *Sue*, which occurs in issue (*isNAME349*). Other examples are as follows.

(4.12) Yes. Last skittNAME352 practice.

(4.13) Back in brum! Thanks for putting us up and keeping us all fNAME211 and happy. See you soon xx

(4.14) Oh unintentionally not bad timing. Great. Fingers cNAME513NAME211 the trains play along! Will give fifteen min warning.

Other nameforms that could be parts of other words (in the *alsowords* list) included *Ads*, *Al*, *Anne*, *Beth*, *Dan*, *Di*, *Ed*, *El*, *Han*, *Les*, *Ron*, *Tom*, *Val*. Examples are *Di* in *Didn't think so*; or *Les* in *Just going to college to enrole shud b back soon unless u want me un pick them up?*. Names which were also other words included: *Mark*, *Bill*, *Carol*, *Gay*, *Cat*, *June*, *Miles*, *Nick*, *Rich*, *Rob*, *Rose* and *Sue*. One example is *rich* in: *You'll end up a workaholic but rich*. Nameforms felt to occur only as names included *Adam*, *Amanda*, *Brendan*, *Caroline*, *Charlotte*, *David*, *Emily*, *George*, *Hannah*, *Jennifer*, *Katie*, *Elizabeth*, *Pamela*, and *Stephen*. Names in this latter group which differed from others in terms of added letters such as *s* or *x* were then removed[1] and some misspellings (such as *Caoline* and *Geoprge*) corrected.

Categorising names into *alsowords* and *onlynames* involved manually checking the list against CorTxt. It was therefore possible to eliminate apparent names that did not actually occur as names, such as *Miles* which was found to refer only to distance, as in: *Can't believe u ushed me from 75miles away!* or as part of *smiles*; or *June*, which occurred only in reference to the month: *We're up wk beg 18th june*. Finally, it was also necessary to order the list so that longer names which included shorter names within them were anonymised

[1] But not from the former list of those which occurred as other words or word parts.

first: *Janet*, for example, came before *Jan*; *Rachel* before *Rach*; *Christopher* before *Chris*. This was because otherwise *Chris* in *Christopher*, for example, would be automatically anonymised before the complete wordform *Christopher* could be, leaving *NAME18topher*.

Two different Perl scripts were then run on the two lists, in order to extract and replace names, either with spaces at the beginning and end (for *alsowords*) so that parts of other words are not included; or without (for *onlynames*). All occurrences throughout CorTxt of, for example, *annie*, were replaced by the code NAME1; all occurrences of *alan* by NAME2.[1]

4.4.3.5 Manual checks: problems and decisions

One observation, once the scripts had been run, was that more names than anticipated should have been placed in the *alsowords* list.

Figure 4.4 Examples of additional names which need to be put in the *alsowords* list

Ross	"All sounds good. Fingers **cNAME513ed**. Makes it difficult to type."
Lisa	"For fear of fainting with the **reaNAME15tion** of all that housework you just did? Quick have a cuppa!"
Kate	"Been running but only managed 5 minutes and then needed oxygen! Might have to resort to the roller **sNAME182s** option!"
Emma	"He said it's great hooking up with u as he always has a good time and that u r really great and fun 2 b with. He knows he has left his charger but it means no **dilNAME68s** when u r back as u have a good reason 4 c ing him. NAME281 is keen 2 come 2 Cardiff 4 my birthday as he also likes Cardiff, he thinks its posh and really great like u, I said don't worry about money and he said he would try and get some together, xxx"
Alan	"Have a good trip. Watch out for **avNAME9ches**. Remember when you get back we must decide about easter. Xxxxx"

Spaces were added to the Perl script for these names, so that they were treated as *alsowords*, but the experience highlights problems in automatic anonymisation. Although it should have been possible to foresee some of the above, it is arguably impossible to predict all possible occurrences, especially for a medium characterised by orthographic variation.

[1] These are not the actual names annoymised with these codes.

Manual checks were needed to check *alsowords* that occurred, as names, within alpha- or alphanumeric strings, names that only occurred in lower-case and so escaped the automation process, and names referring to public figures; as well as to change postal addresses and other details. This was done by looking through CorTxt, message by message. Once a name was found, this was searched for throughout the .txt file using Microsoft Notepad Find and Replace. Judgement was used as to whether it was worth searching for names without spaces[1] and careful notes made of changes.[2]

As well as identifying names on the *alsowords* list missed in the automatic process due to their not being surrounded by spacing, around 200 new names (those that did not appear at all in the lists) were found, largely because they did not occur with initial capital letters. As each was noticed, it was searched for using Find and Replace. Surnames were also identified, some of which had originally appeared in the capitalised namelists. Most references were to firstnames only, and only 47 surnames were identified (compared to over 500 names). The decision to recognise these as distinct from firstnames was made on the basis that using two NAME-tags in a row was misleading and impossible to distinguish from a list of two names without commas, as in:

(4.15) | Well done NAME40 NAME242 and NAME324! luv ya all Xxxx |

At the same time, surnames constituted separate words and so could not be consumed under the same NAME-tag as the firstname.

Initials were not considered a risk and were left in the corpus, both when occurring by themselves and with names or surnames.[3] The only initials changed, however, were a few prolific contributors who consistently signed off messages with an initial. Due to their relationship with me, the researcher, and having checked with the contributors themselves, the decision was made to replace such initials as the following with other initials.

[1] *Chris* could be, as it occurred either as the name or in *Christmas*, but *Tom* occurred too many times in *tomorrow* to enable searches for all occurrences of the string; and *Rob* likewise in *problem*, *probably* and *prob*

[2] These notes were later destroyed as part of the anonymisation process

[3] As with, for example: *Hello, i found my rat fink statement. It says if i repay early i win a cabbage. Do i still get this?? Bank details are SORTCODE ACCOUNTNUMBER miss am SURNAME57, incase you didn't know my name*.

(4.16) | Yo! Howz u? girls never rang after india. L[1]

Decisions as to whether letter-strings were names was made on the basis of their being recognisably common names (*Christopher*, *Stephanie*, *Pam*), my knowledge of the texters and so recognition of less-common names, as well as the position of the word in the message. Some words, especially when uncapitalised and/or abbreviated, were not recognisably names, but their position at the end of the message, or at the beginning following greetings, meant they were classified as names. Similarly, the position of words which could be names, as well as my background knowledge of people involved, also meant that they were sometimes considered not to be names. In the following, *tom* was interpreted as an abbreviation for *tomorrow* rather than the name *Tom*.

(4.17) | Can you let me know details of fri when u find out cos I'm not in tom or fri. NAME139 mentionned chinese. Thanks[2]

To check some items, I contacted message-contributors. One participant, for example, confirmed that *Finn's* was a nightclub not a friend in the following.

(4.18) | I think we're going to finn's now, come

Names of pets or toys were also problematic, but the decision made to anonymise these on the basis that they could aid identification of their owners, that there may be names which referred to animals or objects which could not be distinguished from people, and that if they were obviously pets or toys this would remain obvious even if anonymised. Terms of endearment such as *honey* were retained, but nicknames directed at particular people changed.[3] In one message sent to me, the play on my name is impossible to anonymise without removing the text message: as I was the participant involved, it was not anonymised.

[1] The initial here is already anonymised.
[2] Other occurrences of *tom* were more ambiguous but the following deemed to refer to tomorrow rather than a person: *Hi NAME458 i will get NAME128n to call you tom he has the details for c l k cab*; *So soon - tom will b fab! Sleep well(ish) xxx*
[3] Incidentally, this does not address the problem arising in the opposite situation, where one person uses a particular name to address a range of interlocutors and the nickname becomes part of their idiolect which could be used to identify the sender. This was the case with *bird*, used by one sender to address a number of interlocutors.

(4.19)	Nutter. Cutter. Ctter. Cttergg. Cttargg. Ctargg. Ctagg. ie you

Public figures or institutions were not anonymised[1] and, in many cases, this involved unanonymising names which had been changed automatically.

[1] Where I was unsure as to whether the name referred to a public figure, the name was anonymised.

Table 4.6 Items left unanoymised in CorTxt

Item	**Examples, exceptions and challenges**
Celebrities	*Agatha Christie, Paul Newman, Paula Radcliffe, David Blunkett, Jack Nicholson* The distinction was made, however, between public personae and their private lives: the name Norman Fairclough was retained in relation to his giving a public talk, while the names of other well-known linguists were anonymised when mentioned in private settings.[1]
Shows or films	*Saw Guys and Dolls last night with Patrick Swayze it was great* *Night sweet, sleep well! I've just been to see The Exorcism of Emily Rose and may never sleep again! Hugs and snogs! Xx* *Is that NAME115 on the telly? No its Brdget Jones!",*
Public places	*There are some nice pubs near here or there is Frankie n Bennys near the warner cinema?'* *I shall book chez jules for half eight, if that's ok with you?* *Thanks. It was only from tescos but quite nice. All gone now. Speak soon xxx*
Characters from films or TV series	*Well I wasn't available as I washob nobbing with NAME42 last night so they had to ask Nickey Platt instead of me!;* *Its on in engalnd! But NAME281's telly has decided it won't let me watch it and mia and elliot were kissing! Damn it!*
Reality show contestants	*Jordan got voted out last nite!* It proved difficult to distinguish gossip about reality-show contestants from private individuals[2]. This example refers to *Big Brother*: *Double eviction this week - Spiral and Michael and good riddance to them!*
Words that are not names	*Bill*, as in: *Are there any letters for me. i'm expecting one from orange that isn't a bill but may still say orange on it*. *St Paul's Square*: *Hello! I've just accepted an invite to a restaurant in st paul's square - leaving now - we are a random bunch and should be in the actress and the bishop by nine thirty. Would you like to join us there?* *Em*, which was often an abbreviation of *them* as in: *I've got2ring em on monday to let em know if its beta or not*; although elsewhere a name.

[1] Skelton points out that public figures should receive the same protection in private discourse or as private individuals as other people (in Rock 2001).
[2] Anonymising this name was not only felt to be unnecessary but also misleading, as it suggested a private individual, rather than a public figure, was being discussed.

A line had to be drawn between public and private contact details. Although an extension number was anonymised, as it related to a specific individual at least at the time of texting, the area code in *I rang on an 01823 number, the withheld one's not me. X* was retained on the basis it could not be used for identification. Similarly, with addresses, it was felt identification was only possible where specific house-numbers or postcodes were given so information regarding roads was retained in the following.

(4.20) Ok thats cool. Its POSTALADDRESS, just off either raglan rd or edward rd. Behind the cricket ground. Gimme ring when ur closeby see you tuesday.[1]

4.4.3.7 Summary

The anonymisation procedure adopted here rests on the principle that participants be protected and their identity concealed to their satisfaction, bearing in mind what is feasible without seriously compromising the data: which, in the case of text messaging, contains much personal yet mundane information. The procedure selected involved replacing names and contact details of private individuals with basic codes. The addition of numbers to each code conveys textual links between references to the same name; however, the approach is formal in that the same nameforms are linked, rather than references to the same person. The procedure is also formal in that indicators of identity are replaced at the expense of 'true, complete', but more time-consuming and difficult, anonymisation. It should also be noted that, partly because of the unconventional language of Txt, the process could only be semi-automatic and relied on thorough manual checks, with care taken to acknowledge and outline problems and subjective decisions made. While it is felt that participants are thus adequately protected from risk, the caveat remains (as in any research) that true and complete concealment of identity can never be wholly ensured.

4.5 Corpus Storage and Handling

4.5.1 Initial processing

Only minimum processing was necessary in order to prepare the text messages for analysis. Several text messages were split into sections[2] and were manually joined, an issue also

[1] This can be compared to the message cited previously, in which the address related to a particular individual: NAME238 what number STREET do u live at? Is it 11?

[2] Where the message had exceeded the character limit set by the phone operators who had subsequently sent the text in two parts

faced by Fairon and Paumier (2006) who observe that phone operators' various 'splitting methods' meant messages cannot be automatically joined. Fairon and Paumier also removed messages which did not conform to 'project recommendations'.[1] The boundaries of the current target population are less stringently demarcated than Fairon and Paumier's but a few messages sent by mobile phone companies were removed.

However, business-related messages were retained (see Table 4.1), as were any which mentioned my research or indicated participants' awareness that their messages would be analysed. In the following, the sender greets me (the researcher) at the end of a message to another individual.

(4.21) Hello.How u doing?What u been up 2?When will u b moving out of the flat, cos I will need to arrange to pick up the lamp, etc. Take care. Hello caroline!

This was then mentioned in a message from the above receiver to myself in which she replies to my thanks for a collection of text messages.

(4.22) I take it the post has come then! You must have 1000s of texts now! Happy reading. My one from NAME13 wiv hello caroline at the end is my favourite. Bless him.Xx

Graphic messages were also retained.

(4.23) :-) :-)

(4.24) :) Xx

4.5.2 Storing the corpus

Corpora must be stored so that they can be processed automatically and source information can also be included. Most importantly, however, care must be taken to ensure participants' rights are protected in accordance with the UK Data Protection Act (1998) and that

[1] These included 'commercial messages, forwarded SMS obviously written by another person, SMS written for the attention of our team, non- French SMS, graphical SMS, etc'.

participants are informed of this (see section 4.4.2). It is believed that this information, and the procedures followed in storing and processing data, comply with the Act.[1]

CorTxt is stored both as a text-only file (*.txt*), to enable it to be processed in Wordsmith Tools (Scott 1996), and in a *Microsoft Access* database. The decision to store the corpus in an Access table was made on the pragmatic basis that the programme is easily available, relatively user-friendly, able to generate queries and produce forms, and compatible with Microsoft programmes such as *Word*, *Excel*, or *Notebook* (the latter of which stores text-only files) which meant data could be imported and exported between them. As well as generating queries, *Access* also enables selected information columns (or fields) to be sorted alphabetically. The fields include: *Date*, *Time*, *Sender*, *Receiver*, and *Message*. The inclusion of personal or contextual information, alongside data, is discussed further below, as one aspect of corpus mark-up.

4.5.3 Corpus mark-up

Marking-up a corpus for either textual or contextual information is vital in corpus-compilation not only to structure corpora but to enable finer analyses of subsets within the data (McEnery et al 2006: 75). However, the form that mark-up takes again mixes consideration of practicalities and the need to bear research questions in mind. The only way in which textual information was marked up in CorTxt was that the beginning and end of messages was indicated by speech-marks ("*message*").[2] The relative lack of textual mark-up can be explained, firstly, by practical considerations: as McEnery et al (2006: 75) point out, mark-up tends to be time-consuming and, if displayed, may render corpora less reader-friendly. Also relevant is the nature of the data: not only are texts messages too short to include paragraphs but sentences and clauses are difficult to identify, due to non-standard punctuation. The main factor however is that, as with spoken data, contextual information was more relevant. Contextual information was stored in a table within Access (*context* table) and linked to the table of text message data (*data* table). In order to protect participants' identity, no names were included in the context table; instead, participants

[1] Ethical guidelines acknowledge the need to protect participants' confidentiality in accordance with the Data Protection Act: in Milligan's (1999) words, '[a]ppropriate measures should be taken to store research data in a secure manner'; and consent forms should contain: 'A description of the intended uses, and disposal/storage and documentation procedures for data including an option to agree/disagree with these procedures' (Milligan 1999).

[2] This was done using automatically when messages were exported from Access.

were given codes (P1, P2 etc). These codes were linked to names in a separate file held separately from the data and context tables and, once data-collection and mark-up was finished, destroyed. The context table included the following fields: date of birth, gender, profession, native language, make of phone, how long the user had been texting, whether they used predictive texting and how often they sent messages.[1] For the purposes of the current thesis, however, the information was only used to describe the corpus as a whole, and not to distinguish between individuals.

4.5.4 Normalisation

The present decision not to standardise or normalise the language (to translate the unconventional spellings into a more conventional orthography) was made despite practical advantages. Fairon and Paumier (2006), for example, translate their corpus into 'standardised French' for reasons of readability and usability: not only that researchers perhaps unfamiliar with texting practices can read messages and distinguish between typing errors and unconventional but intended spellings, but also that they can conduct automatic word- and phrase-frequency counts. However, as Sinclair (1991: 21) argues in support of his 'clean-text policy', however logical and justified moves to annotate language may appear, they will not necessarily be accepted by all linguists, and the process remains highly subjective. The decision not to standardise the spelling also recognises the argument that non-standard orthography is involved in the expression of identity (Sebba 2007), the meaning of which would be lost in a 'spell checked' version (Shortis 2007a: 4). If, as suggested in the next chapter, texters are using unconventional spellings to create identity and achieve effect, normalisation strips away an essential part of the interpersonal meaning conveyed.[2] However, with research into unconventional spelling in texting and other forms of electronic communication still in its infancy, decisions made when 'normalising' Txt at the initial stages of this project would threaten to impose pre-existing perceptions of the role and rules of spelling onto an as-yet little understood medium (although see Shortis 2007), and onto future corpus-users.

[1] Information obtained using the questionnaire attached to the consent form. Once information had been entered electronically, the paper questionnaires were detached from consent forms (which included names and signatures) and were destroyed.
[2] See Chapter 5: Respellings in Text Messaging.

4.5.5 Corpus exploitation

The procedures followed for exploiting the corpus are shaped primarily by the intention, as outlined in Chapter 3, to let the texted data drive the findings. However, given the fact that text messaging is a largely unexplored and highly distinctive language variety, as well as the relative lack of existing corpus-driven studies of language variation, the study in practice explores various ways into the data. In so doing, it probes the question at the heart of the distinction between corpus-based and corpus-driven approaches: how to initially identify and retrieve features of interest, and the need to be both pragmatic and flexible in response to a specific corpus. The procedures, summarised below, are explored further in the relevant chapters.

One procedure, drawing on existing research into the social meaning of respellings (Sebba 2007; Shortis 2007a,b), involves the retrieval of spelling variants. These were identified manually from a frequency wordlist compiled through WordSmith Tools (Scott 1996), using the lemmatisation function normally reserved for grammatical lemmas. The spelling variants were then interpreted according to categories outlined in the literature (e.g. Weber 1986; Sebba 2007). This procedure, although largely corpus-based, quantified and in some cases challenged previous statements regarding Txt spelling.

A second procedure involved exploitation of the wordlist, and its redefinition as a wordform list. The list generated in the above procedure was firstly compared to that of the BNC, which revealed significant differences alleviating the need for a keyword analysis. Differences were then explored with reference to the phrases (recurring multi-wordform units) in which significant wordforms occurred. These phrases were identified through the generation of concordance lines, using Wordsmith's Concord tool, which were then sorted alphabetically so that patterns emerged. Concord's Cluster Analysis tool was also used to support the findings, as were retrieval of the most frequent 3-, 4-, 5- and 6-wordform phrases in the corpus. Findings were compared to those emerging from investigation of spoken and written corpora (O'Keeffe et al 2007), and interpreted along similar lines.

A further and very different line of investigation arose from noticing, whilst exploring the data, somewhat unexpected occurrences of everyday creativity in the corpus. As in Carter's (2004) study of spoken creativity, examples to an extent 'leapt out' from CorTxt, while others were retrieved through close, linear reading of the data, facilitated by my familiarity

with participants' backgrounds as well as the manageable size of CorTxt. Whilst this precludes attempts to quantify creativity, the fact that discovery of creative patterns would not necessarily be possible through more quantitative methods highlights the need for a diversity of exploitation procedures.

A final procedure, emerging both from descriptions of Txt as being 'speech-like' and similar observations made of CorTxt, was based on Carter and McCarthy's (2006) descriptive model of spoken grammar. A variety of procedures were followed in retrieving grammatical features from CorTxt. Clausal analysis was carried out on a random selection of 100 text messages, while investigation of situational ellipsis involved a combination of corpus reading, and exploration of concordances generated in WordSmith for auxiliary verbs (*am*, *is*, *will*, *don't*) with only limited and cautious conclusions reached regarding overall frequency. Headers and tails, as described by Carter and McCarthy, presented even greater retrieval challenges, and could be retrieved only by reading CorTxt. Investigation of the deictic determiners *this* and *that* involved searching in WordSmith for the two items, and use of the Set function in Concord: firstly in manually distinguishing between their uses and word classes, and secondly in identifying various references the determiners made. Similar procedures were followed for discourse markers and fixed tags. Response tokens were explored by retrieving in alphabetical order all message-initial items, and judgments made according to form, function and context as to whether each occurrence was a response token. Statements regarding frequencies of response tokens, as well as discourse markers and fixed tags, could therefore be made. Although the process of retrieving previously-identified grammatical features is corpus-based, Carter and McCarthy's model itself emerges from inductive research into spoken data and allowed comparisons with texted data to be made; whilst again focusing on aspects of language overlooked by a word frequency approach. As mentioned above, this and the other procedures making up the eclectic and exploratory approach to exploitation of CorTxt are discussed further in the relevant chapters.

4.6 Chapter Summary

This chapter outlined and explored current corpus specifications and procedures and in so doing paved the way for further research by exploring new challenges raised by the compilation, storage and exploitation of a text message corpus, alongside considerations

also addressed in the compilation of other corpora: corpus size and representativeness, sampling, and ethical considerations. Data collection is complicated by the intimate nature of texting and by technological features, although studies show that individuals are willing to submit messages. Benefits of an approach in which participants are known to the researcher is firstly researchers' familiarity with participants' backgrounds and texting practices; and secondly, that it gives the researcher control over issues of message selection and transcription error, issues which must nonetheless be acknowledged. This data-collection method also has beneficial implications for corpus representativeness, assuming that the best way for corpora to be deemed representative is to carefully define parameters and limit generalisations to the sample. As for corpus size, this is inevitably an issue with texts averaging 17.2 words, but a stable corpus of less than 200,000 words can be justified pragmatically as well as by reference to the number of texts included (11,067) rather than the number of words. Furthermore, texting raises new challenges in regard to ethical issues, complicated already by the lack of ethical frameworks for linguistic researchers and the fact that guidelines must always be interpreted according to particular datasets. Anonymisation is difficult, given the personal yet unconventional nature of texted language, yet this chapter outlined a framework for reference by future researchers. Finally, the various methods used in exploiting the corpus are shaped by the intention to let the data drive findings whilst recognising the need to be pragmatic and flexible when faced with a new and distinctive dataset. A complex balance emerges between *principles* (such as representativeness, protection and delayed human intervention) on the one hand and *practicalities* (what can be done given the nature of the data and goals of the research) on the other, the need for which is also acknowledged in the general literature on corpus design (Sinclair 1991; Hunston 2002; McEnery et al 2006). As well as guiding the present investigation in achieving an appropriate balance, this chapter thereby provides guidelines which subsequent researchers of texting (and other electronic domains) can interpret and adapt to their datasets.

CHAPTER FIVE

RESPELLINGS IN TEXT MESSAGING

5.1 Introduction

Greater understanding of the variation in Txt spelling is necessary given public concern over what is described as 'bleak, bald sad shorthand' (Sutherland 2002, writing in *The Guardian*) as well as alarm over its presumed effects on other writing domains: 'Girl writes English essay in phone text shorthand' (Cramb 2003, in the *Daily Telegraph*). Spelling variation in the following text message from the text message corpus, CorTxt, illustrates the language behind these concerns. Respelt forms include *u*, *r*, and *b* (for *you*, *are*, and *be*); *wil* (for *will*); *bout* (for *about*) and *goin* (for *going*).

(5.1) Looks like u wil b getting a headstart im leaving here bout 2.30ish but if u r desperate for my company I could head in earlier-we were goin to meet in rummer. X [1]

Studies based on text message data suggest, however, that Txt spelling constitutes an appropriate, creative response to features of the texting situation (Hard af Segersteg 2002; Ling and Yttri 2002; Shortis 2007a,b). By situating Txt into a sociocultural model of orthography, we can go further and suggest that respellings in texting are, as in other domains, *principled* and *meaningful* (Sebba 2007). As Shortis (2007a,b) argues, in terms of both form and motivation, the apparently unconventional language of texting thus follows and extends traditional patterns of spelling variation: found in medieval handwritten manuscripts (Bradley 1919), early printing practices (Baron 2000; Leith 1983), Victorian letters (Crystal 2003: 120; Kesseler and Bergs 2003), fanzines (Androutsopoulous 2000), children's writing (Kress 2000) and other forms of electronic communication (Shortis 2007a,b). As in these texts, respelling in texting is seen in this thesis as conveying meaning, through communicating informality, creating illusions of brevity, defining group boundaries, and thus construing texted identity.

The following investigation of CorTxt exploits empirical data to substantiate and refine previously-made assertions in this area regarding the use and frequency of occurrence of

[1] 'rummer' refers, apparently, to a pub.

spelling variation in texting. After defining terms used, the chapter outlines the argument that spelling makes meaning (Jaffe 2000; Sebba 2007). Principles of spelling variation identified in the literature are then outlined, along with their functions, both of which are traced back through existing and historical spelling practices across formal and informal texts (Sebba 2007; Weber 1986). Respellings in CorTxt are then systematically investigated and the possibility raised that text spelling constitutes one realisation, alongside grammatical ellipsis and recurring phrases,[1] of what I call *performances* of spoken informality, brevity and deviance, by which texters develop relationships, define group boundaries and construe texting identities, in culturally-determined ways shaped by the medium. The contributions made by this chapter are therefore twofold: firstly in the empirical and quantifiable claims made on the basis of a large corpus and, secondly, in the interpretation of respellings and other linguistic choices as *performances* of identity.

5.2 Spelling

5.2.1 Terms used

The term *spelling* rather than *orthography* is used in this study to describe variation in texting. This follows the distinction between *orthography* as the conventions of written languages (determining possible letter-sequences and the phonemes that graphemes represent) and *spelling* as their application to particular words, so that the way in which we spell words is shaped by orthographic principles (Sebba 2007: 11).[2] Despite the similarity of the terms (*orthography* and *spelling*), in this investigation of variation in how words are spelt in texting by literate English-speaking people within Britain, it is appropriate to talk of variation in *spelling* rather than orthographic variation, with the point made that 'unconventional' spellings draw on the same orthographic conventions. *Woz*, for example, works as a respelling of *was* because although both follow principles of English orthography, the former adheres more strongly to the expected sound-symbol correspondence.

Terms such as *non-standard* or *deviant* spelling appear across the orthography literature (Jaffe 2000; Sebba 2007). However, *respellings* (as also used by Jaffe 2000 and Sebba 2007) and *spelling variants* or *variation* (used by Weber 1986) are preferred in this thesis: not

[1] See Chapter 7: Wordforms and Phrases; and Chapter 8: Spoken Grammar, Texted Grammar.
[2] This distinction is also maintained, if implicitly, by other researchers (Carney 1994: xvi-xvii; Mitton 1996): Mitton (1996: 2), for example, defining orthography as 'the way that English is spelt'.

least because of difficulties in defining 'standard' and hence 'non-standard' (Stubbs 1988; Milroy and Milroy 1991; Wilkinson 1995). For example, definitions of non-standard spelling often recognise not only its departure from codified and prescribed standards but its use in informal, unregulated situations, as in Androutsopoulos's (2000: 514) definition of non-standard spelling as:

> a general heading for spellings that diverge from standard (codified) orthography and/or do not occur in formal writing.

However, although words tend to have one codified, invariant standard spelling, some escape clear categorisation as standard or non-standard: idiosyncratically-spelt personal names (Carney 1994: 443-466) and product names or logos such as *Beanz Meanz Heinz* are both permitted unconventional spellings within otherwise highly regulated domains. Also omitted are common initialisms and acronyms such as BBC and UNESCO which occur across mainstream writing (Crystal 2003: 120-1; Weber 1986). These variants often enter dictionaries and can perhaps be included in what Sebba (2007: 30) calls *licensed* or legitimate variation such as between *judgment* and *judgement*, to take Sebba's example, or *realize* and *realise*, where 'conventional norms allow for a choice'.[1] This suggests that focusing on what is non-standard (or unlicensed) may overlook accepted (licensed) forms of spelling variation; while the increasing trend of dictionaries to include 'informal' language complicates the association between the standard and the formal, public sphere: *WYSIWYG* and *IOU* occur in dictionaries but would hardly be acceptable in academic essays.[2] In other words, whether spellings are judged to be 'standard' depends on context.[3] Another complication is the notion of (in)correctness that inevitably colours popular perceptions of 'standard' (Wilkinson 1995: 29). Spelling carries acknowledgement of normal or *right* forms (Sebba 2007: 10): as Kress (2000:1) puts it, spelling is often defined as 'knowing how to write words correctly'.[4] Kress's (2000: 8) distinction between *correct* and *accurate* spelling recognises that even where children's spelling does not match

[1] In contrast to unlicensed spellings, which contravene the standard.
[2] 'What you see is what you get', recently spelt in a school homework assignment as *WUSIWUG* where *you* is represented as *u* (Cramb, 2003) and 'I owe you'.
[3] As another example of this, Kress (2000: 139) points out that while children's efforts to transcribe what they hear or to show evaluative meaning through the size or shape of letters may be deemed erroneous, they will encounter similar 'spellings' in the books they read. As this suggests, a model of standard and non-standard spelling cannot adequately account for the occurrence of what many would label 'misspellings' or 'errors' in children's writing at school.
[4] Although, as Sebba acknowledges, spelling is not always correct: 'since it is not a contradiction in terms to talk of deviant or unusual spellings' (Sebba 2007:1).

codified forms, it can still accurately mean what they intend. Finally, whether or not respellings are 'standard' is less important in this chapter than whether they draw on orthographic principles and reflect usages in other texts, as well as the social meanings conveyed.

This chapter therefore uses the less evaluative terms *spelling variants* and *respellings*, to refer neutrally to spelling variants. These terms are also useful in encompassing BBC, UNESCO and IOU, regardless as to whether people consider them standard or correct, and instead focuses on their potential for meaning-making.

5.2.2 Sociocultural model of orthography

The theoretical framework guiding investigation of texted spellings in this chapter is the sociocultural model of orthography outlined and elaborated on by Sebba (2007), and which also underlies, if implicitly, work by researchers such as Jaffe (2000), Androutsopoulous (2000) and Preston (2000). The significance of the model for present purposes lies in the fact that it shows how choices made in spelling are both *principled* and *meaningful*. Other discussions of orthography offer descriptive accounts for pedagogical or remedial purposes which focus on sound-spelling correspondence (Carney 1994) and spelling difficulties (Mitton 1997), or historical accounts which explore standardisation and what spellings reveal about earlier pronunciations (Leith 1983; Scragg 1974). [1] Sebba's (2007) sociocultural model, however, explores spelling as an integral part of particular societies' reading and writing practices which cannot be investigated in isolation from the literacy practices it supports, themselves embedded in wider social and cultural practices (Street 1984).[2] Like literacy, orthography can therefore be seen as a *social practice*, which Sebba (2007) defines as 'widespread and recurrent activity which involves members of a community [in this case, literate with a common, standardised orthography] in making meaningful choices, albeit from a constrained set of possibilities' (Sebba 2007: 31). While

[1] See Sebba (2007: 11-12) and Carney (1994: 1-4) for other overviews of the treatment of orthography by linguists and other researchers.

[2] The extent to which spelling is embedded in ideological notions of literacy is illustrated in clashes in literary practices between missionaries and indigenous linguists when devising orthographies for previously unwritten Mexican languages (Barros 1995). The missionaries, from societies with solitary and silent book-reading practices, favoured alphabetic Spanish orthography, while indigenous linguists argued for 'mural texts' or 'wall-newspapers' and for a phonemic script, reflecting their focus on collectivity and orality (Barros 1995: 282). Different orthographies supported different literacy practices, emerging from different ideologies (Sebba 2007: 24).

the ‘set of possibilities’ regarding the form which spelling variants take is ‘constrained’ because spelling tends to be highly standardised, there is scope for ‘variation and deviation’ (in Sebba’s words) and this possibility of variation makes it feasible to call orthography a social practice, assuming that the *choice* of spelling inevitably carries social meaning. Decisions by individuals or groups to spell according to prescribed usage or to depart from prescription by using respellings to achieve particular effects are thus *meaningful*.

At the same time, variation from conventional norms must follow language-specific orthographic principles to remain recognisable to the community and thus be effective (Sebba 2007: 32). Unconventional spelling in Willans’s (1953) *Down with the skool* (Sebba 2007: 31) makes a rebellious yet humorous point against schooling: yet, to do so, the choice of <skool> conforms to English sound-spelling conventions. In contrast, <zguul>, while recognisable, is not a meaningful deviation from the spelling of *school* because it does not follow orthographic principles and appears, in Sebba’s (2007: 31) words, ‘odd or weird’. In other words, respellings are neither random nor freely chosen but *principled* (Sebba 2007: 33). Similarly, to ensure readability, respellings are as selective as transcriptions of spoken data are (Jaffe 2000: 501), as Androutsopolous (2000: 528) observes in his study of German fanzines. This paradoxical ‘sameness to’ yet ‘difference from’ the standard creates, suggests Jaffe (2000: 506-7), a perpetual tension in that non-standard spelling often represents a stance against standard orthography, or expression of contrasting identity, and yet must operate within principles or conventions established by the standard. Consequently, non-standard spelling is not only constrained by the standard but is often seen (although not by all researchers) as being negatively evaluated in comparison with its legitimate, authoritative counterpart (Jaffe 2000: 499).[1] Nonetheless, the restrictions within which respelling operates heightens its potential for meaning-making.

The implication of this is that respelling in texting will follow orthographic principles and therefore reflect or extend existing patterns of respelling. This suggestion is corroborated by Shortis (2007) who describes choices available to texters as *an extended orthographic palette*: the extension of traditional ‘binary’ choices of ‘standard’ and non-standard’ spellings to a range of uncodified and potentially meaningful options. Users’ choices as to which respellings they select from the extended palette, argues Shortis, is constrained by

[1] Stigma has, as mentioned in the introduction above, been attached to the abbreviation and phonetic spelling seen in text messaging.

contextual pressures such as their technoliteracy (technical and literary expertise), literacy identity (their beliefs as to how electronic communication should be handled), audience, purpose, context, and physical constraints. Shortis recognises that respellings are, as discussed above, also restricted by existing orthographic principles and so conventions are not new; however, he points out that the widespread use of texting and electronic communication have brought such practices to the public fore (Shortis 2007: 15). The popularisation of respellings is documented by the number of text messages written[1] and the level of engagement produced, while the suggestion that spelling practices have been legitimised can be seen in highbrow public debate surrounding Txt. Although, as suggested in the introduction, many commentators deplore it, others laud its merits.[2]

The present study of respelling extends Shortis's work, which is based on limited data, by providing over 11,000 text messages from which respellings are extracted. The following description of respelt forms and conventions, categorised into variants designed to reflect spoken forms and those occurring independently of speech (as in Weber 1986), provides a point of comparison against which respellings in CorTxt can be compared.

5.3 Patterns of Spelling Variation

5.3.1 Reflecting Spoken Language through Spelling

5.3.1.1 Spelling system

Orthographies using an alphabetic system are phonetic not in the sense of accurately reflecting how language sounds but to the extent that there are, to varying degrees across languages, correspondences between graphemes and phonemes: in English, for example, <v> is virtually always pronounced /v/; <igh> is usually /aɪ/. One-to-one correspondences in English are, however, few. '[H]igh on the hit-list of all spelling reformers' (Carney 1994: 229) are the varied pronunciations of <gh> (*enough*, *though*, *ghost*, etc), as well as the graphic representations of /i:/ as <ea>, <ee>, or <e> (in *feature*, *feeble*, *frequent*) among other spellings (Carney 1994: 155-164). Nonetheless, English speakers recognise a link between graphemes and phonemes, and use them to aid spelling and pronunciation (Carney 1994; Kress 2000). Taking into account that much variation disappears once surrounding

[1] See Chapter 1: Introduction.
[2] Crystal (2008a), for example, weighs in on the public debate and concludes of texting, 'I am fascinated by it, for it is the latest manifestation of the human ability to be linguistically creative and to adapt language to suit the demands of diverse settings' (p175).

morphemes are considered (Carney 1994: 15-18), it is possible (as Carney does) to talk of 'divergences' from the following 'default' spellings of phonemes.

Figure 5.1 Default spellings of selected phonemes (from Carney 1994)[1]

- /ɒ/ spelt by default as <o> (92% TF; 95% LF)[2] as in *dog* or *strong*, although <a> is particularly common after /w/ (*want*, *quality*);
- /ʊ/ spelt by default as <u> (32% TF; 54% LF) in *push* and *pull*, or <oo> (64% TF; 35% LF) in *good*, *foot*, *look* (the figures show that <oo> occurs in more high frequency words);
- /əʊ/ spelt by default as <o>, <oe> or <o..e> (TF 75%; LF 85%) in *photo*, *don't*, *mope*, *potatoes*; but with <ow>, the second most common (TF 18%; LF 7%), particularly in more frequent words and 'almost exclusively' in morpheme final position (*mow*, *mellow*, *shown*);
- /aɪ/ spelt by default as <i>, <i..e> and final <-ie>, <-y> (TF 80%; LF 82%) in *time*, *reptile*, *die*, *dry*, *rhyme*, *hyphen*) (only the spelling <igh> is unique to /aɪ/, as Carney (1994:155) notes, with <ight> in monosyllabic words most frequent, over <ite>);
- /ə/ diverges more than other phonemes but is most commonly spelt <a> (TF 35%; LF 30%) in *away*, *media*, *capable*, *admire*, *elephant* compared to, for example, er (TF 15%; LF 12%) in *knocker*, *better*;
- /f/ spelt by default as <f> (TF 84%; LF 77%) in compared to <ph> (TF 11%; LF 8%), in words borrowed from Greek;
- /z/ spelt by default as /s/ (TF ; LF) especially in high frequency words (*please*, *always*, *present*), but not in initial position (*zero*, *zip*).

Associations which these default spellings suggest between graphemes and phonemes have implications for explaining 'phonetic' respellings such as *luv* or *woz*. Although regional,

[1] Carney's (1994) survey of spelling, which maps phonemes to graphemes (and vice versa) is based on a corpus including five word-frequency lists, including the Brown corpus (Kucera and Francis, 1967) and the LOB corpus (Johansson and Hofland, 1989) (thus covering American and British English).

[2] LF = lexical frequency (frequency of *types* with a particular spelling) and TF = text frequency (frequency of *tokens* with a particular spelling) to account for the fact that spelling frequencies depend on the frequency of the words which use that spelling. Only lexical words are included: Carney's frequency counts exclude function words (see Carney 1994:109-112).

social or individual alteration in pronunciation is unaccounted for in the current English spelling system,[1] phonetic respellings used in speech transcriptions, novels and across informal domains such as texting to indicate informality or regional accent include *standard contractions* such as *can't* or *I'm*[2] and *colloquial contractions*.

5.3.1.2 Colloquial contractions

Colloquial contractions are reduced or contracted written forms such as *ya* or *gonna* which reflect informal pronunciation and thus socially-charged speech, as well as emotions of anger or affection: in other words, they establish tone (Weber 1986: 420; Androutsopoulous 2000: 521). As examples show, contractions can involve word shortenings (*you* to *ya*) or combine two words (*going to* to *gonna*). Although, as Crystal (2003: 275) points out, many now have established written forms, they differ from standard contractions in not being formally recognised in dictionaries or schools, not using apostrophes, involving at times just one word, and being largely limited to written discourse representing direct speech such as below.

Figure 5.2 Written discourse exploiting colloquial contractions in representing speech (from Weber 1986: 421 and Crystal 2003: 275/416)

- Comic strips, jokes and humour
- Journalism, including interviews representing direct quotation,
- Dialogue in prose fiction, plays and other forms of literature, as a basis for characterisation, where idiosyncrasy and regional background are reflected in distinctive spelling
- Marketing

[1] In the current system, *the* is spelt <the> both before consonant (pronounced /ðə/) or vowel sounds (pronounced /ðɪ/); *and* is spelt <and> regardless of differences between its weak and strong forms; *bath* is spelt the same throughout the UK, although some pronounce the vowel /æ/ and others /a:/.Representation of regional pronunciations, however, was once commonplace (Leith 1983; Baron 2000; Crystal 2003). Although spelling was standardised through printing and mass education, regionally distinct spellings surviving into the 1700s remain apparent in irregular forms such as *fox* and *vixen*, in which /f/ represents the East Midlands dialect of Middle English, and /v/ the Southern spelling (Baron 2000: 106).

[2] As standard forms, these are not explored here. However, they raise relevant points: firstly, in not representing the full range of variation in informal speech, they highlight the selectivity involved in any representation of speech; secondly, they highlight its arbitrariness: many Elizabethan contractions are used only in old literary language (Partridge 1964): *tis*, *twas*, *ne'er*. Most significantly, they show how the use of spelling exists as a widespread and established means for indicating informality as well as evoking spoken forms. In terms of the form the standard contractions take, it is also of interest to note the use of an apostrophe to indicate the missing letter; as we shall see, this is not found either in colloquial contractions or, as we shall see, the standard contractions found in many text messages

- Popular culture including pop and rock songs: ‘Gotta Serve Somebody’ (Bob Dylan); ‘I wanna hold your hand’ (Beatles); ‘Never Gonna Happen’ (Lily Allen).

Limitations of Weber’s (1986) investigation of colloquial contractions, in respect to the current investigation, is that the material she uses (comic strips, magazine interviews and anecdotes, as well as casual observations of other written material) features American English and sources date from over twenty years ago; however, it appears that even idiosyncratic forms represent patterns reflecting those in present-day British English.

Table 5.1 Colloquial contractions (following Weber 1986)[1]

progressive verbs	-ing → in’	also used for other words ending in ‘ing’—*something*, *building*.
	going to → gonna	*goin’ to* rarely used to form the present continuous
auxiliary verbs	does → -’s did → -’d do → -da	‘What’s that mean?’ ‘How’d it go, handsome?’ ‘Whadda ya call a …?’[2]
	have → a	’ve shortened to: *coulda*, *shoulda*, *woulda*, *mighta*
catenative verbs	got to → gotta have to→ hafta want to → wanna	*a* represents unstressed schwa; while the deletion of the *t*s in *wanna* reflects a regular informal pronunciation rule (that is, the elision of /t/ between two vowels following the sound /n/ as in *winter* or without /n/ as in *beauty*, represented by Weber as [VntV] – [VnV])
personal pronouns	him → ‘im her → ‘er them → ‘em	reflect the regular deletion of /h/ preceding unstressed vowels in rapid informal speech.
	you → ya, yuh, y’	the most frequent is *ya*, *y’* appears as the subject in conversational expressions such as *y’know* and *y’see*. Crystal (2003) lists *ya* among other forms which represent elision in colloquial speech including *dunno*, *yeah*, *yup*,

[1] In keeping with our recognition that spelling follows grammar as well as sound, they are grouped below according to grammatical features (as in Weber), rather than the phonemes they represent.
[2] Another example which could be added to the category of auxiliary verbs is the colloquial contraction *innit* (*isn’t it*) as noted by Crystal (2003:275) which represents the spoken elision in a common discourse marker used similarly to the French ‘ne c’est pas’ in certain regions of the UK, including the south-west of England.

		c'mon, *trific* or *triffic*, *nuf* or *nuff*, *em*, and *innit*—the last sometimes used as a rhetorical device by humorous writers as a light-hearted way of making a point
	you → -cha	reflects assimilation of /j/ after /t/ in *gotcha*, *doncha*, *watcha*. Unlike other forms, *-cha* omits the letter that represents the identity of the pronoun in standard spelling and variants. Crystal (2003) describes *gotcha* as an accurate representation of a colloquial assimilation, and comments on its effective use as a headline in *The Sun* when the British navy sank the General Belgrano during the Falklands War.
others	of → a	*lotta* (lot of); *coupla* (*couple of*); *buncha* (*bunch of*); *sorta* (*sort of*); *kinda* (*kind of*); *outa* (*out of*).
	give me → gimme let me → lemme	object *me* added to verbs, final consonants of the verbs are deleted and the doubling of the consonant is due to the rule which indicates that the preceding vowel is lax.
	deletion of d after /l/ or /n/	*ol'*; *an'*
	aphaeresis of because and about	'*cause*; '*bout*
	unique forms	*c'mon*; *li'l*

Although uncodified, the colloquial contractions follow recognised patterns. Use of <a> in representing schwa, to take Weber's (1986: 424) example, is now a strong pattern in English[1] conformed to in *gonna*, *gotcha*, *kinda* and in the advertising slogan: *Drinka Pinta Milka Day* (Carney 1994: 447).[2] Another way in which colloquial contractions conform to familiar spelling patterns can be seen in *gotta*, where the consonant doubling conventionally marks the preceding vowel as short.[3] As described later, these patterns are reflected in texting.

[1] <a> replaced <e>, which commonly represented schwa through the Middle and Early Modern English periods, due to borrowings from classical languages such as *agenda*, *bacteria*, *parka*, *sofa*, and *yoga*.
[2] As Crystal points out, the letter *r* is also frequently used in such stylistic effects: *larf*, *lorra*, *luvverly*, *har har*, *shurrup*.
[3] A final point made by Weber (1986: 424) of relevance to the present study is the lack of apostrophes in these colloquial contractions (in contrast to the standard contractions outlined above and see footnote 2, p.104),

Other attempts to imitate informal speech include regional respellings and what I call *colloquial respellings*. Regional respellings, described variously as regiolectal respellings (Androutsopoulous 2000: 521) and literary dialect[1] (Ives 1971 in Weber 1986: 420), are representations of regional varieties through spelling. As Weber (1986: 420) explains, there are 'few definite rules' but 'recurrent strategies' she notes include omitted and shifting sounds in *some'ers* for *somewheres* and *mout'* for *mouth*; and the use of <h> to represent 'r-lessness' as in *caht* or describe vowel sounds: *mah* for *my*. Other respellings, however, are perhaps better known as *colloquial respellings*, in that, although non-standard pronunciation is suggested, it does not have regional associations and nor is the word shortened (as with colloquial contractions). Examples include *yeah* or *nope*, lengthier than *yes* and *no* and used widely in informal speech.[2] The distinction between regiolectal and colloquial respellings, however, is unclear. Omission of <h>, for example, categorised by Weber (1986) as a colloquial contraction, is also associated with estuary or cockney English.[3] Nonetheless, bearing this caveat in mind, it is useful to distinguish respellings in Txt suggesting regional varieties from those suggesting features found generally in conversational English. The role of all three in indicating informality and identity is explored below.

5.3.1.3 What attempts to reflect speech mean

Colloquial contractions and respellings 'mean' by reflecting people's spoken language, thus evoking aspects of identity (often derogatorily) and informality through writing: that is, as argued by Androutsopoulous (2000), their function is *indexical*.[4] In other words, although respellings are chosen to reflect certain pronunciations or alterations in speech, sounds are not the ultimate purpose as much as what this means for readers regarding register and/or speaker identity (Jaffe and Walton 2000). This is evidenced by the use of 'eye dialect'

---

although it is retained in other two-word contractions such as *What'd, y'know* and perhaps certain one-word shortenings such as *ol'*.

[1] Due to their use in literary fiction.

[2] The distinction between colloquial contractions and colloquial respellings is not made by Weber (1986), nor by Sebba (2007) or Shortis (2007), but is here deemed a useful one.

[3] Another example is *gawd*, a word which, according to Crystal (2003), suggests a distinct regional or class pronunciation, but could also be described as an informal (regionless) one: 'Oh Gawd', Crystal (2003: 275) notes, 'is less serious than 'O God' and would, one imagines, never appear in a liturgical manual'.

[4] Indexical refers to 'features of speech or writing which reveal the personal (biological, psychological or social) characteristics of a language user, as in voice quality or handwriting' (Crystal 2008b:241)

(such as *sez* for *says*)[1] which does not correspond to particular accents (Preston 2000: 615) but instead suggests lack of education (Ives 1971), in that phonetic spellings are perceived as mistakes made by people who write what they hear.[2] Portrayals of others, found in novel dialogue and in speech transcripts (Preston 1982; 2000), thereby characterise or ridicule people of low educational, economic or social status, and/or those with strong regional accents.

Self-representation, however, is used overwhelmingly in positively affirming personal identity: respelled names such as *Jaqui*, *Mikel* or *Nikki*, for example, represent 'personalized interpretation of a generic linguistic form' (Jaffe and Walton 2000: 507; see Carney 1994: 443-466), as well as affirming group identity, in cases where people involved belong to the same community and share values and identities (Jaffe 2000: 508). Caribbean poetry, for example, often uses non-standard spelling reflecting pronunciation of the poets' creole.

Figure 5.3 Respellings in the creole poem *Dread Beat an Blood* (Johnson 1975, cited in Crystal 2003: 348)

wi feel bad
wi look sad
wi smoke weed
an if yu eye sharp,
read de vialence inna wi eye;
wi goin smash de sky wid wi bad bad blood

Androutsopoulos (2000: 528), to take another example, describes colloquial spelling variation in German underground music magazines as establishing informal, intimate arenas between writers and readers. The difference between this use of colloquial respellings and the derogatory uses above is that they do not contrast with conventional surrounding text (often the language of educated or higher-standing persons) but occur throughout the fanzines in a 'regular, patterned manner'. Fanzines and Caribbean poetry show that, while colloquial and regional respellings used to represent others contast with

[1] The difference between eye dialect (Bowdre 1982) and colloquialisms is that, although eye dialect involves phonetic spelling, it is used to capture more regularly the 'standard' pronunciation rather than indicate non-standard regional or informal pronunciation: it is 'the spelling of a familiar word in a non-standard form, while maintaining the standard pronunciation' (Weber 1986: 418) or, in Sebba's (2007:34) words, '[u]sing sound-symbol correspondences which are conventional for the language, but are the wrong ones for the particular word': eg, <thort> for <thought>'. Eye dialect is returned to below.

[2] Thus explaining the impact of scrawling *skool* on a school desk, as Sebba (2007: 31) points out.

the conventional language of educated persons and can be derogatory, colloquial respellings can also affirm group boundaries, evoking shared identity and intimate, personal and informal relationships. This *performance* of spoken informality, as described here, also occurs in texting.

5.3.2 Visual representations of meaning

5.3.2.1 Shape of words

The argument that writing cannot be seen simply as speech written down is widely accepted among linguists (Baron 2000; Chafe 1982; Halliday 1985), yet there is a tendency, among spelling reformists and the public, to conflate spelling as little more than representations of sounds. However, although it is possible to link English sounds and symbols despite the lack of one-to-one correspondence, as Carney (1994) shows in his systematic survey, factors other than sound shape the spelling system. These include, Leith (1983) explains, *meaning* (hence the identical spelling of /ɛɪ/ and /æ/ in *nation* and *national*) and *grammar* (regular plural forms in English are marked with <s>, rather than <s> or <z>, following their pronunciation). As Kress (2000) argues, spellings can represent direct transformations from the external world (or writers' inner worlds) onto paper. For children, the form of overall words is also important,[1] which Kress (2000: 190-191) illustrates with a child's adoption of a beautiful, careful style in appropriate contexts (such as when writing poems) quite different from her scrawled handwriting elsewhere. Respellings cannot therefore be accounted for simply as attempts to reflect speech, but instead arise from the unconventional shape of words (Weber 1986; Jaffe 2000: 500).

5.3.2.2 Phonetic spellings

One use of phonetic spelling is the aforementioned *eye dialect.* As explained in the previous section, this involves substituting letters in irregular standard spellings for those which more regularly correspond to the particular sound, and thus found with common words more likely to be irregularly spelt: *enuf*, *frum*, *cum*, *tu*, *yu*, *ur* (Weber 1986), as well

[1] The same is true for adult spellers: 'How often do you look at a word you have just written, and say to yourself 'That doesn't look right?' asks Kress (2000: 173), '[i]f, as often happens, I type *langauge* on my computer …, it is the appearance of language in a visual/spatial dimension that has gone wrong, and it is its visual appearance that tells me that something is awry' (Kress 2000: 167).

as *wot*, signalling uneducated speakers and *sez*, which reflects the pronunciation of *say* in spontaneous speech (Crystal 2003: 275).[1]

Phonetic spelling in other contexts creates modern, dynamic, eye-catching effects, solely through the visual effect of its deviance from expected spellings.[2] Many brandnames involve what Androutsopoulous (2000) calls 'grapheme substitutions': the substitution of one letter for another (<z> for <s>; <k> for <c>) such as *Beanz Meanz Heinz* (Crystal 2003: 275). Pound (1925 cited in Carney 1994: 447) identified a 'Kraze for K' in the early 1900s, which Carney suggests has not disappeared: *KitKat*, *Kwik Save*, *Rice Krispies*, *Kleenex*. Many are shorter than corresponding conventional variants, due to a tendency to reduce vowel digraphs and consonant clusters: studies of trademark respellings by Jacobson (1996) and Praninskas (1968) include *Protex* (<x> for <ks>, as in the fanzines); *Tru-Blu*; *Fre-Flo*, *Mildu*; and double-consonant reduction: *Hot-Stuf*, *Chil-Gard*; *Fly-Kil* (Carney 1994:447). Use of letter and number homophones as in 'rebus' puzzles is noted by Carney (1994: 448), who cites *Spud-U-Like*, *U2*, *INXS* and *IOU*[3]. Another well-established example in advertising is the reduction of *ight* to *ite* (Weber 1986, Moon 2008) as in *CocaCola Lite.*[4] As Sebba (2007: 37-8) suggests, homophones were entrenched in graffiti and other non-mainstream domains before texting and computer-mediated communication.

[1] Although eye dialect is used alongside colloquial contractions and respellings to evoke spoken forms in dialect speech, it is included alongside abbreviation as independent from speech by Weber (1986) in that it marks no phonemic alterations but makes meaning (informality and ignorance of, or disrespect for, conventions) through the visual impact of a form gone 'awry' (Kress 2000: 167) which contrasts (like colloquial contractions) with how others' speech is represented (Weber 1986: 418).

[2] Weber's distinction between the use of phonetic spelling as eye dialect and other uses of phonetic spellings, which she labels brandnames, is reflected here, but will not be used in the following investigation of text messaging, on the basis that the use of phonetic spelling in the text messages may not be categorised neatly into either category.

[3] The recognition here of, and distinction between, letter and number homophones and grapheme substitutions follows Androutsopoulous (2000).

[4] *Lite*, initially referring to food or drink with fewer calories, now also 'deprecate[s] something as an over-simplified or weakened form of an original, more challenging or demanding, activity, subject, and so on' (Moon 2008: 18), as in Stallabrass's (1990) *High Art Lite: British Life in the 1990s*; the entertainment guide *London Lite* (www.thelondonlite.co.uk); and *Nintendo DS Lite*, a lightweight version of Nintendo DS. As these examples show, *lite* is used postnominallly (unlike *light*). *Lite* is used in modern, informal contexts concerning entertainment, advertising and technology: BoE reveals *grunge lite*, *Bud lite*, and *Internet lite*, as well as *passion-lite* as a description of contemporary Britain, and also that *lite* occurs overwhelmingly in the American ephemera subcorpus (accessed March 2008). These associations now extend to other respellings: *nite* is used to describe cutting-edge technology needed to deal with 'the night' or events taking place at night-time, rather than the period of darkness itself. A Google search brings up a *Tandoori Nite* Indian restaurant, a *Boogie-Nite Mobile discos and Wedding DJ*, and the *Nite Hunter*, the latest in lamping equipment (with, incidentally, two new designs: *Pro-Lite* and *Supa-Lite*).

The motivation behind phonetic respellings can be to engage with or challenge readers. In advertising, the aim is to attract consumers (Davies 1987; Androutsopoulous 2000: 515), which it achieves not only by creating, as Jaquith (1976) suggests, *modern* images of products through respelling, but because distinct, unexpected spellings contrast with surrounding text, disrupt readers' scanning and so catch their eye (Jaffe 2000: 510). In other words, its impact lies in its *deviance* from expected norms, rather than because of its phonetic representation.[1] Another example of this is its role in underground music magazines explored by Androutsopoulous (2000). Use of <x> to represent <ks>, <cks> and <gs> in English and German words such as *punx*, *thanx*, *sonx* (*songs*), *lyrix* and *demnaxt* (*demnachst* = *soon*) position texts in the subculture and mark them as radical, tough and original (Androutopoulous 2000: 527-528). However, in this case, phonetic spelling not only affirms group identity but creates deviance by marking the subculture as distinct or opposed to mainstream ideology:[2] the two functions of convergence and opposition operate simultaneously. As described in this investigation of texting, texters adopting these forms of respellings similarly affirm membership of texting networks by performing *deviance* from expected norms.

5.3.2.3 Technological constraints and abbreviations

Variation in spelling dating from handwritten (heavily abbreviated) manuscripts throughout medieval Europe (Bradley 1919: 4) is encouraged not only by functions such as information storage which written texts fulfil, but by the inevitable impact of the technology involved in their production (Haas 1996: x-xi in Baron 1998: 136). Although early printing, for instance, played a leading role in standardising spelling and punctuation (Baron 2000: 99),[3] Elizabethan printers, like scriveners before them who were paid by the inch and thus invented lengthier versions of words (*pauvre* for *povre*, for example) (Scragg 1974: 52 in Baron 2000: 98), also varied their spellings due to space constraints and cost. The need to justify the right-hand margin of printed pages and to ensure text fit the page,

[1] Parallels can be found in the description of everyday creativity in speech, whereby expected patterns are disrupted through, for example, the manipulation of idiom as seen in 'It's a question of how many birds we can kill with this particular stone' (Carter, 2004) which makes its impact through the distortion of *form* as much as the meaning of the phrase.

[2] A similar point could be made of graffiti (see Blume 1985). Similar observations have been made of teenagers' texting practices, in that parents and other adults are often unable to penetrate the abbreviated language 'code' (e.g. Ling and Yttri 2002); while, in a broader sense, the popularisation and legitimisation of respellings through texting offers a 'challenge to official educational discourses of literacy' (Shortis 2007a: 11).

[3] It was under Caxton's influence, for example, that the London standard became the printing norm (Crystal 2003: 56)

for example, prompted the use of an elaborate system of abbreviation and variable spelling: *busy* could be spelt *busie*, *here* as *heere* (Baron 2000); *on* as *onn*, *hon* and *ho*; and *say* as *sai*, *say*, *saie*, and *sei* (Bennett and Smithers 1968 in Baron 2000: 104). Strategies employed by printers included:

Figure 5.4 Elizabethan printing strategies (Baron 1998: 98)

- Intentionally using (or avoiding) abbreviations
- Crowding letters together or spreading them out
- Increasing or decreasing gaps between words
- Choosing longer or shorter spellings of words [according to Crystal (2003: 56), printers' strategies led to great variation in the final *e* of many words]
- Adding or deleting words
- Substituting words for phrases, or vice versa.

To an extent, these methods are not dissimilar to those used in texting.[1] A more recent example is the telegram, the cost of which was calculated according to the number of words it contained, encouraging highly abbreviated styles (Crystal 2003: 425). However, abbreviated forms occur in greater number across a wider range of current texts than might be thought. Although people generally tend to place the number of English abbreviations in the low thousands, the *Acronyms, Initialisms & Abbreviations Dictionary* listed over 586,000 entries in its 2001 edition (Crystal 2003: 121), and the extent to which certain abbreviations, such as AIDS, laser, UNESCO, yuppie, and BAA, have entered our lexicon is illustrated by ignorance as to what the full terms of some are.

Other reasons behind abbreviation lie in the value placed on linguistic economy or attempts to achieve concise styles,[2] the desire to convey social identity or be part of the social group to which the abbreviation belongs (Crystal 2003: 120). ROM, RAM, DOS and WYSIWYG, for example, belong to computer scientists, while abbreviations abound in science,

[1] At least the methods which involve shortening the text.

[2] Crystal (2003: 120), for example, documents the occurrence of such abbreviations in the nineteenth century, quoting the New York *Evening Tatler* of 1839 which comments on 'the initial language … a species of spoken short-hand, which is getting into very general use among loafers and gentlemen of the fancy, besides Editors, to whom it saves much trouble in writing'. The abbreviations in question here included *OK* (*all correct*), *PDQ* (*pretty damn quick*); *GT* (*gone to Texas*) and *LL* (*liver loafers*). Crystal (2003:120) concludes that the fashionable use for abbreviations 'comes and goes in waves, though it is never totally absent'. This reinforces the argument that respellings apparent in, for example, studies into texting presented below reflect a long tradition of abbreviation and other spelling variation.

technology and specialist fields such as cricket, drug trafficking, the armed forces, and the media. As argued later in this chapter, it can be questioned whether abbreviation in texting is chiefly motivated by any need to be brief but instead contributes to texting identity through *performances* of brevity, informality and deviance.

Regardless of the technology, domain or motivations involved, however, abbreviations tend to follow certain formal patterns. Defined as shortenings of words which do not affect pronunciation (Weber 1986: 415), abbreviations can be categorised into the following types.

Figure 5.5 Types of abbreviation (Crystal 2003: 120)

- **Initialisms** or **alphabetisms**—'Items which are spoken as individual letters, such as BBC, DJ, MP, EEC'. The majority of abbreviations fall into this category, including those which also take letters from the middle of the words concerned: PhD, GHQ, TV.
- **Acronyms**—'Initialisms which are pronounced as single words, such as NATO, laser, and UNESCO' and distinguished by a lack of full stops, although, notes Crystal, some linguists do not recognise a distinction between these and initialisms.
- **Clipping**—'A part of the word which serves for the whole, such as ad and phone'; the commoner type retains the first part of the word (*demo*, *exam*, *pub*) while with others the last part is kept (*bus*, *plane*), a middle part (*fridge*, *flu*) or two parts: *maths*, *gents* and *specs*. Others show adaptation: *fries* (fried potatoes), and *Bill* (William).
- **Blends**—'A word which is made out of the shortened forms of two other words', such as *brunch* (breakfast and lunch), *heliport*, *smog*, *Eurovision*. Blends can often be identified in scientific terms (*bionic*), brand names and fashionable neologisms (130).
- **Awkward cases**—some that can be used as either initialisms or acronyms such as UFO, or which mix the two devices: CDROM. Some can form part of a larger word, using prefixes: ex-JP; pro-BBC; or are used only in writing: Mr, St)
- **Facetious forms**—TGIF, CMG (Call me God), KCMG (Kindly …), GCMG (God Calls Me God), AAAAAA (Assoc. for the Alleviation of Asinine Abbreviations and Absurd Acronyms).

These abbreviations often bear no relation to spoken sounds: *lb* or *ie*, suggests Weber (1986: 415), can almost be described as 'visual morphemes'.[1] In cases where written abbreviations correspond to spoken phrases, the spoken shortenings arise from the written abbreviation.[2] As we shall see, Txt abbreviations follow and extend these forms.

5.3.3 Summary

The main patterns of respelling, summarised below, include *colloquial contractions*, *colloquial respellings* and *regiolectal respellings* which, together with *eye dialect*, evoke spoken informality. Respellings which make meaning directly through visual representation include *abbreviation* and *visual morphemes*, as well as letter and number *homophones* and *grapheme substitutions*.[3]

Term	Definition	Source
Spelling Variation Dependent on Speech		
Colloquial contractions	'[C]ontracted or shortened forms in writing that reflect colloquial or rapid pronunciations'	Weber (1986: 220-1)
Colloquial respellings	Similar to colloquial contractions, these are attempts to reflect pronunciatons which do not shorten or contract words involved.	Term used in the current thesis.
Regiolectal respellings	Respellings which reflect regional pronunciations.	Androutsopoulous (2000: 521) and Weber (1986: 420)
Spelling Variation Independent of Speech		
Eye dialect, as a	'[T]he spelling of a familiar word in a	Weber (1986: 418)

[1] Visual morphemes is used by Bolinger (1946) to refer 'numerals and other conventional signs' such as &, % or $

[2] Note, out of interest, that the abbreviation *www*, when spoken, is in fact much longer than *World Wide Web* – yet the latter is (I believe) rarely heard. Incidentally, according to Weber (1986: 416), acronyms (which would tend to be shorter when spoken) are not preferred in English over initialims. The Spanish, to take an example from my own experience, seem to favour acronyms more: pronouncing, for example, the IRA and UFOs as acronyms rather than initialisms. However, it seems that English acronyms have recently been increasing (Weber 1986; Leech et al 2009).

[3] Some categories identified in the literature do not prove relevant to my corpus: interlingual spellings ('phonetic spellings of loanwords according to … orthographic rules' Androutsopoulous, 2000:521) and language external symbolism (where writers make 'symbolic use of othographic conventions from other languages' Sebba 2007:39-41).

kind of phonetic spelling	non-standard form, while maintaining the standard pronunciation … [t]hey serve to hint that the overall tone of speech should be interpreted as different from the tone of conventional speech, usually in the direction of rustic and uneducated, often jocular and even grotesque … [E]ven though they themselves do not represent a particular pronunciation, they often appear in texts next to dialect spellings that do'	
Other phonetic spellings	Letter and number homophones and grapheme substitutions used in brandnames and subcultural material.	Androutsopoulous (2000); Weber (1986).
Abbreviation	'the shortening of the written form of a word or words without concomitant shortening of pronunciation'	Weber (1986: 415) see Crystal (2003)
Visual morpheme	'numerals and other conventional signs in our written language' but also where the letters 'do not relate directly to sounds in the full spoken word'	Bolinger (1946) and see Weber (1986)

5.4 Respellings in texting

5.4.1 Previous studies of texting

Respellings in CorTxt must be viewed against the backdrop of attention paid to text spelling by the media and in existing research. Popular conceptions that texting is highly abbreviated is reinforced by Crystal (2008a) who draws on online texting dictionaries as well as the reported practice of family and friends: *THX* (*thanks*); *MSG* (*message*); *XLNT* (*excellent*); *B4* (*before*); *C U l8r* (*see you later*); *c%l* (*cool*).[1] Although, as Crystal acknowledges, online lists do not necessarily reflect real usage, they find some support in the following sociological and linguistic studies into texting which draw, to varying extents, on attested data and focus primarily on respelling.

[1] As shown by the examples given, Crystal's (2008a) list of abbreviations inlcudes a large number of phrases reduced to word-initial letters, also including: *BTDT* (*been there, done that*), *OMG* (*oh my god*), and *yysw* (*yeah, yeah, sure, whatever*).

Table 5.2 Respellings in the existing text messaging literature

Study	**Corpus specifications**	**Respellings**
Doring (2002)	1000 messages in German	Doring notes few acronyms or abbreviations in her corpus.
Hard af Segersteg 2002	1152 messages in Swedish[1]	The corpus comprised speech-like spelling, omitted punctuation and spelling, and both conventional and unconventional abbreviations (as well as grammatical reduction, particularly of subject pronoun). Apparently unique to text messaging (according to Hard af Segersteg) were consonant writing (where vowels are omitted) and typos caused by predictive texting, as well as excessive space omission.
Kasesniemi and Rautianen 2002	Nearly 8000 messages in Finnish	The authors note a lack of traditional grammar or punctuation in their Finnish corpus of text messages, with words being shortened or run together and inflectional endings omitted, and they support the possibility of negative effects on children's (and especially boys') writing. However, Kasesniemi and Rautianen (2002: 184) also highlight the 'variety of repertoires of SMS expression and teens' creativity in adapting their messages to circumstances', with messages to different interlocutors varying in their style, content, form, and thus the degree to which words are abbreviated or standard. The authors also note diachronic changes, with earlier text messages reading more like postcards or telegrams but, although boys were still writing plainly and briefly, girls' texts were increasingly full of (compressed) expressions, references and suggestions, social softening and 'padding material'.
Grinter and Eldridge (2001; 2003)	477 messages written in English	The authors identify 166 unique short forms grouped into 4 categories: 1. Standard abbreviations; 2. Traditional or ad-hoc abbreviations; 3. Dropping of single letters; 4. Phonetic spellings.

[1] Of relevance to the current thesis, is the fact that the messages in this and other studies were not written in English. To take Hard af Segersteg's corpus as an example, perhaps particular to Swedish was the use of foreign words (that is, largely English ones) and compound splitting.

		They note that the two most common methods for shortening words involved the use of letters, symbols or numbers to make appropriate sounds (37%) and the use of abbreviations (47%). However, they also note note that the teenagers in their study were shortening simple everyday words such as *tomorrow*, *weekend*, *school*, *Internet*, *lessons*, *homework*, rather than the complex acronyms suggested in dictionaries and in contrast to the earlier trend noted by Weber (see above).
Thurlow and Brown 2003	544 messages in English	In terms of 'new linguistic forms', Thurlow and Brown identify 1) shortenings, contractions, G-clippings and other clippings; 2) acronyms and intialisms; 3) letter/number homophones; and 6) accent stylisations. However, despite the fact that 82% of participants claimed to use abbreviations, Thurlow and Brown (2003: 6) find only 1401 examples of abbreviation in his data (around 3 per message, or less than 20% of message content); just 509 typographic symbols, mainly kisses and duplicated exclamation or question marks and only 39 emoticons; and few letter or number homophones (73). Instead, there are frequent 'accent stylizations or phonological approximations' such as *novern* for *northern*; 'onomatopoeic, exclamatory spellings' such as *haha!* and *WOOHOO!*; as well as words such as *wakey wakey* and *wotcha*, designed to add prosodic impact' (Thurlow and Brown 2003: 6-7)—suggesting the influence of interactive demands, rather than of technological constraints. Thurlow and Brown (2003: 7) also note what they considers a high use of apostrophes (192, or one in every three messages): 'it does not appear that the apostrophe is quite dead!'.
Oksman and Turtianen (2004)	Nearly 8000 messages in Finnish	They report language play and newly-coined phrases, as well as the spoken expressions, or 'written orality', also noted in Italian text messaging by Fortunati (2001:314)
Faulkner and Culwin (2005)	337 messages in English	Faulkner and Culwin (2005), whose study was unusual in involving participants in their twenties rather than teenagers, found little evidence of what they call 'textish', claiming that 196 of their 337 messages contained no text-specific language and they pointed to the rise of predictive texting as a factor.

Despite evidence for features noted in the press, the above literature also suggests unconventional spelling is more limited than popularly assumed (Doring 2002; Grinter and Eldridge 2003; Thurlow and Brown 2003; Faulkner and Culwin 2005), accounted for by technological factors (predictive texting), interpersonal concerns, and individual style. Changing perceptions, even among teenagers, may be another factor: as mentioned in Chapter 2, Taylor and Harper (2002) find the value placed on text messages can depend on the language used, and that abbreviating to avoid sending more than one message can lower their value. Such studies support arguments regarding the over-exaggeration of abbreviation in public, non-academic minds; and highlights, to use Thurlow and Brown's term, the 'unremarkable' nature of Txt.[1]

Where abbreviations and other variant spellings are used, they appear from the literature to emerge as appropriate, even creative responses to features of the communicative situation: the combination of physical constraints and informal intimate functions (Hard af Segersteg 2002; Kasesniemi and Rautianen 2002; Thurlow and Brown 2003; Crystal 2008a). For Hard af Segersteg (2002), drawing on Grice's (1975) 'conversational maxims', Txt results from the principle of least effort. Informal and speech-like reductions in texting are therefore strategies towards reducing time, effort and space whilst expressing informality and personal intimacy (Hard af Segersteg 2002: 232). As well as 'brevity and speed', Thurlow and Brown (2003: 14-15) focus on the effect of texters' manipulation of the technology, identifying 'paralinguistic restitution' which replaces stress and intonation and 'phonological approximation' (such as capitalisation and multiple punctuation: *what???!!!*), which creates informality and playfulness (Thurlow and Brown 2003: 14-15). A final motivation noted in the literature is that of construing identity and defining group membership (Kasesniemi and Rautianen 2002; Ling and Yttri 2002; Thurlow and Brown 2003). As Thurlow and Brown note, such user-related motivations may occur simultaneously, and override the need for brevity and speed.

[1] Txt is unremarkable, according to Thurlow and Brown (2003: 15) in the sense that most non-standard forms are 'semantically recoverable' and reflect texters' awareness of the need to be intelligible: consonant writing, for example, recognises the information load or 'semantic value' carried by consonants while phonological representation is relatively easy to interpret; as well as being reflective of abbreviations used in other informal written texts where speed and time are issues (his example is the fridge door note); or where a certain informal tone is needed (in a magazine, or an advert). They are neither a feature of electronic language in general, nor of texting in particular, but of certain 'affordances' across language mediums. Finally, Thurlow and Brown argue that the impact of text messaging on language is unlikely to be unremarkable: 'new linguistic practices seldom spring from nowhere, neatly quashing pre-existing forms and conventions. Just as technologies do not replace each other, nor is it really possible to imagine communicative practices breaking completely, or that dramatically, with long-standing patterns of interaction and language use' (Thurlow and Brown 2003: 4).

The perception that abbreviations in texting are unique is also challenged (Shortis 2007a,b). Kesseler and Bergs (2003), for example, compare their corpus of valentine text messages with love letters written by lower-class 'fallen' girls at the London Foundling Hospital in the 1800s.[1] The two are similar in that both tend to be written on the spur of the moment and combine emotion with everyday purposes. Kesseler and Bergs (2003) found that, like texting, Victorian letters contained phonetic spellings such as *bcoz* (*because*) and *luv* (*love*); graphic symbolisations of kisses (*xx*) and roses, similar syntax, shared metaphors and figures, and grammar and spelling mistakes. They see in the text messages a simpler, more explicit style, and greater use of iconic and symbolic signs, but conclude that popular perceptions of an earlier, literate age are based unfairly on letters composed by highly-literate, public individuals who knew their writings would be retained. The grammar and spelling of lower-class writers, however, was no better than today's. Spelling variation in texting is therefore a continuation, or development, of earlier practices (Kesseler and Bergs 2003). Their findings are reflected by Shortis (2007a,b) who describes texting as extending, popularising and legitimising an existing yet devalued 'orthographic palette' of spelling options.

The following investigation of CorTxt supports the above findings through systematic and large-scale categorisation of respellings across 11,067 text messages.

5.4.2 Overview of CorTxt

The first observation when looking through CorTxt is the *absence* of unconventional spellings from many messages.

(5.2) Everybody had fun this evening. Miss you.

(5.3) Men like shorter ladies. Gaze up into his eyes.

(5.4) No, but you told me you were going, before you got drunk!

(5.5) Sorry NAME352. I will be able to get to you. See you in the morning.

[1] The number of valentine text messages was, at least at the time of the study, exceeding handwritten cards (Kesseler and Bergs 2003).

(5.6) Mystery solved! Just opened my email and he's sent me another batch! Isn't he a sweetie?

(5.7) I'm parked next to a MINI!!!! When are you coming in today do you think?

As Crystal (2008a: 20) points out, unabbreviated messages tend to be shorter than abbreviated ones. Other, often longer messages in CorTxt reveal several unconventional spellings: most of which seem familiar. In the example below, we see the contraction of *kind of* into the colloquial *kinda*; and a graphic representation of kisses (*xx*), as well as the less familiar clipping of *tomorrow* into *tomo* (noted also by Grinter and Eldridge 2003).

(5.8) Kinda. First one gets in at twelve! Aah. Speak tomo xx

In the next, we find the colloquial contractions *ditchin* and *comin*; and examples of phonetic spelling in *coz*, *u*, and *cumin*; the standard abbreviation of *Saturday*; and the less standard deletion of space between *thank* and *you*. [1]

(5.9) Thankyou for ditchin me i had been invited out but said no coz u were cumin and u said we would do something on the sat now i have nothing to do all weekend i am a billy no mates i really hate being single

The following investigation extends these initial observations through systematic categorisation of respellings across CorTxt, to show how they are shaped by features of the communicative situation: the need to save space, time and cost, and other technological constraints such as predictive texting; the desire to create an informal, speech-like tone; and attempts to mark group membership; and how users adopt and extend existing methods and forms of variation to do so. This suggests that texters have choices as to how they spell and that they are actively involved in meaning-making and identity formation: in other words, respelling plays a part in the performance of texted identity.

[1] A final point concerns the lack of capital letters used in mid-sentence positions, chiefly (within the above examples) *i* but also for proper nouns. This, however, can be put down to the difficulties involved in capitalising letters other than those following a full stop, and will not be explored here.

5.4.3 Methods

5.4.3.1 Challenges in identifying and categorising respellings

WordSmith Wordlist was used to generate wordlists, which were sorted manually into probable spelling groups,[1] automatically stored by *WordSmith* in order of the frequency of occurrence of each group. The headword of each group was that judged the most standard or original form, so that respellings could be explained in terms of their derivation from the standard. Incidentally, the standard form was generally the most frequent. For example, *are* (with 1,054 occurrences) was considered the headword of a group comprising *r* (422 occurrences) and *ar* (2). Where a respelling was more frequent, such as those below, the standard form remained the headword.

Figure 5.6 Groups where the headword is less frequent than the respellings

because	(with 80 occurrences compared to 226 occurrences of *cos*)
until	(with 46 occurrences compared to 104 of *til*)

The only cases in which the headword was not the full, standard form were standard contractions (such as *I'll*) and the colloquial contraction *gonna*. As two-word forms were not included in the list, nor were two-word contractions such as *wanna* (*want to*) and *kinda* (*kind of*). *Gonna* is included not as a respelling but because it is respelt as *gunna* and *gona*. Similarly, standard contractions such as *I'll* occur where respelt (*ill*), as do other respellings of standard contractions without apostrophes, and *ur* for *you're*. As with *wanna* and *kinda*, multi-word units such as *of course* (respelt 3 times in CorTxt as *course*) or *by the way* (shortened 8 times to *btw*) do not occur in the list. That this distorts the frequency list is acknowledged.

5.4.3.2 Ambiguous or 'shared' respellings

Ambiguous respellings included respellings with more than one referent, such as *2* (which can refer to *to, too* or the numerical value); and *ur* (*your* or *you're*/*you are*).[2] These had to

[1] The term spelling group refers to a set of words felt to be alternative spellings of the same word, which were grouped using WordSmith's lemmatisation function. However, the spelling groups can be contrasted with the grammatical lemma, which encompasses inflected forms of the same word (*goes* and *went* as inflected forms of the lemma *go*, for example). Lemmatisation is done manually in Wordsmith with a combination of the F4 and F5 keys: F5 selects the words to be lemmatised or grouped, and F4 matches them together.

[2] *ur* was considered to be a respelling of *you're* in order to include it in the investigation, although it could perhaps more accurately be a respelling of *you are*.

be checked through sorting concordance lines, and adjusted manually. Concordance lines, for example, showed that *2*, occurring 910 times, occurred 690 times as *to*, 61 as *too*, and 159 as the number, which the word-frequency list was adjusted to reflect, with *2* (*to*) for example added to the spelling group *to*.

Figure 5.7 Selected concordance lines for *2*

1	journey. Xx" "hi bet its	**2**	in the mornin hope u
2	't throw it away now it is	**2**	important - the life u h
3	ane thing i've ever been	**2.**	If theres a better festi
4	ere. Howz u? Wot u up	**2**?	I'm just watchin tv.
5	bout 2 stores shut - the	**2**	i wanted 4 bags! No s
6	am sure he is missing u	**2**	- I miss u loads xxx"
7	hink?as much as i want	**2,**	i dont want 2b a bad

Other 'ambiguous' examples occurring in the top 150 groups included the following. As the list shows, they include 'shared respellings' (where more than one word is respelt the same way, as in *problem* and *probably*, both respelt as *prob*) as well as respellings which also constitute the standard form of another word (*ill*, for example, as a respelling of *I'll* and a word meaning to be sick; *no* as a shared respelling of *know* and *number*, and as a negative marker).

Figure 5.8 Other ambiguous respellings

d: a respelling of *the* (21) and *do* [with an apostrophe] (12);

da: a respelling of *the* (6) and part of a drum roll: *da da!* (5), and the Italian *da* (3) [in *da vinci code*];

an: a respelling of *and* (19), and as the article *an* (253);

ur: a respelling of *you are* (*you're*) (177) and *your* (286)

4: a respelling of *for* (357), and the number (74)

no: a respelling of of *know* (57) and *number* (50) and as a negative (633);

ill: a respelling of *I'll* (46) and meaning 'sick' (30);

com: a respelling of *come* (2) and as in *dot com* (6);

bin: a respelling of *been* (22) and a container for rubbish (7);

prob: a respelling of *problem* (51) and *probably* (45)

tank: a respelling of *thank* (2) and as the vehicle (2);

cause: a respelling of *because* (4) and referring to reason (2);

were: a respelling of *we're* (27) or the verb *were* (163);

mess: a respelling of *message* (2) or meaning 'chaos' (6).

yr: a respelling of *your* (13), *you're* (4), or *year* (4)

eve: a respelling of *evening* (54) or in *Xmas* or *New Year's eve* (5)

st: a respelling of *street* (25), *saint* (13) or *student* (2)

A couple of respellings proved difficult to identify and label. For example, it was difficult to decide whether *wk* in the following referred to *work* or *week*.

(5.10) Hi. Wk been ok - on hols now! Yes on for a bit of a run. Forgot that i have hairdressers appointment at four so need to get home n shower beforehand. Does that cause prob for u?"

(5.11) Hi, wkend ok but journey terrible. Wk not good as have huge back log of marking to do

It was also difficult to interpret *dey* in the following.[1] Although it could also be explained as a respelling of *there*, it was categorised as a respelling of *today*:

(5.12) swhrt how u dey,hope ur ok, tot about u 2day.love n miss.take care.NAME365

(5.13) NAME306, ow u dey.i paid 60,400thousad.i told NAME613 u would call NAME613. NAME365

5.4.3.3 Hapaxes and misspellings

The question as to whether to exclude variants which occurred only once on the basis that these could be 'mistakes' or one-offs was also considered, but the decision was made to include them. This was firstly because, of the 28 hapaxes in the groups selected for study, most added interesting and creative examples to patterns observed elsewhere in CorTxt, as well as in the literature. Furthermore, hapaxes deemed to indicate individual style (rather than common tendencies) were useful in showing the degree of individual choice. Secondly, few hapaxes could be considered possible 'misspellings', in the sense that words were accidentally or unwittingly respelt. These were *thrus* instead of *thurs*, a common error of transposition (Mitton 1996: 91); and *untill*, either an error of insertion or a genuine 'error of

[1] Presumably written by the same texter (NAME635), and so presumably both with the same referent.

competence' by a texter who does not know whether *until* has one l or two (Mitton, 1996:57) [1] (although *til* was 'mispelt' fairly regularly as *till*, 104 and 70 occurrences, respectively). However, it was less feasible to label as misspellings *okej* (*okay*); and *thou* (*though*). The hapaxes are categorised below.

Figure 5.9 Spelling hapaxes in CorTxt
(all examples occur once unless otherwise stated)

Consonant writing:	*cn* for *can*, *hd* for *had*, *tht* for *that*, *knw* for *know*; *snd* for *send*; *wnt* for *went*; *gv* for *give*, *lk* for *look*; *bettr* for *better*, *missd* for *missed*, *nthn* for *nothing*, *thks* for *thanks*;
Abbreviation:	*getin* and *geting* alongside *gettin* (17)
Phonetic spelling:	*dun* for *done*; *alrite* for *alright*; *afta* for *after*, and *ova* for *over*;
Colloquial contractions or respellings:	*appy* for *happy*; *fil* for *feel*; *everythin* for *everything*; *liddle* for *little*; *mora* for *tomorrow*, alongside *2mora* (14) and *morrow* (6); *soo* and *soooo* as well as *sooo* (5) as respellings of *so*;
Liensed variants:[2]	*hallo* as well as *hello* (419).

Transpositions and some insertions were judged to be unintentional or uniformed 'mistakes'. These were not, however, restricted to hapaxes and included: *iin*, *annd*, *jurt* (*just*), and *tomorrrow*, which all occur twice, *ot* (*to*) and *hte* (*the*), occurring three times and *adn*, ten times. This judgement was made on the basis, justified below, that no social meaning or function was apparent in the respelt forms. However, it must be acknowledged that the extent to which it is possible to determine whether spellings are 'deliberate' or mistaken is questionable.

[1] Mitton's (1996) overview of 'slips and typos' is written with a QWERTY keyboard in mind, but it is not hard to see that the same errors could occur with a mobile phone keypad. He describes the 'incorrect' doubling of a letter as both an 'error of competence' and an 'error of performance' where the texter unwittingly presses the key twice.
[2] A term used by Sebba (2007: 30) to refer to cases where 'the conventional norms allow for a choice', another example being *judgment* or *judgement*.

5.4.4 Results and Discussion

5.4.4.1 Spelling groups: an overview

The following investigation was based on the top 150 spelling groups occurring in CorTxt, taken from the top 250 headwords.[1] While narrowing attention to some extent on grammar, rather than content words, the arbitrary restriction made the task more manageable and, by focusing on the most frequent words, ensured that most respellings also occurred frequently in CorTxt.

Two immediate observations made from the lemmatised or grouped list are the tendency for more frequently-used words to be respelt; and the choice open to texters as to how and why they respell. The wordlist in Appendix 5.1 includes the 250 most frequent headwords and shows that words which occur more frequently in texting are more likely to be respelt than those occurring further down the frequency list. The fact that thirty-three of the top fifty groups have spelling variants, compared to only nineteen of fifty less frequent examples, ranked 200-250, can be explained not only in that common words tend to be more irregularly spelt and thus open to being respelt phonetically (Weber 1986: 418) but also because the respelling of frequently-used words also ensures recognition (between texters) of variants used. This extends Grinter and Eldridge's (2003) observation that texters in their study shortened everyday words such as *tomorrow* and *school* (rather than complex acronyms listed by Crystal 2008a), to the fact that texters also regularly respell common functional words such as *you* (*u*, *ya*, *ye*, *yer*), *to* (*2*), and *the* (*d*, *da*, *te*, *th* and *ze*).

As mentioned above, investigation was limited to the 150 most frequent spelling groups, excluding words in the original 250-word list which were not respelt (Appendix 5.2). The top 50 most frequent spelling groups are shown below in Table 5.2 to illustrate the nature of respellings and their distribution across headwords. In each case, the frequency of the group includes the combined frequency of the headword and its variants. Evident in this list is the tendency not only for more frequent words to be respelt, but to be respelt in more varied ways. Although many of the more frequent words in the list have one or two respellings, some have between three and seven and these tend to be higher in the frequency list. The fact that *tomorrow* stands out in the list below with 20 variant forms is perhaps due to the fact that it is not only frequently used (in texting) but that its length

[1] The top 250 headwords included words which were not respelt, whilst the top 150 groups were only those with respellings.

renders it open to different forms of respelling. The other noticeable feature, concerning these variants, is the frequent co-occurrence of respellings which reflect speech with those described as eye dialect and abbreviation, as alternative or 'competing' respellings of the same word. This is illustrated by *ya* and *u* as respellings of *you*. In other words, texters have choices not only in *how* they respell but *why* they do so: reasons connected with brevity, spoken informality and visual effects. The above points are returned to towards the end of the investigation.

Table 5.3 The top 50 most frequent groups including variant forms

	Headword	**Freq of grp**	**%**	**Variant forms**
1	YOU	7,884	4.14	u(3043),ya(256),yer(14),ye(9),uu(2)
2	TO	4,976	2.61	2(690),ot(3)
3	X	3,689	1.94	xx(833),xxx(635),xxxx(11),xxxxxx(3),xxxxxxx(2),xoxox (1),xxxxx(42)
4	THE	3,553	1.86	d(21),da(6),th(8),hte(3),te(2),ze(2)
5	AND	3,171	1.66	n(182),an(19),adn(10),amd(2),annd(2)
6	IN	2,387	1.25	iin(2)
7	FOR	2,057	1.08	4(357),fer(2)
8	HAVE	1,993	1.05	av(8),hve(6),ave(5),hav(106)
9	BE	1,567	0.82	b(375)
10	ARE	1,478	0.78	r(422), ar(2)
12	GOOD	1,265	0.66	gud(40),gd(25),goodo(3)
11	SEE	1,255	0.66	c(248)
13	JUST	1,240	0.65	jus(18),jst(6),jurt(2)
14	I'M	1,216	0.64	im(280)
15	SO	1,159	0.61	soo(1),sooo(5),soooo(1)
16	WILL	1,124	0.59	wil(12)
17	THAT	1,118	0.59	tht(1),dat(4),tha(2)
18	YOUR	1,111	0.58	ur(286),yor(2),yer(9),yr(13),u'r(2),ure(1)
19	NOT	1,106	0.58	nt(4)
20	DO	1,061	0.56	d'(12),du(5)
21	OK	1,040	0.55	okay(75),okey(27),k(33),okej(1),okie(12)
22	WITH	1,002	0.53	wiv(82),wid(4),wiht(3)
23	WAS	952	0.5	woz(4),ws(3)
24	YES	952	0.5	yep(69),yeh(22),yea(15),yeah(382),yup(13)
25	CAN	894	0.47	cn(1)
26	TOMORROW	888	0.47	tomoz(9),tomorro(6),tomorow(4),tomora(3), tomo(361),morrow(6),mora(1), tom(24),2mora(14),

				tomoro(10),2morrow(9),tmw(9),2morow(4),2morro(4), 2mrrw(4),2moz(3),2mrw(2),amoro(2),tomorrrow(2),2mo ro(42)
27	WHAT	839	0.44	wot(148),wat(37)
28	KNOW	776	0.41	no(57),knw(1),kno(2)
29	ABOUT	768	0.4	bout(123),abt(4)
30	HOPE	746	0.39	hpe(2)
31	BACK	733	0.38	bak(38),bac(7),bck(3),bk(11),bek(2)
32	NO	714	0.37	nope(19),nah(18)
33	NOW	713	0.37	nw(9)
34	GOING	710	0.37	goin(60)
35	ITS	696	0.37	it's(240),tis(21)
36	HOW	692	0.36	hw(2)
37	DON'T	671	0.35	dont(198),dnt(4)
38	I'LL	628	0.33	ill(46),il(21),i'l(30)
39	THIS	598	0.31	dis(6)
40	WHEN	589	0.31	wen(75)
41	THERE	588	0.31	ther(2)
42	NIGHT	586	0.31	nite(105)
43	FROM	568	0.3	frm(9),fm(3)
44	HAD	562	0.3	hd(1),ad(2)
45	TOO	554	0.26	2(61)
46	COME	537	0.28	cum(8),com(2)
47	WELL	520	0.27	wel(5)
48	ONE	514	0.27	1(88)
49	SORRY	502	0.26	sori(3),sos(12)
50	SOON	499	0.26	sn(6)

5.4.4.2 Forms of spelling variation

The forms which the variants take, despite their varied and seemingly complex nature, can be grouped into four categories: letter substitution, letter omission, letter appellation and letter transposition, with each subdivided into vowels, consonants, and numbers. For example, *thanx* can be described as the 'substitution' of <ks> with <x>; *jst* as the 'omission' of <u>; and *soooo* as the 'appellation' of <o>'s. In several cases, one respelling exhibits two or more patterns. In *plz*, for example, the <s> is replaced by <z> (substitution) and the vowels <ea> are omitted (omission).[1] The full list of patterns followed can be found in Appendix 5.3, of which a selection is presented below.

[1] This is discussed further in section 5.4.4.7 below

Figure 5.10 Categories of respelt forms

Letter Substitution

- Vowels
    - <o> for <a> (/ɒ/) in *wot* and *woz*;
    - <u> for <o> du (/uː/); *cum*, *luv* and *gunna* (/ʌ/);
    - <a> for <er> (/ə/) in *betta*, *numba*, *lata*; <a> for <ou> (/ə/) in *ya*; <a> for <ow> (/ə/) in *tomora*);
    - <e> for <ou> (/ə/) in *ye*;
    - <i> for <ee> (/ɪ/) in *bin*;
    - <o> for <au> (/ɒ/) in *cos* and *coz*; <o> for <ow> (/əʊ/) in forms of *tomorrow* (*tomoro*);
    - <u> for <oo> ((/ʊ/) in *gud*; <u> for <oul> ((/ʊ/) in *wud*, *cud*;
    - <er> for <our> (/ɔː/ /ə/) in *yer*; *ur* for *your* (/jɔː/)
    - <u> for *you*
    - <ite> for <ight> (/aɪt/) in *nite*, *2nite*, *mite*
- consonants
    - <z> for <s> (/z/) in *coz*(24), *cuz*(3), *cz*(2), *plz*(4), *plez*(3), *howz*(8), *woz*(4)
    - <d> for <th> (/ð/) in *wid* and *dis*; <f> for <ph> (/f/) in *fone*; <v> for <th> (/ð/) in *wiv*; <x> for <ks> (/ks/) in *thanx*
    - <b> for *be*; <c> for *see*; <r> for *are*; <n> for *and*; <ne> for *any*
    - <v> for *very*
- numbers
    - <2> for <to> (tuː) in *2morrow*, *2nite*, *in2*
    - <4> for <for> (/fɔː/) in *b4*
    - <8> for <ate>, <eat> (/ ɛɪt/) in *m8* and *gr8*

Letter Omission

- vowels
    - vowel omission (gd, hve, thx, frm, lv, plz, cld, wld)
- consonants
    - double letter reduction (beta, tomoro, i'l, stil, wel, gona)
    - final letter omission (an, com, goin, bac, wil, jus)
    - first letter omission (<h>, <th>) (av, ello, em)
    - mid-position omission of h (wat, wen, wot, te)
    - other mid-position omissions (gona, thx)

- syllables
    - final syllable omission (tomo, prob, num, mess)
    - first syllable omission (k, bout, cause)

Letter Appellation

- o appellation (so → soooo)
- p appellation (yeah → yep; no → nope)
- y appellation (please → pleasey)
- doubled letter (summort, summing, untill, till)

Letter Transposition

- *adn* for *and*
- *hte* for *the*
- *ot* for *to*
- *thrus* for *thurs* (*Thursday*)

The only category which does not appear to follow formal patterns is transpositions, thus confirming the suggestion that these can best be described as 'mistakes'. In contrast, substitutions, for example, involve systematic replacement of less frequent spellings with what Carney (1994) describes as 'default' or most frequent spelling for the sound: <o> for <a> (/ɒ/) in *wot*, *woz* and *coz*; <u> for (/ʊ/) in *gud*; *wud*, and *cud*; as well as for (/ʌ/) in *cum*, *luv* and *gunna*; <a> for (/ə/) in *betta*, *numba*, *lata*; and *tomora.*[1] In the case of consonants, it is perhaps not frequency which governs the choice of respelling but the more marked or prototypical association of certain letters with certain sounds. <z>, for example (which conventionally occurs in less frequently occurring, more salient words than <s>), is associated with /z/; and <f> with /f/, rather than <ph>, historically restricted to words of Greek origin (Carney 1994: 228-230).

The need for brevity is an obvious motivation for substitutions and omissions. This is inevitable in the case of *omissions*, but most *substitutions* also involve reductions in the number of letters used. In many cases, one letter replaces two (or three), showing a straightforward reduction: <d> for <th> (*wid*), for example, <u> for <oo> (*gud*) or *ite* for

[1] What the current investigation admittedly ignores is the spelling's *context*. For instance, as Carney points out, although the default spelling of (/ʊ/) may be <u>, it tends to be spelt <oo> before /d/, just as /ɒ/ is typically spelt <a> after /w/. Another example is that although <igh> is the default spelling of /aɪ/ only in 'very restricted' contexts, one of the contexts is before /t/: making *night*, for example, the 'default' spelling rather than *nite*.

ight. Although others appear to involve one-for-one letter exchanges, it can be argued that some form of reduction is taking place. The substitution of <a> for <o> in *wot*, for example, also involves omission of <*h*>,[1] while the replacement of <o> with <u> in *cum* and *luv* also involves omission of the final <e>.[2] The only example which does not involve a reduction in letters is the replacement of <a> with <o> in *woz* (for *was*). However, as argued later and as shown by choices available to texters in how and whether they respell words, it should not be assumed that abbreviation and the need for brevity is the only, or even the predominant, motivation. Certain substituions, for example, show alteration in pronunciation. <a> reflects the schwa in *betta* and *lata*; while <er> in *yer* replaces <our> in *your* (/ɔ:/) to reflect the spoken /ə/. Nor, clearly, do appellations perform abbreviating functions and instead create speech-like expressions which illustrate other motivations behind respellings. The apparent functions behind these formal patterns are detailed below.

5.4.4.3 The functions behind spelling variation

The functions of variation which emerge from categorising these formal patterns can largely be described following those emerging from the literature.

Figure 5.11 Functions of spelling variation in CorTxt

- **Colloquial contractions** (Weber 1986) in *CorTxt* include: *n*, *av*, *yer*, *wiv*, *goin*, *bin*, *allo* and *fink*.
- Other **colloquial respellings** include *goodo*, *pleasey*, *nope* and the informal variants of *yes*: *yep*; *yeah*, *yup*.
- **Regiolectal respellings** (Androutsopoulos 2000) include *summat*, *summort*, *sumfing* and *summing* for *something*;
- Examples of **phonetic spelling** include: *2*, *4 u*, *b*, *c*, *gud*, *woz*, *coz*, *thanx*, *wot*, *nite*, *cum*, *luv*, *fone*, *cud*, and *wud*.
- **Abbreviations:** clippings include *tomo*, *cause* and *bout*; while the final letter is omitted from a number of words including: *are*, *have*, *will*, *just*, *all*, *back*;
- **Consonant writing**, in which the vowels are omitted, is evident in *gd* (*good*), *jst* (*just*), *thks*, *thnx* or *thx* (*thanks*), *bk* (*back*), *frm* (*from*), *wk* (*week*), *lv* (*love*), *pls* or

[1] Although, as <h> is part of the consonant cluster <wh> rather than contributing to the vowel sound (Carney 1994: 328), it is not the case that this could be described as the substitution of <ha> for <o>.
[2] Dropping the <e> would be preferred, according to the principles of English orthography (cf Carney 1994: 129) in order to represent a short vowel sound.

plz (*please*), *cld* (*could*), *wld* (*would*), *nxt* (*next*), *txt* (*text*), and *wknd* (*weekend*).

- The **omission of apostrophes** is evident in the otherwise standard contractions: *im* (*i'm*) and *ure* (*you're*) (categorised as a type of abbreviation).
- **Visual morphemes** (Bolinger 1946) are present in the symbol *x*, or a series of *x*'s (and o's), which occur frequently, apparently to represent kisses in order to sign messages off.

Categorisations of forms which fulfil these functions are outlined below[1] and discussed in turn below. As evident from the list, transpositions are not seen as fulfilling functions in the text messages.

Figure 5.12 Functions and the forms which realise them

1. abbreviations
    a) apostrophe omission (im, its, dont ill)
    b) double letter reduction (2morow, wil, beta, gona, hapy)
    c) mid letter(s) omission
        i. h (nigt, wat, wich, wen, tnx)
        ii. c from ck (bak)
        iii. n (thx, thks)
        iv. e (havnt), <a> (hve, yeh), ou (abt), <i> (finshed)
        v. ee (wkend), eek (w'end),
        vi. lea (pse)
        vii. shd
        viii. bday, b'day
    d) final letter(s) omission
        i. e (sum, luv, ar)
        ii. h (o), k (bac), s (thur), w (kno)
        iii. rs (thu), gh (thou)
    e) final syllable(s) omission (fri, num, mess, eve, tomo, tom)
    f) initialisms (v, sth)
    g) substitutions
        i. sos for sorry
        ii. z for rrow (tomoz)
    h) standard abbreviations (pm, xmas, 1st)

2. consonant writing: omission of
    a) a (tht, bck, hd, lst)

[1] The transpositions and miscellaneous respellings will not be included in this discussion.

b) e (txt,next, snd, wnt)

c) ea (pls, spk, msg)

d) i (gv, hs)

e) o (2mrw, dnt, nt, hpe, nw)

f) oo (gd, sn, lk)

g) ou (cld, wld, shld, shd)

3. eye dialect

a) number homophones

i. 1 for one (/wʌn/)

ii. 2 for to (/tu:/)

iii. 4 for fore (/fɔ:/)

iv. 8 for at, ate, eat / ɛɪt/

b) letter homophones

i. b for be (/bi:/)

ii. c for see (/si:/)

iii. r for are (/ɑ:/)

iv. u for you

v. f for ph (/f/)

vi. ne for any (ɛnɪ)

c) ur for you're and your (/jɔ:/) – not phonetic

d) other homphones

i. ite for ight (2nite, rite, mite)

ii. y for why (/waɪ /)

iii. yt for ight (/aɪt/) (myt)

iv. no for know (/nəʊ/)

e) schwa represented

i. a for er (/ə/) (lata, afta, ova, betta, numba)

f) letter o

i. o for <a> (/ɒ /) (wot, wots, wot's, woz, coz, cos)

ii. o for ow (/əʊ/)

g) letter u

i. u for au (/ʊ/)

ii. u for <o> (/ʊ/)

iii. u for <o> (/u:/) (du)

iv. u for <o> (/ʌ/) (cumin, cum, luv, dun)

v. u for oo or oul (/ʊ/) (gud, cud, wud, shud)

h) other vowel sounds

i. ar for <a> (/ɑ:/) (arvo)

ii. or for ough (/ɔ:/) (thort)

iii. or for our (/ɔ:/ /ə/) (yor)

i) consonants

i. x for <ks> (/ks/) (thx)

ii. z for <s> (/z/) (coz, cuz, plz)

iii. doubled letter (m) (summat, which shortens the previous vowel sound)

j) eye dialect or clipping (tho, although)

4. visual morphemes

a) xx, xxx, xxxx

5. colloquial contractions

a) first letter omitted

i. h (ad, ell, ere, avin)

ii. th (em)

iii. a (n)

b) first syllable omitted (though, cos, till, bout, k, morrow, mora)

c) mid syllables omitted (satdy)

d) mid letters omitted (probaly)

e) weak sounds represented

i. shwa (/ə/)

1. a for <e> (/ə/) (da)
2. a for ou (/ə/) (ya)
3. a for ow (/ə/) (tomora)
4. a for to (amora)
5. d for th (/ð/) (d, dat, wid)
6. e for ou (/ə/) (ye)
7. er for our (/ɔ:/ /ə/)

ii. /ɪ /

1. i for ee (/ɪ /) (fil, bin)

f) f for ph (sumfing)

g) t for th (/θ/ /t/) (tank)

h) z for th (ð) (ze)

i) v for f (/f/) (arvo)

j) final letter omitted

i. d (an, n)
ii. e (th, d)
iii. g (avin, goin, mornin, somethin)
iv. t (jus, tha, tex)

k) final syllable omitted (prob, morn, mo, avo, arvo)

l) tis

m) something (summort, summat, summing)

6. colloquial respelling
    a) letter addition
        i. p (yep, nope)
        ii. o (soo, goodo)
        iii. y (pleasey)
    b) ah for <o> (/æ/, /ə/) (nah)
    c) er for or (/fə/) (fer, yer)
    d) dd for tt (liddle)
    e) rr for ll (herro)

7. typo
    a) doubled letter (annd, iin, tomorrrow)
    b) substitution (jurt)
    c) transposition (ot, hte)

8. miscellaneous

5.4.4.4 Colloquial contractions and respellings

As the figure shows, many colloquial contractions noted by Weber (1986) occur in CorTxt. Omission of final letter <g> from progressive verbs can be seen in *goin*; and of <d> after <n> in *an* and *n*. First letter omission of *ad* and *avin* reflects the dropping of <h> in personal pronouns (‘*im*, ‘*er*, ‘*em*); and the various spellings of *you* are also identified by Weber (1986); as is aphaeresis of *ok* (*k*), *because* (*cause*, *cos*, *coz*) and *about* (*bout*). The prevalence of <a> in representing schwa, noted by Weber (1986: 424) and Carney (1994: 204) as a strong pattern in English, is conformed to (in CorTxt) in *ya, da*, *tomora* and *amora*, as well as *gonna*, *gunna*, *gona*), although it is interesting to note that *ye* (an older respelling of schwa), *yer* and *fer* also occur, although much less frequently, reflecting the wider representation of schwa across lexical words (Weber 1986: 424; Carney 1994: 204). Other patterns in texting echo others noted by Weber (1986), although with some difference

in terms of form: deletion of 't' in *jus*, *tha* and *tex*, for example, echoes the deletion of <d> in *ol'* and *an'*, and can be seen in novel dialogue: 'What's tha'? A rat?' (Doyle 1993). *Probaly* is a relatively new respelling which, according to *Urban Dictionary*,[1] also occurs elsewhere; and *sori* reflects use of <i> in representing short /i/ sounds (as in *bin*). Rather than categorised as new or unique to texting, these must be described as extensions of existing functional patterns, reflecting those seen in other electronic forms of communication.

The implications are that colloquial contractions used in texting not only reflect existing *formal* patterns, but that texters also fulfil functions historically and widely fulfilled by these contractions: namely, they are striving to set the tone of the message, indicate emotions or construe identities through their texting practices, adopting and exploiting well-established linguistic devices for doing so. As in other texts, therefore, the effect created is of intimacy, informality and nonchalance. This effect is added to by the use of *colloquial respellings* such as *yep*, *nope*, *goodo*, *pleasey*, *nah* and *liddle* (representing how *little* might be spoken to a child).[2]

Regiolectal respellings fulfil similar functions. Replacement of <th> with <d> in *da*, *dat*, *wid* and *dis*, for example, reflects Caribbean respellings seen in Linton Kwesi Johnson's poetry: 'wi goin smash de sky wid wi bad bad blood'. More complex respellings such as *summort*, *summat* and *sumfing* (for *something*), *arvo* (for *afternoon*), *satdy* (*Saturday*), and *sanks* (*thanks*) can also be described as regional respellings. *Summort* appears to be an attempt to reflect South-West pronunciation of *summat* (itself from Yorkshire) and *sumfing* is estuary English (Crancher 1993). *Arvo* and *avo* (and perhaps *evo*), *sanks* and *satdy* represent spoken Australian regionalisms; *tank* (*thank*) and perhaps *te* (*the*) an Irish or New York accent; and *herro*, a respelling representing the pronunciation of East Asian speakers (according to *Urban Dictionary*). These regionalisms draw on English orthographic principles: doubling the consonant to indicate a previous short vowel in *summort*, use of graphemes with close correspondence to the sound of the spoken form (<f> for /f/ in *sumfing*; <v> for /v/ in *arvo*; <t> for /t/ in *tanks*), and omission of syllables not sounded in

[1] *Urban Dictionary* (www.urbandictionary.com) is an online dictionary of slang, to which users contribute newly created words or word meanings. It should by no means be relied on as a trusted authority, but is useful in determining words which potentially have some currency outside the corpus used in the current study.
[2] These are defined as colloquial respelling rather than contractions simply because, although fulfilling a similar function to contractions, they do not involve word shortenings.

fast, informal speech. As mentioned earlier, there is no clear divide between colloquial and regional spellings: it coud be argued, for example, that *yer* and *fer* are regional representations. Nonetheless, these respellings seem to show that texters are willing to *lengthen* words in order to establish an informal arena and perform speech-like intimacy, and thus highlight the importance of the interpersonal over physical constraints.

Messages which include colloquial contractions, colloquial respellings and regioletal variation include the following. The use of other kinds of respellings alongside these (homophones *r* and *u*, for example) illustrates the extent to which respellings co-occur.

(5.14) Hello beautiful r u ok? I've kinda ad a row wiv NAME99 and he walked out the pub?? I wanted a night wiv u Miss u xx

(5.15) Thought praps you meant another one. Goodo! I'll look tomorrow xx

(5.16) NAME79 says that he's quitting at least5times a day so i wudn't take much notice of that. Nah, she didn't mind. Are you gonna see him again? Do you want to come to taunton tonight? U can tell me all about NAME79!

5.4.4.5 Phonetic spellings

As with colloquialisms, phonetic spelling in CorTxt reflects existing patterns of variation: those found not only alongside colloquial contractions but also in advertising, brandnames, and other less regulated spaces such as fanzines or graffiti. In contrast to colloquialisms, intended effects are not necessarily created through reference or connections to alterations in spoken language and thus to oral identity, but effects are instead achieved through visual representations of meaning (Kress 2000). Texters use number and letter homophones, as in other domains, often combining the two in a way seen in the music-group *U2*: *b4*, *m8*, *cu* and *ur* (*you are*). Other grapheme substitutions include: *ite* for *ight* (in *mite*, *2nite*, *rite* and *lite*), which as we have seen is a default spelling of /aɪ/ except in monosyllabic words;[1] the more unusual *yt* in *myt* (*might*) which extends the spelling of /aɪ/ with <y> in *my* and *dry*; and consonant substitutions: <x> for <ks> in *thx*; and <z> for <s> in *coz*, *cuz*, and *plz*. As mentioned above, these consonant forms are the most salient spellings of particular

[1] *Lite* occurs twice in the corpus, but its headword (*light*) does not occur within the top 250 words.

phonemes, rather than the most frequent spellings. Vowels used in eye dialect include <a> for schwa (*lata*, *afta*, *ova*, *betta*, *numba*), an orthographic practice in English noted by Weber (1986) and Carney (1994) (and see colloquial contractions above);[1] <o> for <a> in *wot*, *woz* and *coz*; <u> for <o> in *cum*, *cumin*, *luv*, *dun*; and *thort* for *thought*. Most are familiar from other domains: *ite* and *wot* from advertising; *woz* from graffiti; *gud* and *luv* from informal personal writing; <x> from Androutsopoulous's fanzines. Other forms in CorTxt may seem less familiar, but they extend existing patterns or practices: representations of schwa in words like *betta*; <u> for <oo> or <oul> in *cud*, *wud*, and *shud*; *cumin* (*coming*), and *myt* (*might*).

Seemingly new letter homophones include *ne* (*any*) with 8 occurrences and *y* (*why*) with 7 occurrences:

Figure 5.13 Concordances for *ne* (*any*) and *y* (*why*)

1		E543 SURNAME44 by **ne**	chance? X ps hope
2	" "aus arnt in the top 3	**ne**	more neither r englan
3	1 lol. Don't think there r	**ne**	others. U goin 2 car
4	"Mam said dont make	**ne**	plans for nxt wknd c
5	ack at lunch c if ive got	**ne**	post, no worries c u
6	How it goin? If you have	**ne**	probs just give us a
7	!i hope u have another 1!	**ne**	way,wud u like it ba
8	tuff, lol" "Lol, it cant be	**ne**	worse than it already

1	thers fancy it?" "Been?	**Y**	D'you want 2 know?
2	DRESS" "Love u loads!	**Y**	didn't u take new pho
3	f swifties" "yep dont no	**y**	hes dun fuck all - as
4	holiday at mo, which is	**y**	my start date was to f
5	I couldn't poo more dad.	**Y**	u keep sending mess
6	av a gud w'end at home?	**Y**	wud i b savin my txts
7	a bell later if u want. X"	"**Y**	r england so shite?

Use of *ur* in respelling *your* and *you're* is not phonetic, as indeed nor were some eye dialect forms in Androutopoulous's (2000) fanzines.[2] However, in CorTxt, *ur* occurs more often as

[1] The use of <a> to represent schwa in words like *numba* is an example of eye dialect and not colloquial contractions, in that they represent the standard pronunciation of the word rather than a regional accent (e.g. *summort*) or an informal contraction (as in *kinda*).

[2] There is also one example of *thanxs* in the text messages, which more closely resembles Androutsopoulous's non-phonetic spelling *H-Blockx*.

a respelling of *your* (286 occurrences) than it does of *you are* (177 occurrences). Below are randomly-selected occurrences of *ur* as the possessive.

Figure 5.14 Selected concordances for *ur* (*your*)

```
1                hope u can sit down in ur dress!! Have a good
2               te an age how are u and ur work av a wonderful d
3             ak at all! How hav u laid ur essay out? I've define
4             ese up u could get rid of ur bra! She grabbed his
5                u, hope 2see you outwi ur mates on satdy xx"
6                rnwall so i remembered ur offer! No worries if diff
7               e a reality check. Hows ur day been? Did u do a
8             hich one(s)) don't forget UR name ta NAME84"  "
9               erson take advantage of ur boozed up state + do
10            old. Have lovely hol with ur sis. X s ps done v littl
11          y you a pint! Xx"  "I'll do ur hair4u4free! No final, it
12               ether to come in. Enjoy ur eve s xx ps doom n
```

Use of <ur> to represent /jɔ:/ supports the argument that *ur* may be a respelling of *you're*, again representing /jɔ:/, rather than a phonetic respelling of the full form, *you are*. This observation regarding *ur* as *your* and *you're* further supports the assertion that it is the visual representation or shape of phonetic spellings which users recognise and which matters in conveying meaning. Many of these forms also serve to reduce characters, however, and it is difficult with some to know how best to define them: *tho* for *although*, *kno* for *know*, and *tomoro* for *tomorrow* for example, could be categorised as clippings or phonetic spellings.[1] Messages using phonetic respelling (and other respellings) include the following.

(5.17) Thnx dude. u guys out 2nite?

(5.18) Ok that would b lovely, if u r sure. Think about wot u want to do, drinkin, dancin, eatin, cinema, in, out, about... Up to u! Wot about NAME408? X

(5.19) Hey! Congrats 2u2. id luv 2 but ive had 2 go home!xxx

[1] Other variants which simultaneously perform more than one function include the omission of *g* in the final position of progressive verbs—a colloquial contraction, but also a further example of a clipping encouraged by the predictive text device. The same could be said of *jus* (for *just*).

5.4.4.6 Abbreviations

Familiar initialisms such as *v* (*very*) or *sth* (*something*) and standard abbreviations such as *pm*, *xmas*, *no* (for *number*) or *1st* are fairly infrequent in *CorTxt*. Other abbreviated forms in the text messages depart in marked ways from respellings elsewhere. However, as with other functional categories, in most cases unusual forms emerge from extending or combining existing patterns. For example, double-letter reduction, producing *stil* (*still*), *gona* (*gonna*), and *beta* (*betta = better*), although appearing unusual, has been seen to occur not only as a medieval printers' trick but in advertising slogans such as *Hot-Stuf* (Carney, 1994:447);[1] and the omission of other final letters such as <e> in *hav* and *sum* can be explained as part of recognisable phonetic spelling.[2] To take another example, although the following forms are unconventional in that they are not widely used elsewhere: *kno* (know), *bac* (back), *thou* (though), *num* (*number*), *mess* (*message*), *thu* (*Thursday*), *tomo* and *tom* (*tomorrow*), they can be described as extensions of the practice of clipping (Crystal 2003).[3]

Other unfamiliar forms include *sos* for *sorry* and *tomoz* for *tomorrow*. *Sos* is a variant of *soz*, with both *soz* and *tomoz* recent, largely online abbreviations of, respectively, *sorry* and *tomorrow* (*Urban Dictionary*). Another example of an unfamiliar form is the frequent apostrophe omission, which happens more often than in other discourses and may be encouraged by technological constraints of the mobile phone. *Its* exists both as a contraction of *it is* (with an apostrophe) and as a possessive determiner (without an apostrophe) and is often an area of confusion for writers (Crystal 2003: 203) but apostrophe omission is evident in texting not only in *its* but in contractions not so widely reported to confuse: *dont*, *ive*, *havent*. Incidentally, concordance lines generated for *its* in CorTxt reveals no occurrences of determiner *its*. Instead, *its* is used exclusively as a contraction of *it is*.

Figure 5.15 Selected concordances for *its*

```
1                e away..."  "Well yeah its there if i want it, dull
2               ing the rugby? Dunno if its england tho thinking
```

[1] The respelling *stuf* does in fact occur once in *CorTxt*.

[2] *sum* and *hav* represent the phoneme /ʌ/ better than *sume* where, by English orthographic conventions, the <e> lengthens the vowel.

[3] It may be the case that clippings are favoured by the predictive text devices particular to mobile phones, whereby the phone predicts the most likely letter sequence as you type. Pressing the 8 key (tuv) and then the 6 key (mno) produces *to*, the third press suggests *too* as the most likely sequence, but by the fourth, *tomo*. Another example may be the final letter reductions in *i'l*, *stil*, *wel*, *com*, *hav* and *jus*, where the phone has apparently recognised the word before the last letter has been typed and thus affords a convenient abbreviation.

```
3                ad my quota."  "mate if its because of the menth
4                another e-mail address. Its EMAILADDRESS not
5                 t id send u a message, its not an e mail its a te
6           iv it all say its excellent! Its NAME281's favourite
7                s hope hey"  "Sorry but its an oops."  "Oh no. it
8               ay on dvd that i bought. its a goo
9                called me this morning, its all cool....I got a job!
10              tin u n co in da pub but its not goin 2 happen ju
11                 eer.."  "Hey NAME242, its NAME67. how ru?i h
12              on..x"  "yes, as long as its before of after work..a
```

Omission of certain mid-letters is similarly unusual: <h> from *nigt*, *wat*, *wich*, *wen*, *tnx*; <c> from *bak*; <leas> from *pse*; and <n> from *thx* and *thks*. In some examples users seem to rely on the *shape* of the word, rather than the sound, to convey meaning, in a way outlined by Kress (2000). Consonant writing can be seen to rely on similar resources, that of shared awareness of a word's visual appearance, and the role of consonants rather than vowels in conveying this (Crystal 2003: 425). Consonant writing in CorTxt includes the omission of <a> (*tht*, *bck*, *lst*), <e> (*txt*, *nxt*, *snd*), <ea> (*pls*, *spk*) <o> (*2mrw*, *nt*, *hpe*), and <ou> (*cld*, *wld*, *shld*). The use of <x>'s to represent kisses, a common feature of personal correspondence (Kesseler and Bergs 2003), is similarly a visual device. The inclusion of <o>'s within some <x> sequences (*xoxox*) is interesting and can be interpreted as circles indicating eternity or forever: hence, everlasting kisses or love.

Messages with abbreviated forms include the following. Again, in most cases, abbreviations combine with other kinds of respellings.

(5.20) V skint too but fancied few bevies.waz gona go meet NAME211&othrs in spoon but jst bin watchng planet earth&sofa is v comfey; If i dont make it hav gd night

(5.21) Ok. Can be later showing around 8-8:30 if you want + cld have drink before. Wld prefer not to spend money on nosh if you don't mind, as doing that nxt wk.

(5.22) That's a shame! Maybe cld meet for few hrs tomo?

The above investigation shows how respellings in texting (colloquial contractions and respellings, regionalisms, phonetic respellings and abbreviation) fulfil certain functions

which can be compared to functions similarly fulfilled by respellings elsewhere. It is also evident that respelt forms in CorTxt largely follow orthographic principles of English and thus reflect historical and current spelling practices. We have also seen that where respellings seem unfamiliar, they generally result from extensions of existing practices to new wordforms. The extent to which respellings can be considered to reflect or extend existing patterns is listed in detail in Appendix 5.4, and main points summarised below.

5.4.4.7 New and combined forms

Respellings that do not widely occur elsewhere and to varying degrees extend existing patterns include the following.

Figure 5.16 New forms emerging from existing practices

- consonant writing (*bck*, *txt*, *spk*, *hpe*, *cld*);
- the omission of <h> in mid-position: *wat*, *wen*, *wich*, *wo* and *nigt*; the tension between *wat* and *wot* as competing variants of *what* is discussed below;
- the forming of new clippings: *tomo*(361), *tom*(24), *thu*(6), *thou*(1), *num*(7), *tho*(176), *altho*(2);
- final letter omissions such as *bac*(7) and *kno*(2), although again these reflect early printing practices as well as the necessary omission of final <e> in respellings such as *luv* in order to maintain the short vowel sound;
- the frequency of double letter reductions (eg, *stil*, *gona*, *beta*) although reflecting early printing practices as well as spelling in brandnames;
- mid letter omissions (*shd*(3); *pse*(2))
- the letter homophones *ne* (for *any*), *y*(29), *myt*(6) and *ur* and *ure* (for *you're*)
- *cud*, *shud*, *wud*, as extensions of the respelling *gud*.
- widespread apostrophe omission (*wont*, *im*, *cant*).

Other seemingly unusual forms actually result from combinations of orthographic principles: in other words, these 'combined' forms display more than one function at work simultaneously. These include *av*, where <h> has been dropped, suggesting a spoken form, while the omitted <e> appears likely to be a clipping. This stands in contrast to *cum*, a recognised example of eye dialect which requires the omission of <e> in order to represent

the short vowel sound and which, incidentally, produces a clipped form. A full list of combined forms can be seen in Appendix 5.5, and they include the following.

Figure 5.17 Combined forms

a. Combined forms with only one function

d (*the*)	<d> for <th> (/ð/) and final letter omission (<e>)
iv (*I've*)	apostrophe omission and final letter omission (<e>)
2nite (*tonight*)	<2> for <to> (/tu:/) and <ite> for <ight>
dnt (*don't*)	vowel omission and apostrophe omission
tho (*although*)	first syllable omission and <o> for <ough> (/əʊ/)

b. Combined forms with more than one function

mora (*tomorrow*)	first syllable omission, double letter reduction (r) and <a> for <ow> (/ə/) colloquial contraction + abbreviation
coz (*because*)	first syllable omission, <o> for <au> (/ɒ/), final letter omission (<e>), and <z> for <s> (/z/) colloquial contraction + phonetic spelling
cumin (*coming*)	final letter reduction (<g>) and <u> for <o> (/ʌ/) colloquial contraction + phonetic spelling

5.4.5.8 One implication: choice in texting

'Competing' respellings refers here to alternative respellings of the same word. *You*, for example, can be spelt conventionally as *you* (4560 times); or *u* (3043 times), an example of eye dialect for stylistic effect; or *ya* (256), *ye* (9) or *yer* (14), used to reflect regional or informal spoken forms. Other examples include the following and the full list can be found in Appendix 5.6.

Figure 5.18 'Competing' respellings

av(8), hve(6), ave(5), hav(106),
r(422), ar(2),
gud(40), gd(25), goodo(3),
jus(18), jst(6),
woz(4), ws(3),

tomoz(9), tomorro(6), tomorow(4), tomora(3), tomo(361), morrow(6), mora(1), tom(24), 2mora(14),
tomoro(10), 2morrow(9), tmw(9), 2morow(4), 2morro(4), 2mrrw(4), 2moz(3), 2mrw(2), amoro(2), 2moro(42),
wot(148), wat(37),
its(435), tis(21),
2nite(45), tonite(10), 2night(12), 2nigt(3),
hallo(1), ello(9), allo (5), helo(6), herro(3),
thanx(32), thx(5), tank(2), sanks(2), thnx(2), thks(1), thanxs(1), tnx(2),
cud(48), cld(19),
txt(94), tx(6), tex(2),
luv(46), lv(4), lov(4),
wud(38), wld(22),
cos(226), coz(24), cause(4), cuz(3), cs(2), cus(2), cz(2),
shud(11), shld(7), shd(3),
pleasey(4), plez(3), pls(35), plse(3), plz(4), pse(2),
giv(10), gv(1),
gunna(11), gona(6),
sth(9), somethin(8), summort(3), sumfing(2), summat(2), summing(2),
whats(48), wots(17), wats(7), wot's(3),
msg(7), mess(8),

d(21), da(6), th(8), te(2), ze(2)
n(182), an(19),
havin(24), avin(6),
mornin(23), morn(18),
l8r(20), lata(5), l8er(2),
mite(9), myt(6),

Some competing respellings illustrate how one function (abbreviation, for example, or eye dialect) can be achieved in different ways by exploiting more than one orthographic principle. The various ways in which words can be colloquially-contracted is illustrated in *ya*, *ye* or *yer*, which illustrate different ways of representing schwa;[1] and in *havin* (in which the <g> is dropped) or *avin* (in which the <h> is also dropped), as well as in *and*, contractions of which are captured by *an* or *n*. Respellings of *something* show several regional pronunciations: *summort*, *sumfing*, *summat* and *summing*. Different pronunciations of *the*, to take more examples, can be represented in various regional respellings: *d*, *da*, *th*, *te* and *ze*. Contrasting forms of eye dialect of *later* are *l8r* and *lata*; and of *might*, *mite* and *myt*. Others seem to indicate choices open to texters in reducing the number of characters

[1] Although *yer* may be a regional representation of a rhotic /r/.

used, depending perhaps on how much they wish to abbreviate: *thurs* or *thu*, for example, and *pls*, *plse* or *pse*. The choice between clipping or consonant writing in some examples may depend on texters' use of predictive texting. *back*, for example, is abbreviated either as *bac* or *bk*; *give* as *giv* or *gv*; *have*, as *hav* (21) but also twice as *hve*, and as *av* (8); *tomorrow* as *tomo* or *tmw*. Weekend occurs variously as *wkend*, *w'end*, and *w'kend*.

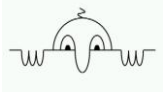
Others, however, suggest choices between two functions of respelling: the aforementioned variants of *you* (*ya* and *u*) are echoed in the spoken chatty form of *please* as *pleasey*, versus the brief abbreviated *plz*. Elsewhere, more conventional eye dialect form contrast with consonant writing in, for example, *would*, which occurs both as eye dialect, *wud*, but also as consonant writing, *wld*. Similarly *could* is spelt either *cud* or *cld* (7 times); and similar examples include *good* (*gud* or *gd*); and *love* (*luv* or *lv*).

The choice of *wot* or *wat* as variants of *what* is interesting as an example of conventional versus apparently new spelling forms. *Wot* is an example of phonetic respelling used in grafitti, as in 'Wot, no butter?' (Crystal 2003: 275).[1] Whether the less conventional form, *wat*, is a competing form of phonetic respelling or simply an attempt to cut down on characters is difficult to determine. The form is reflected also in *wen*, for *when*. A similar process may be in evidence with *come*, which occurs both as the conventional eye dialect form *cum* but also as the less conventional clipping *com*.

5.5 Chapter Summary

The above investigation draws on *CorTxt* to explore the meaning-making potential of respellings in texting. The observation that most spelling variation in *CorTxt* follows or extends one or more orthographic principles of the English language and thus reflects historical and current spelling practices is significant in providing empirical support for those who challenge the assertion that spelling in Txt is random or unprecedented and who argue that, by choosing to spell in principled ways which deviate from expected conventions, texters use respellings in *meaningful* ways. The assertion that texters have

[1] 'Wot no …?' was a popular post-war graffiti in Britain, accompanied by the picture of a figure peering over a wall, reproduced below, which commented on the lack of various items in the aftermath of the Second World War. According to Crystal (2003), it also occurred in similar forms in other countries.



choice, vividly illustrated through 'competing' respellings such as *u*, *ye* or *ya* for *you* and which underpins sociocultural understanding of creative respellings, implies texters actively and creatively choose how to present themselves, albeit constrained by contextual factors and orthographic principles.

It is naturally not possible to determine what texters intend to mean through respellings, or how they are interpreted by interlocutors, but research into other writing domains (such as interview transcripts, fanzines or graffiti) allows us to speculate on what spelling in texting *may* mean. As in other attempts to reflect spoken or regional pronunciations, colloquial contractions and respellings are likely to fulfil an indexical function and create an arena for participant relationships which, as in spoken interaction, are informal and intimate. Informality is thus construed by texters drawing on their awareness of features of everyday, face-to-face conversation, creating an expressive and wordy language which by most accounts can be captured through Fortunati's (2001: 314) term of 'written orality'. At the same time, however, texters also draw on existing patterns of abbreviation which contrast somewhat with the speech-like language described above and so add what could be described as *illusions* of brevity to what are often lengthy, expressive messages. The purely visual device of phonetic spelling also disrupts the relaxed intimacy with unconventional and eye-catching forms which evoke the originality and radicality of advertising, graffiti and underground subcultures. Through signalling deviance from expected, mainstream norms, these respellings affirm both group identity[1] and distance from other social groups. Putting all this together, we can see that texters express attitudes and emotions, define relationships, and construct texting identities through construal of spoken informality, brevity and deviance.

Texters' choices, I suggest, can be described in terms of *performativity* (Butler 1990). That is, through their choice of respellings, texters draw on awareness of other language domains in construing texting identities through *performances* of spoken informality, of brevity and of deviance. This perspective recognises, as does Shortis's (2007a,b) metaphor of the *orthographic palette*, the active role that texters play in choosing how to mean, how surrounding discourses and conventions constrain these choices, and how the choices made can change according to purpose and audience. However, performativity also extends Shortis's metaphor not only by beginning to explain *what* spelling choices may 'mean' but

[1] A role ascribed by Kasesniemi and Rautianen (2002), among others, to the language used in text messaging.

also by encompassing everyday creativity as well as structural ellipsis and features of texted grammar. That is, the concept of performativity allows us to go beyond the narrow focus on spelling suggested by previous text messaging research. The full implications for texting are explored in Chapter 9.

Of relevance now is the decision not to normalise spelling in *CorTxt*. This decision recognises the argument that respellings in texting are a potentially significant form of meaning-making which would be 'severely misrepresented in a reset, spell checked, cleaned-up typeset equivalent' (Shortis 2007a: 4). In other words, if texters use respellings to communicate informality, express attitudes, define group boundaries and *perform* texting identities, normalisation removes an essential part of interpersonal meaning.

CHAPTER SIX
EVERYDAY CREATIVITY IN TEXT MESSAGING

6.1 Introduction

This chapter reveals how everyday creativity such as idiom manipulation occurs not only in spoken interaction but also plays an important role in texting. Creativity as a normal if not prevalent feature of everyday speech is well-documented by linguists (Tannen 1989; Cook 1994, 1997, 2000; Crystal 1998; Carter 2004b; Maybin and Swann 2006). It is seen (by most) to include not only the deliberate and amusing play with idioms, metaphor or morphology described by Carter (2004b) as *pattern re-forming* but also the more covert repetition of interlocutors' vocabulary and grammar choices which, as described by Tannen (1989) and Carter (2004b), structures everyday conversation and which Carter terms *pattern forming*. The above studies also raise largely unresolved yet thought-provoking questions, including the extent to which 'successful' creativity relies on receivers' interpretations and responses or whether it can be recognised solely through form, whether *all* manipulation of forms can be deemed creative, and whether links can be made between creativity and context (Carter 2007; Maybin and Swann 2007).

Texting would nonetheless seem a fertile ground for everyday creativity: not only because its evolving nature and relaxing of prescribed language use encourage creative manipulation, but because text messages often resemble conversations and are sent between intimates for informal, phatic purposes.[1] Research into spoken creativity suggests it is indulged in not only 'for fun' but also to create convergence between speakers, indicate group membership, display identity, offer new perspectives and challenge interlocutors (Cook 2000; Carter 2004b). Contexts in which creativity occurs are therefore likely to be informal where relations are equal. Another function of creativity, particularly in developing social contexts where conventions are suspended or evolving, is the forging of identities and styles (Carter 2004b: 200). Despite this, investigation into everyday creativity in computer-mediated communication is only beginning (North 2006; 2007) and creativity in texting interpreted narrowly as referring to respelling (Hard af Segersteg 2002; Crystal 2008).

[1] See Chapter 2: Research into Text Messaging

This chapter therefore draws on CorTxt to show that everyday, speech-like creativity not only occurs in texting but plays an important role because of, rather than despite, technological constraints. It begins by exploring what is meant by *language play* and *creativity*, and discussing methods and issues involved in investigating creative language. Exploration of repetition as a pattern-forming choice in CorTxt highlights linguistic and contextual differences between spoken and texted interaction, while analysis of pattern-reforming choices suggests that the desire to be creative often overrides concerns with brevity and signals texters' metalinguistic awareness. As well as extending linguistic research which limits texted creativity to respelling, this investigation of texting addresses Carter's (2007) recognition of the need for greater understanding of the effect context has on creativity, whilst drawing to linguists' attention the potential for creativity in texting and its pedagogical implications.

6.2 Defining terms

6.2.1 *Language Play* or *creativity*?

6.2.1.1 Limitations of language play

My preference for the term *creativity* over *language play* is motivated by the latter's focus on deliberate, humorous language use illustrated by prepared or published texts (or anecdotal exchanges such as below), rather than collaborative, contextualised, everyday uses implied by *creativity*.

Figure 6.1 Example of 'ping-pong punning' (Crystal 1998: 2-3)

Janet:	And so there was a sort of confrontation between Crumble and Splash …
Jane:	Catfrontation, you mean. (Laughs.)
….	
Peter:	Near cat-astrophe, if you ask me (Groans all round.)
Janet:	I wasn't asking you, Peter!
Peter:	Sorry, I didn't mean to be categorical. (More groans all round.)
Jane:	This sounds like it's becoming a catalogue of disasters. (Peals of laughter).

The playful aspects of language use invoked by *language play* usefully highlights the fact that language is not always (or predominantly) concerned with information transferral, but

also with amusement or enjoyment, as illustrated above. Association of language play with the seemingly 'trivial' purpose of play also helps explain why mainstream linguistics has, until recently, largely ignored it (Crystal 1998: 1; Cook 1997), and its emergence as a serious topic recognises not only its prevalence in everyday language, but the wider significance of 'playing' within society (Cook 2000: 97-101). As this suggests, language play described by Cook and Crystal also has effects on perceptions and social relationships: used, according to Cook (2000), to collaborate and compete. However, recognition of overt and humorous language play as a fundamental part of social activity does not allow for patterning which operates on less conscious levels to fulfil similar interactional functions (Tannen 1989; Carter 2004b). Another limitation of Crystal and Cook's research, for current purposes, is their focus on verse, fiction and prepared performances rather than improvised 'play' developing through conversation. Although both researchers are in theory concerned with the role of language play within social relationships, their focus on largely decontextualised texts does not prioritise how creativity arises through social interaction. *Creativity*, as used by Carter (2004b), more effectively captures these aspects.

6.2.1.2 Creativity

Typical definitions of creativity run along the lines of 'having the ability or power to create' (Collins English Dictionary 1994) and 'involving the use of skill and the imagination to produce sth new or a work of art' (Oxford Advanced Learners Dictionary 2000). However, as Carter (2004: 24-30) argues, although creativity has since the 1700s been associated with originality, individualism, genius and invention, its semantic roots suggest this has not always been the case.[1] Modern Western ideas of the individual and autonomy restrict current notions of creativity to talented individuals, obscuring discussion of collaborative, contextualised creativity emerging from everyday interaction (Carter 2004b: 30). Carter's (2004: 54-8) argument that our parallel association of creativity with written 'literary' language arises primarily through the greater value placed on writing over speaking, and the greater priority granted to its analysis, reflects arguments made by linguists exploring spoken grammar (Brazil 1995; Halliday, 1989; Biber et al 1999; Carter and McCarthy 2006;

[1] The idea, for example, of the artist or writer as a lone *individual* is a recent concept which disregards the collaborative, co-creating work of writers such as Shakespeare or, indeed, of artists such as DJs who sample previous artists' work (Pennycock 2007); similarly, *invention* derives from the Latin *in-venir* (to come into) and implies an *intervention*, a building on existing knowledge within the particular circumstances and needs of a particular social group rather than any novel construct (Carter 2004b).

O'Keeffe et al 2007).[1] The implication of Carter's work is not that speech is necessarily as creative as literature, but that literary written language exploits the potential for all language to be manipulated and all language users to be creative (Swann 2006: 18). The instinctive reaction to this may be to object to any notion that Shakespeare has parallels with everyday conversation, but the important point is that Carter's use of *creativity* is free from consideration of social or cultural merit. Instead, the same resources which Shakespeare draws on to craft imagery and complex ideas are exploited spontaneously in conversation to fulfil less explicitly socially-valued tasks such as signalling involvement.

This definition of *creativity* therefore has a larger remit than *language play*, encompassing covert pattern construction as well as deliberate, playful manipulation of forms. If creativity need not be novel or humorous, however, the question remains as to what it does involve and how we can recognise it.

6.2.2 Defining creativity

Creativity can be defined as the manipulation of language form, in unexpected and yet contextually appropriate ways. It is therefore through manipulation of forms, rather than semantic meaning, that intended effects are achieved. This definition, to borrow Maybin and Swann's (2007) terms, can however be broken down into *textual* (formal), *contextual* and *critical* (evaluative) aspects. This chapter takes the view that creativity must initially be identified through formal properties, but that probable contextual and critical factors must be considered when exploring each instance.

6.2.2.1 Textual aspects

That creativity can be defined (at least in part) as 'manipulation of language form' is generally accepted across the literature, even among those who do not take a textual (data-based) approach (Cook 1994; 1997).[2] Swann (2006: 6), for example, suggests that in creative language '[s]ome formal aspect of language – such as sound, rhythm, grammar,

[1] Carter (2004: 56-7) draws on Halliday (1985/1989) to point out that, with the advent of tape recorders, not only can spoken language now be explored, but so can Halliday's assertion that spoken language is in fact richer, more fluid, and more spontaneous than writing.

[2] Cook's assertion, for example, that natural language can be 'preparatory, repetitive, artificial, removed from reality, and focused on the rules of the game' (Cook 1997: 230) leads him to argue against the accepted wisdom of current English language teaching (ELT) practices which prioritise meaningful communication over form-based instruction. The fact that language users often focus on grammar and phonology rules rather than semantic meaning can be seen, Cook argues, in the apparent significance of formal properties in texts as diverse as nursery rhymes and advertising.

meaning – is highlighted and this makes the utterance stand out'. Forms which are 'highlighted' through being repeated or non-canonical impact on listeners, thus achieving intended social effects. In Cook's (2000: 137) words, 'meaning and effect are dictated by patterning of forms, rather than the other way round'; and this is echoed by Carter's (2004: 81) claim that creativity: 'depends for its effects on particular patterns of language form'. Although seemingly excluding consideration of language play on a semantic level, as in the creation of fantasy worlds, the formal definition in fact allows us to position fiction as a function of language manipulation (Cook 2000: 58). Cook argues that our ability to disassociate syntax from semantics and operate them separately enables us to use syntax as a basis on which to create syntactically-correct but semantically-unreal situations. Cook's (2000) argument is limited somewhat in that he applies it only to children's stories and nonsense rhymes, in which the disassociation of semantics and syntax is evident, and it is not clear how it relates to more realist fiction. Nonetheless, for present purposes, our definition of creativity as the manipulation of forms would allow us to explore how it is used, alongside images, sound, and semantics, to create fictions, much as it is used to signal involvement.[1]

The validity, however, of assuming all language manipulation to be creative, including (as Carter and Tannen do) covert patterning, is questioned on the grounds that it renders the concept diluted and meaningless (Swann 2006: 9). Maybin and Swann (2007), who would not describe all repetition as creative, avoid this by perceiving occurrences of creativity as performances framed by unexpected language forms and marked off from everyday interaction. However, as Carter points out in personal communication to Swann (2006: 9), the issue here is how creativity is defined and valued. If creativity is seen as collaborative and co-creating, building on existing knowledge and within particular contexts rather than as novel inventiveness, then it can be perceived as a common feature.[2] The problem remains however that defining creativity so broadly risks the interesting paradox that it is both unusual or unexpected *and* common (Pennycook 2007: 583-4).[3]

[1] Similar points can be made regarding the hindering of identification of creative practices at the level of discourse (Holmes and Marra 2002).
[2] Swann's (2006) assertion that 'while it [everyday creativity] may have certain interactional effects, it is not creative; however, it may be drawn on in literature to creative effect' suggests an association of creativity with the originality, individualism, genius and invention arises from Western ideas of the individual and of autonomy (Carter 2004b: 26). See section 6.2.1.2 above.
[3] Pennycook describes it as an: 'apparently paradoxical state of affairs whereby the breaking of language norms may be the norm (which undermines the notion of the norm)'.

The argument can nonetheless be made that the decision to label all language manipulation as inherently creative neglects consideration of its functions and how it is perceived by participants (Carter 2004b: 148). Holmes and Marra (2002: 70-73 in Swann 2006: 21-26), for example, identify functions of humour in the workplace through sociocultural clues which indicate how utterances are received and responded to. A sociocultural approach to creativity acknowledges, Swann (2006: 26) concludes, that people in different cultures and contexts may interpret humour differently; and also that participants' perspectives can differ from external researchers (Carter 2004b: 151). This in turn raises questions as to whether features can be defined post-hoc by researchers as creative if, used unconsciously by participants, they pass unnoticed at the time. I would argue that they can, on the basis that although most other linguistic features operate below the consciousness of language users, this does not preclude their functions or significance.[1] The extent to which sociocultural clues can or should be incorporated into predominantly formal investigations is returned to below.

6.2.2.2 Contextual aspects

The contextual level can be understood both in reference to cultural backgrounds shared by participants and to the fact that creativity is seen as arising from interaction in local contexts (Maybin and Swann 2007: 498). The role of shared cultural knowledge is illustrated in the following exchange, in which a hostess apologises for a lack of vegetables.

Figure 6.2 Shared cultural knowledge (Carter 2004b: 100)

<S 01> And so I'm afraid we're a it sort of erm challenged greenwise..
<S 02> Greenly challenged.
<S 01> We're greenly challenged so erm sorry about that.

As Carter points out, creativity here not only draws on British concerns with political correctness and speakers' awareness as to how politically-correct terms tend to be phrased, but also shows how creativity can be jointly created within the context of particular interactions: the second speaker picks up on the first's use of *greenwise* and corrects it to *greenly*, which the first accepts.

[1] A good example of a feature of language escaping native-speaker intuition is semantic prosody (Louw 1993) whereby things that are *caused*, for example, are nearly always bad: *fires, problems* (Stubbs 1994).

The need to explore spoken creativity in terms of its functions in particular contexts emerges from recognition that creativity can only be defined as such if the manipulation of formal properties is both unexpected within the context of the interaction, and appropriate to it (Carter 2004b: 148; North 2007). This raises again the question as to whether identification of language use as creative depends on its interpretation as such by listeners.[1] For North, the impact on other participants online is highly significant, and she distinguishes 'successful humour' as that which builds social cohesion by responding to and being responded to by other participants (North 2007: 548). Participants need not explicitly signal that they find instances of creativity 'coherent' but this must at least become apparent by their smooth insertion into the exchange.

Another implication of the role of context is that the nature and function of creativity may differ across contexts.[2] In North's (2007) investigation of online MSN discussions, for example, she notes that creativity may be encouraged by the particular need in online chat to keep talking (North 2007: 544). Inevitably, creative language occurs more frequently in certain contexts such as intimate discourse, where participants share experiences and ideas within symmetrical relationships (Carter 2004b: 151-164). Online communication, for example, has been described, in North's words, as 'particularly playful', not only because of its textual resources but because time delays allow participants to craft responses and prepare puns (North 2007: 546). The fact that previous messages can be viewed at leisure must also contribute to participants maintaining creative threads. Online creativity is also encouraged, as Carter (2007: 598) suggests, by the need to construe new identities and negotiate interactive practices. As we shall see in CorTxt, however, creativity does not always differ across contexts solely in terms of degree,[3] but also in different ways.[1]

[1] Carter's call for further research into the role of context in shaping creative language use focuses (among other things) on the need to document listener characteristics and contributions as well as those of speakers, despite inherent problems in accessing and evaluating such information (Carter 2007: 606). Doing so, suggests Carter, would reflect the shift in literary studies towards the idea that 'definitions of literature' lie as much with readers as with writers; and the shift from one abstract and timeless *value* to various context-dependent *values* (Carter 2007: 601). The shift means that conversational creativity need not be evaluated in comparison with socially-valued literary works but in terms of its own value as perceived, perhaps, by participants in the interaction (Maybin and Swann 2007: 514). These are important points which challenge or at least contextualise the notion of literary merit and which have implications extending beyond the remit of this thesis.

[2] Carter's (2004) stated aim is therefore to explore creative functions across various social and generic contexts (in and beyond CANCODE), involving different speakers and types of interaction, including IRC (Internet Relay Chat) interaction.

[3] Carter's focus appears to be on 'the extent to which creative language use is evident in different speech genres' (Carter 2004b: 149).

Incorporating context into investigations of creativity is however far from straightforward. Practical problems include accessing and recording contextual factors, especially those relating to participants' intentions and motivations (Carter 2007: 606), as well as initially defining what constitutes context. North (2007: 550), for example, draws on Colby (1987: 452) in suggesting online creativity be understood not only within the context of the current exchange, but within the wider, online experience shaping participants' practices.[2] The issue is arguably more complicated with spoken interaction, in which speakers may draw on various interactions or shared experiences (Maybin and Swann 2007: 514). Similarly, contributors to CorTxt tend to have relationships extending to other mediums of communication including face-to-face (see Chapter 2) which potentially shape their intentions and interpretations when texting. The other issue is that, even where researchers can access various contextual factors, those they foreground may not be those prioritised by participants (Maybin and Swann 2007: 514). Carter (2004b) acknowledges that his conclusions rest on informed speculation over essentially private, contextually-bound exchanges. The present attempt to relate uses of creativity to the nature and affordances of texting similarly acknowledges that it is only one post-hoc interpretation of the interaction.

6.2.2.3 Critical aspects

The third level to creativity encompasses the evaluative or interpersonal functions fulfilled or foregrounded by creativity (Maybin and Swann 2007: 498).[3] Functions which creative language fulfils in everyday interaction were described previously as those not only of fun or amusement but also signalling involvement (Tannen 1989), creating convergence between speakers, indicating a sense of belonging or group membership, displaying identity and offering new perspectives (Carter 2004b) as well as conflict or competition (Cook 2007). According to Maybin and Swann, creativity serves to signal or frame evaluative

---

[1] For example (but see below), the repetition of nouns (but not the determiner they initially occur with) as a cohesive devise serving to make connections between messages given the time delay (section 6.4.2.3) and the ability of texters to comment on their own utterances and, unlike spoken utterances, leave a record of this (section 6.5.5 below).

[2] In fact, North (2007: 545) suggests that in the MSN discussions the only context that exists is that constructed textually and collaboratively by participants. This raises the more fundamental challenge which questions the notion of an external and unchanging context which impacts in observable ways on internal features of the text (Maybin and Swann 2007: 501). Maybin and Swann highlight a contradiction in studies of creativity which simultaneously acknowledge, as Carter does, that creativity varies between interactions depending on shifting purposes, expectations and relationships and yet seeks to explain language forms through fixed contextual factors. However, their challenge also seems to be one they direct more generally at the investigation of large language corpora and lies out of the scope of the current discussion.

[3] In their words, 'the potential for linguistic creativity to foreground, in various ways, the kinds of critical/evaluative stances that are evident in all language use' (Maybin and Swann 2007: 498).

functions or performances, in turn evaluated by the audience. The role that creativity plays in fulfilling these functions (in spoken and texted interaction) is explored in greater depth later in this chapter.

Of significance here, however, is the question as to why manipulation of language patterning in the form it takes is necessary for marking stance and other interpersonal functions. The reason, according to Cook (2000: 137-8), lies in the randomness of the relationship between form and meaning. Forms combine randomly to produce meanings through which users convey information or perform tasks. In certain situations, however, where language is needed primarily to collaborate or compete, the focus is not on conveying content and, in the absence of any need to be factually accurate or meaningful, grammatical or phonological patterns themselves become the focus of concern. This can be seen, for example, in nursery rhymes, pre-fight exchanges, and verbal duelling in the British House of Commons, where 'words are used like the ball in a contest' (Cook 2000: 67). As formal patterns take on greater importance, they are no longer dictated by semantic meanings but become independent concerns: form bypasses semantic meaning to directly produce intended social or interpersonal effects (Cook 2000: 137). Given our observations above regarding different values and interpretations of creativity, the effects must be understood in highly-contextualised terms. However, although Cook's theory can only remain as Swan (2006) puts it, '(informed) speculation' and is not based on wide or systematic data study, it seems particularly relevant to texting where such factors as interpersonal functions, relaxed prescription regarding standard forms, and lack of resources other than the written word combine to encourage a focus on, and manipulation of, form.

6.2.2.4 Summary

Discussion of everyday creativity raises challenging issues not yet fully resolved. However, while discussion of these issues enhances our understanding of creativity, in order to proceed with investigations of this phenomenon, decisions must be made by each researcher regarding how creative language is to be defined. In this chapter, creativity is defined as 'the manipulation of language form in unexpected yet contextually appropriate ways to achieve an intended effect'. This reflects the aspects of creativity identified by Maybin and Swann (2007): textual, contextual and critical; and Carter's (2004: 148) recognition that formal definitions of creativity are insufficient without consideration of

local purposes. However, the definition also allows for creativity to be initially identified and retrieved in terms of formal properties: a method compatible with the focus on data in this thesis.

6.3 Creativity in Texting: setting up the study

6.3.1 Data and methods

One strength of the present study lies in its basis in naturally-occurring data. Other linguists take different approaches, shaped by their purposes and theoretical starting-points. As mentioned previously, Cook (2000) and Crystal (1998), despite real differences in their treatments of language play,[1] share a concern with prepared and recited language play such as jokes and nursery rhymes, rather than spontaneous context-bound instances emerging through interaction. As such, their data is limited largely to published or well-known sources and, in Crystal's case, anecdotal evidence. In contrast, Tannen (1989) draws on recorded conversations between herself and groups of friends and acquaintances. Social characteristics of her participants are, however, limited in range and the type of interaction largely small, friendly private gatherings. Use of CANCODE to support his arguments is therefore one of the greatest strengths of Carter's (2004) work (Swann 2006: 9). As well as the benefits of size, CANCODE is organised by context type and interaction type, labelled where necessary in terms of the conversation's main activity. Unlike Tannen, Carter (2004b) is able to explore creative functions across social and generic contexts, involving various speakers and interaction types.[2] As Swann (2006) notes, data in CANCODE is mainly limited to informal talk in non-institutional settings although Carter (2004: 163-4) includes workplace and business contexts from outside CANCODE.

The current study draws in a similar fashion on a corpus.[3] In terms of interaction types within CorTxt, the purposes to which text messages are put are chiefly, according to sociological and linguistic research, informal, interpersonal, and typical of spoken

[1] Crystal's is a populist account of the significance and prevalence of language play; Cook's a theoretical discussion situated in the wider context of play.
[2] However, again in contrast to Tannen, although it is clear as Swann (2006) points out that interactions between both women and men are included, Carter gives no information about the gender balance in his sample, nor about other social factors such as class and ethnicity. It is not therefore possible to consider social, cultural and contextual factors that might affect the types of creativity people engage in and their interactional effects, and this could be seen as limiting the generalisability of Carter's claims.
[3] Although no attempt is made to suggest that CorTxt and CANCODE are comparable in size, range or in terms of the sophistication of their annotation: nor that the achievements of the studies based on them are comparable.

interaction between intimates.[1] Although CorTxt is not divided by context or interaction type, it would be possible through analysis of text message content to identify main activities within exchanges and thus attempt to match uses of creativity to specific communicative functions and contexts. For the purposes of the current study, however, the corpus is explored as a whole and creative features matched where possible to functions and affordances of the medium.

Identifying and classifying creativity remains a contentious issue, as Carter observes. Given the impossibility of retrieving creative features automatically through surface formal features (Carter 2004b: 148),[2] creativity in CANCODE is retrieved by systematically dividing the corpus into cells and by 'reading' rather than frequency counts and concordance lines.[3] Limitations include the impossibility of quantification, despite Carter's claim to have combined qualitative with quantitative methods. Although he suggests that creativity 'leaps out at researchers from almost every transcript of the data' (Carter 2004b: 90), selected examples can only be, as he acknowledges, 'taken on trust' and there are unavoidably few replicable figures.[4] However, given the continuing relative lack of access to 'intimate discourse', as highlighted by Cook (1994), the fact that such examples are identified at all is of great value and there is no inevitable need for studies of creativity to be quantitative. The great advantage of Carter's text-driven approach, in comparison to corpus-based methods, is the sense that Carter's investigations are led as much by data as by preconceived theory; and, in comparison to corpus-driven methods, that features he uncovers are not retrievable through quantitative methods.

Methods used to identify creativity in CorTxt are similarly qualitative, and examples of creativity accessed by reading. The fact that creative instances were primarily identified through formal characteristics can also be defended on the practical grounds that creativity

[1] See Chapter 2: Research into Text Messaging.
[2] My attempt to find creativity automatically from CorTxt through searching for metalanguage (*literally, I mean*), intensifiers (*indeed, absolutely*), vague language (chiefly, *a bit*) and even conjunctions (*and*) was largely unproductive.
[3] An approach described in Chapter 3 of this thesis as a *text*-driven approach.
[4] The same also applies to Tannen (1989). Tannen supports each observation she makes with an example from her own data, but whether it is one of many such examples or an isolated occurrence is difficult to ascertain. The nearest we get to an overview of Tannen's larger bank of data is her statement that 'not all transcripts show a high percentage of repeated words but many do, and all show some'. Other studies, such as Carter (2004: 79-80) and Crystal (2003: 413) reproduce Tannen's oft-quoted example of a discussion on 'eating' in which phrases repeat, but the fact that one stretch of discourse fits the pattern can hardly be very convincing.

in texting is arguably less easily identifiable than in speech through sociocultural clues:[1] responses such as laughter or groaning may occur verbally yet be omitted from text messages themselves. At the same time, the (limited) contextual knowledge of this researcher also serves to identify puns or shared jokes. As with other studies of creativity, identification and interpretation therefore relies on speculation, albeit cautious and informed by CorTxt and by previous research into everyday creativity.

6.3.3 Outline of the investigation

This investigation is structured according to categories of creative language identified by Carter (2004b). Creative patterns of various kinds, Carter (2004: 101-9) observes, potentially form and can be traced throughout everyday language, and choices motivating these patterns fall into two categories: *pattern-forming* structures such as repetition and *pattern re-forming* choices as seen in puns. Although the two categories may combine, the following investigation begins by exploring pattern-forming choices, before moving on to pattern re-forming choices. For each, their use and function in spoken interaction is explored first (Carter 2004b; Tannen 1989; North 2007), before turning to investigation of CorTxt.

6.4 Pattern-forming choices in texting

6.4.1 Pattern-forming in spoken language

Research by Carter (2004b) and Tannen (1989) forms the basis of this investigation of pattern-forming creativity. Carter's pattern-forming choices correspond roughly to the use of repetition described by Tannen (1989) as an 'involvement strategy' and their work describes repetition as prevalent in everyday conversation.

The significance of repetition has long been recognised, although neglected somewhat by twentieth-century concerns with our ability to create novel sentences (Chomsky 1957). Nonetheless, descriptions by linguists such as Bolinger (1961) and Hymes (1981) of repetition as part of the structure or grammar of language (cited in Tannen 1989), have since been corroborated by corpus investigation which draws attention to repetition in the sense of frequently-occurring phrases and collocations, as well as by in-depth studies of

[1] See the more theoretical defence in sections above.

spoken interaction (McCarthy 1988; Tannen 1989). According to Tannen (1989: 97), repetition is 'a limitless resource for individual creativity and interpersonal involvement', underlying other involvement strategies: rhythm, figures of speech, indirectness, ellipsis, tropes, narrative. According to Cook (1994, 1997), the significance of repetition is such that it needs to be drawn on in language teaching.

Defining prepatterned language, formed through repetition, is challenging.[1] The focus in the present investigation is on repetition within the confines of the exchange being studied: that is, where words or phrases are repeated within that, or the next, turn. Distinctions made by Tannen include whether repetition is 'self-repetition' or allo-repetition' (the repetition of others across turns), whether it involves repeating phonemes, morphemes, words and phrases or larger discourse structures, and its position on a scale of fixity from exact repetition to paraphrase. Repetition with variation, such as the following, is most common (Tannen 1989: 54).

Figure 6.3 Repetition with variation in pronouns (Carter 2004b: 107)

<S02> [laughs] cos *you come home*
<S03> *I come home*
<S02> *You come home* to us

Another category of current relevance is Tannen's (1989) 'patterned rhythm',[2] which serves to create intensifying, list-like intonation as seen in the following description of pandemonium in a railway station.

Figure 6.4 Patterned rhythm (Tannen 1989: 67)

six thousand, five thousand, six thousand, ten thousand people come in

The patterned rhythm, Tannen argues, not only stresses the number of people involved but also, through the randomness of the numbers, the degree of chaos caused.

[1] According to Tannen, it encompasses a range of language forms including fixed expressions or formulaic language passed down the generations; coinages that do not last beyond a particular exchange; and words or phrases repeated only within that, or the next, turn: a question, for example, which is repeated then answered. See also Wray 2002, 2008 and Chapter 7of this thesis.
[2] '[I]n which completely different words are uttered in the same syntactic and rhythmic paradigm as a preceding utterance' (Tannen 1989: 63).

The specific functions of such repetition in conversational discourse are broadly categorised by Tannen (1989) into five areas. As well as functioning to facilitate language *production* and *comprehension*, repetition is involved in *connection*, showing how utterances and ideas are linked to earlier discourse, as well as playing an evaluative role through its indication of emphasis, as in the following.

Figure 6.5 Evaluative role of repetition (Tannen 1989: 50)

And he knows Spanish,
And he knows French,
And he knows English,
And he knows German,
And HE is a GENTleman.

As Tannen suggests, the repetition of the frame: *and he knows* intensifies the repeated part and makes his achievement sound impressive; while its divergence from the pattern intensifies the final sentence. As well as binding participants together, repetition also binds participants to the discourse: it facilitates *interaction*. This, according to Tannen, includes getting or keeping the floor, providing back-channel responses or showing listenership, as the following example illustrates.

Figure 6.6 Repetition used to show listenership (Tannen 1989: 61-2)

Deborah: It's expressed in language.
Chad: It's expressed in language
David: It's expressed in language.

Other functions include gearing up to answer or speak, including someone who did not hear a previous utterance and humour, as in this echoing of grammatical frame.

Figure 6.7 Repetition and humour (Tannen 1989: 63)

Deborah: Rover is being so good.
Steve: I know.
Peter: He's being hungry.

These functions combine to fulfil the larger function of creating involvement (Tannen 1989: 52).

6.4.2 Repetition in Texting

6.4.2.1 Overview

Repetition appears to be used in texting for some functions identified in speech: to create involvement by achieving certain evaluative effects otherwise difficult to convey through this restricted medium (intensification, for example); and by linking responses to previous text messages (Tannen's function of connection). However, unlike spoken interaction, there is less evidence that texters use repetition to facilitate production, perhaps because while speech occurs in real time, texting is subject to delay, and participants have time to plan and to edit what they produce as well as being limited by space rather than time constraints. Similarly, in managing texted interaction, certain concerns typical of spoken interaction can be avoided: getting and keeping the floor or stalling or gearing up to answer or speak, and repetition does not appear to be used for these functions. Instead, in contrast to research into spoken data, this investigation also reveals repetition of style and text message structure. These observations suggest that repetition in texting differs in some respects from that of spoken interaction due to contextual differences between the two means of communication.

6.4.2.2 Self-repetition for evaluation

Repetition of grammatical structure or syntactic frame in texting appears to fulfil two main functions: self-repetition plays an evaluative role while allo-repetition (the repetition of others) shows listenership and connects speakers' ideas, as well as being evaluative. In this example of self-repetition, we see repetition of the syntactic frame *with the stick*.

(6.1) Am watching house – very entertaining – am getting the whole hugh laurie thing – *even* with the stick – indeed *especially* with the stick.[1] [2]

Repetition of the frame 'with the stick' emphasises *especially*, the word slotted into the frame and, as such, it intensifies the texter's evaluative stance on the subject. The intensifier 'indeed' emphasises the structural parallelism of the two phrases, and highlights the way in which the effect of the second part builds on or exploits the first. In other words, *even with the stick* is not a mistake which should be corrected to *especially with the stick* but, instead, the second adds to the first to create the overall intended meaning: that the texter finds

[1] All examples of data are taken from my corpus, CorTxt, unless otherwise referenced.
[2] This is in reference to the medical drama *House*, in which Hugh Laurie plays a doctor with a limp and a walking-stick.

particularly attractive something which should apparently detract from Hugh Laurie's charm. The syntactic parallelism in this example illustrates the apparent role for repetition in indicating stance despite (or because of) the constraints of the medium. Similar examples include the following.

(6.2) I think I'm waiting for the same bus! Inform me *when* you get there, *if* you *ever* get there.NAME14 x

(6.3) They can try! They can get lost, *in fact*. Tee hee

Repetition of the frame *you get there* in the first appears to highlight the texter's uncertainty over whether their interlocutor will arrive, and thus their lack of faith in the bus system, by emphasising the use of *if* in contrast to *when*, and conveying the idea in a relatively short phrase. In the second, *They can try!* can be interpreted as meaning 'they can try but I don't like it and/or they won't succeed' which seems to inspire the stronger *They can get lost*, the strength of which is emphasised by its being slotted into the same grammatical frame as the previous, phrase. Other examples are as follows.[1]

(6.4) I jus hope its true that NAME281's missin me cos i'm *really* missin him! You haven't done anything to feel guilty about, yet. Xx

(6.5) Just re read it and I have no shame but tell me how he takes it and if he runs I will blame u 4 ever!! Not *really* 4 ever just a long time xx

The following examples of 'patterned rhythm' have a similarly intensifying function as the above.

(6.6) Received, understood n acted upon!

(6.7) Miss ya, need ya, want ya, love ya.

[1] But are not commented on.

Texters' repetition of syntactic frame and their use of patterned rhythm to evaluate statements through intensification or emphasis highlights an apparent role for repetition in indicating stance through a highly interpersonal medium which can in many other ways be seen as constricted.

6.4.2.3 Repetition across turns: listenership and connection

Repetition across turns appears to link utterances and ideas to earlier discourse, or of one speaker's ideas to another, and in the process often shows listenership as well as playing an evaluative role by providing emphasis. Potential confusion arising where phrases or words are *not* repeated in responses (in order to show which part of a previous text message is being responded to) is arguably greater than you might expect in synchronous spoken exchanges and is illustrated by the lack of repetition in the following exchange.[1]

(6.8)	A:	Have you got Xmas radio times. If not i will get it now
	B:	No. Yes please. Been swimming?
	A:	Okay. No no, just shining on. That was meant to be signing, but that sounds better

Examples in which phrases are repeated across turns include the following.

(6.9)	A:	Can you use foreign stamps for whatever you send them off for? Xx
	B:	Can not use foreign stamps in this country. Good lecture x.

(6.10)	A:	Cool, we shall go and see, have to go to tip anyway. Are you at home, got something to drop in later? So lets go to town tonight! Maybe mum can take us in.x
	B:	Okey doke. I'm at home, but not dressed cos laying around ill! Speak to you later bout times and stuff. Xx

(6.11)	A:	Hiya, had a good day? Have you spoken to NAME76 since the weekend? Xx
	B:	Had a good day – hope clinic went well, going 2 a lecture on Freud later. Have texted from NAME76 but did not hear anything late yesterday so thought he did not want 2 no – got mixed feelings about him – what do u really think? X x

[1] Which, of course, also shows that text messages do not always rely on repetition.

(6.12)	A:	Hello! How r u? Im bored. Inever thought id get bored with the tv but I am. Tell me something exciting has happened there? Anything! =/
	B:	Not a lot has happened here. Feels very quiet. Beth is at her aunts and charlie is working lots. Just me and helen in at the mo. How have you been? X

(6.13)	A:	I am back. Bit long cos of accident on a30. Had to divert via wadebridge. I had a brilliant weekend thanks. Speak soon. Lots of love xxx
	B:	I had a good time too. Its nice to do something a bit different with my weekends for a change. See ya soon xxx

The last example in particular also illustrates the role of repetition in showing listenership by linking the second texter's sentiment to that of the first, perhaps to make clear that they are reciprocating in kind.

The need to connect ideas between text messages also means that, rather than pronouns being used as a cohesive device, nouns mentioned in a first text message are often repeated in the second but the article or determiner is dropped: perhaps for reasons of brevity or speed. Examples include the following.

(6.14)	A:	Aah bless! How's your arm?
	B:	Arms fine, how's Cardiff and uni? xx

(6.15)	A:	Have you heard about that job? I'm going to that wildlife talk again tonight if u want2come. Its that2worzels and a wizzle or whatever it is?! Xx
	B:	Hello. No news on job, they are making me wait a fifth week! Yeah im up for some woozles and weasels... In exeter still, but be home about 3. X

(6.16)	A:	I don't quite know what to do. I still can't get hold of anyone. I cud pick you up bout 7.30pm and we can see if they're in the pub? Xx
	B:	Thats a bit weird, even NAME281?- where is the do supposed to be happening? But good idea, sure they will be in pub! X

(6.17)	A:	Hey there! Glad u r better now. I hear u treated urself to a digi cam, is it good? We r off at 9pm. Have a fab new year, c u in coupla wks! X
	B:	Camera quite good, 10.1mega pixels, 3optical and 5digital dooms. Have a lovely holiday, be safe and i hope you hav a good journey! Happy new year to

you both! See you in a couple of weeks! Xx

What is interesting about the above is that the repetition is not particularly speech-like. In face-to-face interaction, it could be assumed that ellipsis, pronouns or other anaphoric references (such as *there* instead of *in pub* and *It's* instead of *Camera* in the above two examples) would be used.[1] Noun (and structural) repetition in texting connects utterances in a second text message with relevant parts of the preceding text message, encouraged perhaps by the asynchronicity of texting, as well as the fact that, unlike MSN discussions, previous text messages are not available for reflecting on while the next are written.

Other examples of repetition across turns explicitly combine a role in connecting text messages with an evaluative function, which again (as with self-repetition) highlights the need for such language manipulation given the lack of paralinguistic and other resources. In each of the following, the second texter responds to the first by repeating a phrase and then using the same syntactic frame to extend the sentiment in an emphatic way.[2]

(6.18)

A:	Thanks honey but still haven't heard anything I will leave it a bit longer so not 2 crowd him and will try later - great advice thanks xxx hope <u>cardiff is still there</u>! Xxx
B:	Sounds like a plan! <u>Cardiff is still here</u> *and still cold*! I'm sitting on the radiator!

(6.19)

A:	Hi hope u r both ok, he said he would text and he hasn't, <u>have u seen him</u>, let me down gently please xxx
B:	Hiya, sorry didn't hav signal. <u>I haven't seen</u> *or heard* from NAME79 and neither has NAME281, which is unusual in itself! I'll put NAME281 on the case and get him to sort it out! Hugs and snogs. Xxx

(6.20)

A:	Did he say how fantastic I am by any chance, or anything need a bigger life lift as losing the will 2 live, do you think <u>I would be the first person 2 die from N V Q</u>? XXX
B:	He said that he had a right giggle when he saw u again! <u>You would *possibly* be the first person2die from NVQ</u>, *but think how much you could NAME330 for*! Xx

[1] The occurrence of un-speech-like ellipsis and the relative infrequency of place adverbials is explored further in Chapter 8: Spoken Grammar, Texted Grammar.
[2] The repeated phrase is <u>underlined</u> in each example, and the extension *italicised*.

(6.21)	A:	Yeah that sounds good, even the dancing! I'm not too bad. You? NAME209 has been offered some experimental chemo drug that might giv him a yr, wivout it he's probably got3months. It's terrible, i can't imagine what it's like2be told something like that! Xx
	B:	Shit that is really shocking and scary, cant imagine *for a second*. Def up for night out. Do u think there is somewhere i could crash for night, save on taxi?

In most of these, repetition occurs with variation in terms of deictic reference (*hope Cardiff is still there? – Cardiff is still here!*) and pronoun (*have u seen him – I haven't seen or heard from NAME79*) as well as sentence mode (generally from interrogative to declarative) and in terms of spelling and abbreviation (*the first person 2 die – the first person2die; can't imagine – cant imagine*).

Repetition of lexis, as well as grammatical structure, is illustrated by the following.

(6.22)	A:	All done? All handed in? Celebrations in full swing yet?
	B:	All done, all handed in. Don't know if mega shop in asda counts as celebration but thats what i'm doing!

Transformation of grammatical structure from two initial questions into statements serves not only to clarify which of A's three questions B is responding to, but also shows that B is responding fully to A's question: B is 'listening'. Similarly, repetition of the lexical item *celebration* shows how B's otherwise somewhat irrelevant description of her shopping trip relates to A's initial query, and indicates a touch of irony.

Another example of lexical repetition is the following.

(6.23)	A:	How have your little darlings been so far this week? Need a coffee run tomo?Can't believe it's that time of week already ...
	B:	All was well until slightly disastrous class this pm with my fav darlings! Hope day off ok. Coffee wld be good as can't stay late tomorrow. Same time + place as always?
	A:	That's the trouble with classes that go well - you're due a dodgey one … Expecting mine tomo! See you for recovery, same time, same place xx

B:	See you there! Exx

Lexical repetition in this example occurs in the transformation of *your little darlings* in the first text message to *my fav darlings* in the second. The lexical switch from *little* to *fav* serves, on the surface, to positively evaluate the *darlings* although the context suggests some irony. Repetition of *darlings*, however, also serves to link the response back to the initial query, in the same way as the repetition of *coffee* links the response back to *a coffee run*. In the third text message, *well* links A's response back to B's claim that *All was well*, and the sign-off *same time, same place* echoes B's question: *Same time + place as always?* A's *see you* is also echoed by B in the final text message. This lexical repetition signals texters' involvement with the other and with the interaction, as well as fulfilling the practical function of linking responses to previous text messages.

A related and final phenomenon is how repetition of a phrase can sometimes serve to present it as a concept which can then be evaluated, through its transformation from a clause or verb into a noun. This can be seen in the following.

(6.24)	A:	... Are you in the pub?
	B:	sorry, no, have got few things to do. may be in pub later.
	A:	I like to think there's always the possibility of being in a pub later.

In the above, discussion of whether or not texter B is in the pub becomes a discussion of the concept of *being in a pub later* largely through repetition of the phrase, and this allows evaluation: in this case, an apparently positive view of being in a pub. A similar process occurs in the following through repetition of the adjective *cheap* and its transformation into a noun.

(6.25)	A:	Well, I was about to give up cos they all said no they didn't do one nighters. I persevered and found one but it is very cheap so i apologise in advance. It is just somewhere to sleep isnt it?
	B:	I like cheap! But i'm happy to splash out on the wine if it makes you feel better..

To sum up, repetition in texting is used both to evaluate and, where it occurs across turns, to indicate connections between text messages. The way in which utterances are linked

through repetition is in some respects non-speech-like and thus shows how creativity is adapted to fit demands or affordances of texting. Meanwhile, the evaluative role played by repetition supports the suggestion that everyday creativity acts as a resource for the expression of emotion and stance within an otherwise constrained medium.

6.4.2.4 The imitation or repetition of style

The role of repetition in shaping texting style (in that responses imitate the style established in initial text messages) can be explained by the fact that the medium is new and evolving, as well as by its interpersonal nature and its role in delimiting group boundaries. The following playful example illustrates the tendency for texters to adopt interlocutors' styles, in this case a formal, old-fashioned style.

(6.26)		
	A:	good evening to you mrs NAME33! here is me, NAME298! i contact u with solemn news which may be difficult to bear! it appears, from my end, that birthday beverages will have to proceed in my absence. i truly regret this […]
	B:	Greetings me, NAME298! Consider yourself excused.
	A:	[…] anyway, many good evenings to u! s
	B:	And several to you sir. Xx

The formality and dated style of the first text message can be seen in the address and the vocabulary choices: *contact*, *solemn*, *bear*, *beverages*, *proceed*, *absence* and *regret* as well as the use of fully-formed, elaborated or unabbreviated sentences: *i contact u with difficult news that may be difficult to bear*. The second texter picks up on these, responding with the old-fashioned greeting *Greetings* and the address *sir*, and the formal *Consider yourself excused*. Other aspects, however, show the varied and evolving nature of texted style: despite the general formality, A chooses *u* over *you* and includes no capitals, and B ends the text message with *Xx* (and no signature). The suggestion made here is that texters are still negotiating how texts should be written, and a definitive style has yet to emerge: in the meantime, texters may be encouraged to pick up on their interlocutor's style. A further illustration of this can be seen in the following example.

(6.27)	A:	Can you do a mag meeting this avo at some point?
	B:	What time. I'm out until prob 3 or so
	A:	4 oclock at mine. Just to bash out a flat plan.
	B:	Ok cool. See ya then.

In the above, the style of the first text message is brief and functional and the following text message adopts a similar tone. At the same time, both texters retain some speech-like informality: *mag* is echoed by *prob*, and *bash out* by *Ok cool*. *See ya then*; and both use numerals and abbreviation: *avo*, *3*, *4 oclock*. In other words, the texters share a similarly mixed style.

One way in which style is imitated lies in the nature and extent of abbreviation, phonetic spellings and colloquial contractions. The following, for example, show a similar elision of first-person pronouns and other initial elements by both texters (and the echoing of the colloquial contractions *wanna* in the first and *cos* in the second).[1]

(6.28)	A:	Forgot you were working today! Wanna chat, but things are ok so drop me a text when you're free / bored etc and i'll ring. Hope all is well, nose essay and all xx
	B:	Am not working but am up to eyes in philosophy so will text u later when a bit more free for chat...
	A:	S'fine. Anytime. All the best with it.
	B:	Well am officially in a philosophical hole, so if u wanna call am at home ready to be saved!

(6.29)	A:	Are we doing the norm tomorrow? I finish just a 4.15 **cos** of st tests. Need to sort library stuff out at some point tomo - got letter from NAME583 today - access til end march so i better get move on!
	B:	Yes but can we meet in town cos will go to gep and then home. You could text at bus stop. And don't worry we'll have finished by march … ish!
	A:	Ok. Not sure what time tho as not sure if can get to library before class. Will try**.** See you at some point! Have good eve.
	B:	And you! Will expect you whenever you text! Hope all goes well tomo xx

[1] Although, as with the other examples, the extent to which this is inspired by their interaction, rather than being simply the personal style adopted by both or a reaction to the confines of texting is not clear.

It is interesting to note that in the less functional segment of the latter exchange, the first-person pronoun is retained: *access til end march so i better get move on!* met by the playful: *And don't worry we'll have finished by march … ish!*

4.2.5 Echoing text message structure

Message structure, another area open to repetition, is perhaps most clearly seen in its closing, where 'sign-offs' are often echoed in subsequent text messages.

(6.30)	A:	I am back. Bit long cos of accident on a30. Had to divert via wadebridge. I had a brilliant weekend thanks. Speak soon. Lots of love xxx
	B:	I had a good time too. Its nice to do something a bit different with my weekends for a change. See ya soon xxx

As with the choice of syntactic frame, it could be argued that the echoed sign-off is an attempt to reciprocate sentiment. Other examples include the following.

(6.31)	A:	What time you coming down later? xxx
	B:	It wont b until 2.15 as trying 2 sort house out, is that ok? Xx
	A:	Yeah, that's fine! It's £6 to get in, is that ok? Xx
	B:	Yes fine xx

(6.32)	A:	Hey there! Glad u r better now. I hear u treated urself to a digi cam, is it good? We r off at 9pm. Have a fab new year, c u in coupla wks! X
	B:	Camera quite good, 10.1mega pixels, 3optical and 5digital dooms. Have a lovely holiday, be safe and i hope you hav a good journey! Happy new year to you both! See you in a couple of weeks! Xx

(6.33)	A:	Hello! How r u? Im bored. Inever thought id get bored with the tv but I am. Tell me something exciting has happened there? Anything! =/
	B:	Not a lot has happened here. Feels very quiet. Beth is at her aunts and charlie is working lots. Just me and helen in at the mo. How have you been? X
	A:	*Ok*. Not much to do here though. H&M Friday, cant wait. Dunno wot the hell im gonna do for another 3 weeks! Become a slob- oh wait, already done that! X

B's *Speak soon!* below could be seen as a way of showing that they are now ready to end the exchange, by echoing A's own (ignored) earlier closing.

(6.34)		
	A:	I have had two more letters from NAME187. I will copy them for you cos one has a message for you. Speak soon. Xxxxx
	B:	Goodo! Yes we must speak friday - egg-potato ratio for tortilla needed! Xx
	A:	Okay but i thought you were the expert
	B:	Yes obviously, but you are the eggs-pert and the potato head… Speak soon! Xx

A similar process occurs in the following.

(6.35)		
	A:	I have 2 sleeping bags, 1 blanket and paper and NAME187s phone details. Anything else?
	B:	I think just yourself … Thanks and see you tomo xx
	A:	And maybe some pressies
	B:	Maybe?! Say hi to NAME150 and find out if NAME247 got his card. Great escape or wetherspoons? Xx
	A:	Great escape. I fancy the bridge but NAME150 needs her lager. See you tomo x

6.4.3 Summary of repetition in texting

The above overview shows how texters exploit pattern-forming structures found in spoken interaction and adapt them to demands and constraints of texting. In other words, the way in which language is manipulated differs between spoken and texted interaction, due to distinct demands and affordances. As an asynchronous means of communication, texting draws on repetition in linking ideas or responses to those in previous text messages in a somewhat non-speech-like way. Furthermore, as texting is relatively new and evolving, with prescriptive language rules to some extent held in abeyance, the style and structure chosen by texters may be influenced by that of interlocutors and this, perhaps, has a similar effect of achieving convergence. Finally, repetition also plays an important role in the creation of intensifying or evaluative effects otherwise difficult to achieve, given the limited resources available to texters: a significant observation returned to in the following section.

6.5 Pattern re-forming choices

6.5.1 Pattern re-forming choices in spoken language

6.1.1 Definitions and descriptions

Pattern re-forming choices described by Carter (2004b) correspond more closely (than do pattern-forming structures) to language play described by Crystal (1998) and Cook (2000), as well as to typical notions of creativity. This section outlines Carter's definition of this overt creative pattern and shows how he relates its impact to that of non-core vocabulary and idiomatic language, all of which operate along clines of intimacy, intensity and evaluation. Four types of pattern re-forming choices are introduced and structure the investigation of texting.

Pattern re-forming choices, suggests Carter (2004: 102), create patterns which draw attention to themselves because, unlike pattern-forming choices, they break with expected, conventional patterns. They are therefore more overt, deliberate and transparent than pattern-forming choices. Although like pattern-forming structures they create common viewpoints, sometimes jointly co-produced and often adopted by interlocutors, pattern re-forming choices may involve novel, innovative change introduced by single speakers. This kind of creativity is the focus of North's study of online discussions, where creativity is defined as: 'novel and unexpected uses of language that are crafted to be relevant to a constantly changing context' (North 2007: 540). As well as prompting laughter, which is Crystal's (2000) focus, overt instances of creativity indicate intimacy, express evaluative and affective viewpoints (just as pattern-forming structures do), and change how participants view the world (Carter 2004b). For example, although the focus of North's study is primarily 'spontaneous conversational humour' (North 2007: 540), she goes on to show how 'successful' humour provides social and textual cohesion.

The notion of core and non-core vocabulary is central to understanding how creative pattern re-forming choices achieve the above effects (Carter 2004b: 119).[1] As Carter points out, non-core vocabulary can express greater intimacy, intensity and evaluation than more expected, simple lexical choices, which he illustrates with how the use of *pop off* by care assistants in describing the death of residential home residents increases the group's

[1] Core vocabulary constitute the basic, normal, simple words of a language, which can be identified through means of criteria such as their often being superordinates, having clear antonyms and weak collocational patterns and connotations.

intimacy and familiarity; and how ‘You didn’t just win. You annihilated us’ after a sports game strengthens intensity (Carter 2004b: 117). Choices of figures-of-speech over core vocabulary, as Carter points out, similarly indicate intimacy, intensity and evaluation: *floods of tears*, for example, intensifies the extent to which someone cries. As McCarthy and Carter (1994) put it, idioms do not simply describe but comment in positive or negative ways. The association of non-core and idiomatic language with intimacy, intensity and evaluation is not, however, straightforward: it is, for example, possible for core vocabulary itself to take on evaluative, intimate or intense meanings.

Creative patterns listed below nonetheless work in a similar way, injecting fresh meaning into existing idioms, metaphors or words. Manipulated idioms, for example, have more affective impact than canonical versions due to their unexpectedness. They may also be more contextually relevant, as speakers respond to contextual demands.

The following forms of pattern re-forming creativity occur in spoken interaction.

Figure 6.8 Types of pattern re-forming creativity (Carter 2004b: 109)

- Metaphor invention or extension;
- Speaker displacement of fixedness, particularly idioms and formulaic phrases;
- Morphological inventiveness; and
- Verbal play, punning and parody through overlapping forms and meanings.

Examples of these emerge in other studies of creativity (North 2007) and, as explored below, in texting. However, differences between the spoken and texted mediums allow for creativity of a different nature to occur in the latter. A fifth pattern that texters draw on, for example, can be termed *metacomment*, by which texters reflect on or reinterpret previous utterances.

The following investigation of CorTxt explores the five areas in turn. As evident in examples, these pattern re-forming structures rely on immediate context: not only on previous turns but on events taking place and participants’ shared knowledge.

6.5.2 Metaphor extension

6.5.2.1 Linguistic definition of metaphor

Attempts to explore metaphor are complicated by continuing debate over the nature of metaphorical language. Given the scope and focus of this study, the following discussion does not claim to do justice to the issues but instead seeks to secure an operational definition of metaphor extension reflecting that used in other studies of everyday creativity. Metaphors are traditionally identified in linguistic terms as a disjunction between two otherwise unrelated areas: the *Topic* (subject matter) and *Vehicle* (word or phrasing used), resulting in images such as ships ploughing through waves (Cameron 2006: 46). In Carter's (2004: 119-120) words, metaphors are 'analogies'. However, with the shift in the 1980s from literary to cognitive perspectives of metaphor (Lakoff and Johnson 1980), came the understanding that we habitually understand things in terms of other things and that this way of perceiving the world is embodied in systematically-organised metaphors. In other words, metaphor need not be deliberate, conscious or noticed. This raises questions as to 'where metaphor stops' (Cameron 2006: 51), which in turn raises questions as to how metaphor can or should be defined. However, the focus of the current study is not conventionalised metaphors, unless they are 'crafted to be relevant' (to take North's 2007: 540 expression) to the particular context. The assertion that conventionalised metaphors are not of themselves 'creative' is based on a similar premise as North's (2007: 542) suggestion that well-established emoticons and acronyms in online communication are no longer creative *per se*. However, neither are wholly novel metaphors expected to emerge from everyday conversation (Cameron 2006: 47) or electronic communication. The primary focus of this study is therefore the extension of existing metaphors in contextualised ways, reflecting Carter's (2004: 119-125) choice and discussion of metaphors from CANCODE. Carter (2004: 124) illustrates the functions of metaphor extension in creating intimacy, intensity, and evaluation[1] with *elephant bottom* which works simultaneously on the three clines to informally exaggerate and evaluate the bottom in question (Carter 2004b: 124).

[1] This affective role of metaphor, according to Cameron (2006: 50), is 'more important in discourse than is commonly realised': creating 'insider' language; saving face by, for example, injecting humour; and creating an atmosphere of warmth and support: all of which could be said to signal intimacy. To take another of Cameron's examples, metaphors are evaluative in the sense that they provide new insights or encourage the adoption of a particular perspective on the Topic of the metaphor; and are thus used wherever participants in interaction have communication problems to solve, such as when explaining new concepts or expressing ideas for which the actual word is not known.

Problems must be acknowledged. Firstly, distinguishing extended, contextualised metaphors from dead or conventionalised ones is complicated by the fact that what constitute conventionalised metaphors varies between socio-cultural group (Cameron 2006: 48). Post-hoc analysis by external researchers arguably reveals metaphors which extend ones *they* recognise as conventionalised, although their meaning within the context can be explored through their impact on participants.[1] In Carter's example, for instance, laughter accompanies observations regarding the speaker's *elephant bottom*. Rather than drawing a line between the conventionalised and novel, this suggests metaphors require different processing loads dependent in part on shared contextual and background knowledge. Understanding *elephant bottom*, suggests Carter (2004: 123), would generally require less processing than *genetic roulette* as the associative link between genetically-modified food and roulette requires knowledge of the scientific issues. The implication is that once metaphors become conventionalised in a community, little or no processing is required in order to understand the analogy, which correspondingly reduces its affective impact.

The danger here is that of circularity: that metaphors identified through reference to their function then form the basis of investigation into functions fulfilled by metaphor. Researchers' identification of metaphor must therefore be driven by forms which appear to have the potential for metaphorical processing (Cameron 2006: 47). However, linguistic resources Cameron (2006) suggests are exploited in English to create metaphor cannot reliably be used to identify extended metaphor.

6.5.2.2 Metaphor Extension in Texting

Extension of metaphors in texting, bearing in mind the difficulties involved in identification, ranges from contextualised root analogies through extended metaphor to seemingly more novel metaphorical reference. In the following, the metaphorical use of 'ploughing' is applied to housework, illustrating both the extent of the task (heavy and physically challenging) and the texter's approach (determined).

(6.36) Hope ur head doesn't hurt 2 much ! Am ploughing my way through a pile of ironing ! Staying in with a chinky tonight come round if you like.

[1] Although this raises similar issues to those discussed in sections 6.2.2.1. and 6.2.2.2 above regarding the use of sociocultural clues.

The extent to which the texter is aware of their metaphorical language is unclear, but it seems fair to suggest that use of *plough* rather than *work* has a more intense and attitudinal effect, an effect perhaps difficult to convey through more core vocabulary given space constraints and lack of paralinguistic resources. The following similarly invoke intense and emotional stances, through entrenched, physically vivid metaphors.

(6.37) on NAME461 hen night. Going with a swing

(6.38) Yes watching footie but worried we're going to blow it - Phil Neville?

(6.39) It's justbeen overa week since we broke up and already our brains are going to mush!

In the following, metaphorical reference is extended (in each case by adding the underlined word) to fit the particular situation: the writing of an essay on philosophy. The texter's sentiment is intensified through the exploitation and contextualisation of vivid metaphor.

(6.40) Well am officially in a philosophical hole

(6.41) am up to my eyes in philosophy

Another example of metaphor extension is the measurement of time in *sleeps*:

(6.42) I to am looking forward to all the sex cuddling.. Only two more sleeps x

(6.43) I do know what u mean, NAME281 is the king of not havin credit! I'm goin2bed now. Night night sweet! Only1more sleep! Xx

The first is part of an exchange between a couple who have only two more lonely nights to endure before their reunion. The texter measures this time in *sleeps*, disrupting expectations of how time is measured and suggesting that it is the number of times they must sleep alone that mark the other's absence, thus making the sentiment more intense and intimate. This is reminiscent of Dylan Thomas's 'A Grief Ago', where time is measured by the passing of emotion.

Another metaphor, reminiscent of *elephant bottom*, is this.

(6.44)	I don't think he has spatula hands!

Spatula hands illustrates what Cameron calls 'condensation of comparisons into modifier + noun metaphors': [1] he has hands like spatulas, presumably big, square and clumsy, and whether the texter is aware of *spatulate* as a description of hands is hard to tell.[2]

One deliberate, highly contextualised extension of metaphor in CorTxt is the following announcement of a safe arrival in Birmingham, after a tedious journey on public transport.

(6.45)	A:	'Allo! We have braved the buses and taken on the trains and triumphed. I mean we're in b'ham. Have a jolly good rest of week xx
	B:	You intrepid duo you! Have a great time and see you both soon. Xxxx

Here, a journey by public transport is dramatically described in terms more normally associated with war or struggle.[3] It uses a 3-part structure associated with rhetorical devices (*We have braved the buses and taken on the trains and triumphed*), which suggests a deliberate attempt to be dramatic. This is acknowledged by B, who responds: *You intrepid duo you!*

Metaphorical extension on the discourse level includes the following allusion to radio communication.

(6.46)	Roger that. We're probably going to rem in about 20

To sum up, use of conventional or novel metaphors allows texters to achieve highly intimate, intense or evaluative effects in what is otherwise a relatively constrained medium. In many cases, this is also achieved in few key-presses (*spatula hands*), although brevity is

[1] This construction could be surmised as being popular in texting due to its relative brevity: although it is difficult to identify and retrieve instances.
[2] The implication with this text message is that the texter is in fact responding to a metaphor used previously – as B's messages to A are not included in the corpus, this cannot easily be checked.
[3] While this may not be conventional conceptual metaphor, it perhaps draws on cultural dissatisfaction with transport services, and the perception that providers of public transport are working against, rather than for, the public.

not always a priority: *We have braved the buses and taken on the trains and triumphed*. This highlights the interpersonal role of texting rather than its function of conveying information, as use of potentially novel or extended metaphors such as *spatula hands* or *sleeps* suggests shared backgrounds and intimate relationships, and hints at the co-production of metaphorical images. In short, Txt is not a dry, formulaic language but one that in fact incorporates and extends metaphorical language in order to inject humour, establish evaluative stances or convey expressions of intimacy.

6.5.3 Speaker displacement of fixedness

Word play involving 'creatively disfiguring' larger units or 'chunks' such as idioms or fixed expressions include (in CANCODE) the following discussion of cold relations between a couple.

Figure 6.9 Idiom manipulation in spoken language (Carter 2004b: 94)

<S 01> He's at it again but he really wants you know just to sit down.
<S 02> Like they just talk about how they both feel.
<S 01> Out of the frying pan into the deep freeze this time.

The idiom is used to capture the worsening marital situation and, in replacing fire with freeze which plays with the association of hot and cold with feelings and with the general 'kitchen' theme, increases its relevance to the domestic situation (Carter 2004b: 96). This again works simultaneously along three clines: of intimacy, in invoking a shared cultural idiom; intensity, through *frying pan* and *deep freeze*; and evaluation, of the couple's headstrong, reckless nature. Phonological repetition in *frying* and *freeze* is also interesting, while *this time* suggests the speaker's awareness that he is adapting a conventional expression.

Although the main focus in the investigation of texting is on idiom manipulation, displacement of other fixed expressions is illustrated in the following text message.

(6.47) | I just cooked a rather nice salmon a la you |

With *a la you* the texter conveys the fact that they have cooked following a recipe or suggestion from the interlocutor, and the language makes this is an approving, stylish compliment.

Manipulation of idioms in texting includes the following. The first shows how reference to part of a well-known idiom is sufficient to recall the whole expression:

(6.48) Early bird! Any purchases yet?

Although this phenomenon is seen in spoken interaction, it shows how often lengthy idioms can be abbreviated when used in texting. With very few key presses, the above idiom conveys a positive (rather impressed) evaluation.

The following turns a familiar expression on its head.

(6.49) Yes see ya not on the dot

Choice of *not on the dot* seems marked in comparison to *a few minutes late* or *not on time*. It is difficult to determine the exact nature of the effect intended but it could be to emphasise the lateness, to imply that they or their interlocutor are usually on time, or to contradict an earlier promise. The repetition of the vowel sound /o/ in *not*, *on* and *dot*, and the rhythm of the phrase, is also interesting to note (Tannen 1989; Cook 2000).

The following shows a similar twist, albeit one which makes a negative structure affirmative.

(6.50) Good good, billy mates all gone. Just been jogging, again! Did NAME42 enjoy concert? X

This interpretation of *billy no mates*[1] perceives the idiom to have analysable components which allow *billy mates*, in reference to the friends Billy once had: with the implication, presumably, that the texter now feels friendless or alone.

[1] *Billy no mates* is an expression describing somebody with no friends.

The next three show the flexibility inherent in supposedly fixed idioms, not only in a structural sense but in their application to particular situations.

(6.51)	Did you show him and wot did he say or could u not c him 4 dust?
(6.52)	I've got some salt, you can rub it in my open wounds if you like!
(6.53)	Big brother's really scraped the barrel with this shower of social misfits

In the first, use of *c him 4 dust* succinctly evokes the image of someone quickly and recklessly leaving a situation, without regard for those he leaves behind: again, an impression otherwise difficult to convey through texting. In the second, the texter exploits the well-known idiom *rubbing salt in an open wound* to ironically point out some insensitive comment the other texter made, but the new form of the idiom makes it more marked, noticeable and personal. The last, a comment on the contestants of the Channel 4 reality show, is chiefly interesting for the way it combines two expressions in one sentence: *scraping the barrel* and *shower of social misfits*, linked perhaps through the water theme as shower otherwise seems a marked choice. Again, the phonological repetition is interesting: /s/ in *scraped*, *social* and *misfits*; /sh/ in *shower* and *social*, perhaps encouraged by the earlier alliteration *Big Brother*.

Co-construction of idiom manipulation is illustrated in the following exchange in which two texters discuss plans for B to meet A and friends the following week.

(6.54)	A:	Sorry, left phone upstairs. OK, might be hectic but would be all my birds with one fell swoop. It's a date.
	B:	Can help u swoop by picking u up from wherever ur other birds r meeting if u want. X
	A:	That would be great. We'll be at the Guild. Could meet on Bristol road or somewhere - will get in touch over weekend. Our plans take flight! Have a good week x

A refers to the idiom *in one fell swoop* and mixes it with another idiom, *two birds with one stone*. B picks up on this idiom and manipulates it into: *Can help u swoop by picking u up from wherever ur other birds r meeting if u like* in order to offer A and friends a lift. This

close attention paid to A's words heightens the intimacy and level of involvement, and A responds by extending the bird metaphor with: *Our plans take flight!*.

Linguistic awareness is indicated in the following text message through the texter's metalanguage. The added *and all that!* calls on all the texter's linguistic, idiomatic reserves in defence of a friend.

(6.55) Don't give a flying monkeys wot they think and I certainly don't mind. Any friend of mine and all that!

Similar awareness is evident in the following.

(6.56) Guess which pub im in? Im as happy as a pig in clover or whatever the saying is! X

(6.57) here is my new address POSTALADDRESS-apples&pairs&all that malarky

To sum up, although the lengthy expression evident in most texted idioms seems to take precedence over the need for brevity, their occurrence not only indicates intimacy and heightens intensity but also captures highly evaluative attitudes and emotions that would perhaps otherwise be difficult to express in this constrained medium: *I've got some salt, you can rub it in my open wounds if you like!* There is also evidence that idiom manipulation can be co-constructed and that texters are aware of their exploitation of language.

6.5.3 Morphological creativity

Morphological creativity involves the creative reinterpretation or manipulation of morphemes (Carter 2004b: 98). According to Carter, neologisms, perhaps a 'survival' mechanism for when words escape speakers, are common in conversation and, far from disrupting interaction, can be adopted by interlocutors. The practice also shows speakers' morphological awareness (Howden 1984). As Carter's (2004: 99-100) examples show, the *y*-suffix is particularly productive in creating neologisms, although suffixes such as *-wise* are also used.

Figure 6.10 Neologisms in CANCODE (Carter 2004b: 99-100)

<S 01> You said you wanted the little ones [bowls] as well. Want the little ones?

<S 02> Not really … sort of *salady* … that fruit bowl would be ideal.

<S 01> They're well sort of like lycra, *elasticky* sort of stuff.

Carter's example of *greenly challenged* (reproduced earlier) also shows how words can be co-created, and the need for shared cultural knowledge in doing so: the assumption, in this case, that *greenly challenged* is politically-correct.

Examples of morphological creativity in texting include these.

(6.58) see you then, we're all christmassy here!

(6.59) Printer is cool. I mean groovy. Wine is groovying

(6.60) Are you driving or training?[1]

(6.61) Thanks lotsly!

(6.62) Beerage?

While the first example is fairly conventional, the others appear more novel. Example 6.59, which illustrates the creation of a verb from an adjective in *groovying*, is also a play on the words *cool* and *groovy*; for some reason, the texter changes their initial choice of *cool* to *groovy* to describe the printer,[2] and then over-extends this to describe the wine as *groovying* rather than *cooling*. The practice of changing nouns to verbs, as with *training* in example 6.60, is taken to an extreme in the following transformation of the drinking game *I have never* into a (progressive) verb.[3]

(6.63) The wine is flowing and i'm i have nevering..

[1] *Training* here refers to 'taking a train'.
[2] Apparently because their interlocutor had earlier expressed their dislike of the word *cool* and their preference for *groovy*.
[3] This example also illustrates repetition of the grammatical structure.

In example 6.61 above, the adverb *lots* is given an ending typical of adverbs, *-ly*, perhaps to intensify the sentiment or make it sound more intimate; and, in example 6.62 *beer* is given a suffix typical of nouns, *-age*, perhaps with the intention of referring to an event or thing larger than the drink itself. A similar process produces *workage*.

(6.64)	No worries, hope photo shoot went well. have a spiffing fun at workage. x.

As texting does not take place in real time, we could surmise that created words are not part of the survival mechanism suggested by Carter. Nor do they necessarily save time or space as many require extra key presses and risk not being recognised by predictive text devices. The purpose instead must be to evaluate, intensify sentiment or heighten the sense of intimacy. The ability to achieve this through neologism also indicates texters' morphological awareness.

6.5.4 Puns and playing

Puns such as *There must be a raisin for it being there*, cited earlier, are common across language domains, as Carter shows, but it is perhaps their ubiquity which renders them, like sarcasm, 'simultaneously reputable and disreputable' (Carter 2004b: 91).[1] Nevertheless, while some puns work simply on phonological or grammatical ambiguities such as the above, or with homonyms, others are semantically dense. North (2007: 548) describes as punning the intermingling of a serious discussion of money with a humorous play on vegetables, starting when a comment about being poor is met by the suggestion that the speaker could eat turnips, and culminating in: *So is it the Turnip or the swede that is the root of all evil?*. North suggests that such verbal play contributes in a rich and complex way to textual and social cohesion: creating, for example, common viewpoints. As with speaker displacement of fixedness, puns are self-referential and draw attention to themselves, while the laughter and comment they elicit suggests speakers' metalingual awareness.

Texted puns in CorTxt are similarly deliberate, overt and humorous. This first exchange took place prior to B attending a public lunch with the Queen.

[1] Puns, according to North (2007: 546), 'function by disrupting the normal relationship between form and meaning, where a chance coincidence of form stimulates an unexpected interpretation': that is, they exploit similarities between words or phrases.

(6.65)	A:	Have you been practising your curtsey?
	B:	I've got it down to a tea. not sure which flavour

The initial query appears humorous, not least because B is male and unlikely to curtsey. B responds to the humour with *to a tee*, changing it to *tea* (perhaps with reference to the meal he was to have with her majesty); a pun then extended to *not sure which flavour*. This seems to cast doubt on the effectiveness or nature of his curtsey, which could be interpreted as his level of general preparation for the event. Other examples of contextualised, humorous, intimate language play are seen below.

(6.66)	A:	Ooh, 4got, i'm gonna start belly dancing in moseley weds 6.30 if u want 2 join me, they have a cafe too. F
	B:	Not sure I have the stomach for it ...
	A:	Yeah right! I'll bring my tape measure fri!
	B:	Ho ho - big belly laugh! See ya tomo x

(6.67)	A:	I have had two more letters from NAME233. I will copy them for you cos one has a message for you. Speak soon. Xxxxx
	B:	Goodo! Yes we must speak friday - egg-potato ratio for tortilla needed! Xx
	A:	Okay but i thought you were the expert
	B:	Yes obviously, but you are the eggs-pert and the potato head… Speak soon! Xx

The punning in the second example differs from the others in that the word *eggs-pert* not only works as a pun but is itself a morphological creation, similar to *Al Capigeon* and *The godfeather* (Swann and Maybin 2007; see also Swann 2006: 4) coined in conversation to describe a scruffy but aggressive pigeon among a group a family are feeding.

While the chief function of texted puns may be to amuse, the close attention paid to interlocutor's words and the effort made to manipulate them seems to heighten the intimacy of the exchange, showing listenership and creating pattern through the discourse. It could also be suggested that, unlike speakers, the fact that texters have time to read and reread text messages and plan responses may facilitate such wordplay.

6.5.5 Metacommentary

Another resource for creativity in texting lies in the exploitation of features of the online written medium: the ability for texters to refer back to and comment on their previous utterances and, in contrast to spoken utterances, leave a record of their full contribution. This is referred to here as metacommentary.[1] In other words, the observation that 'mistakes' are not always deleted although texting can be edited indicates creativity, while the durability of what is essentially a written medium heightens the possibility of re-readings and re-interpretations. We saw this with repetition of the syntactical frame in *even with the stick – indeed, especially with the stick!* and the suggestion that the second part built on the first to create the overall meaning. Other examples include the following.

(6.68) Ok. Not much to do here though. H&M Friday, cant wait. Dunno wot the hell im gonna do for another 3 weeks! Become a slob- oh wait, already done that! X

(6.69) … and don't worry we'll have finished by march … ish!

(6.70) I'll have a look at the frying pan in case it's cheap or a book perhaps. No that's silly a frying pan isn't likely to be a book

(6.71) Shall I bring us a bottle of wine to keep us amused? Only joking! I'll bring one anyway

In (6.68), *Become a slob- oh wait, already done that!* reinforces the texter's frustration with their situation (and perhaps with themselves) while (6.69) seems to be an ironic comment on the optimism of being *finished by march*. The metacomment in (6.70), part of a discussion about buying Christmas presents and perhaps designed to amuse, may have emerged simply as a consequence of the texter reading the text message back and noticing the ambiguity. Finally, (6.71) is in reference to a planned holiday photo viewing and seems

[1] The terms metalinguistics and metalanguage to describe language which refers to itself is widely used in linguistics, although the term meta *comment* less so. According to admittedly unverified online sources, meta comment was coined by the comedian and television presenter Michael Palin. Drawing on the technique of meta reference in fiction, whereby characters display their awareness of being in a fictional work, meta comment refers specifically to commentary which shows characters' awareness of being within the scene of a film. Here, I use the term somewhat differently, to refer to explicit comments made by language users about their own language use, through which they show their linguistic awareness.

to be an attempt by the texter to both amuse and not offend. Others (not commented on) include these.

(6.72) Hey mr NAME240 NAME353 and I are going to the sea view and having a couple of gays I mean games! Give me a bell when ya finish x

(6.73) I'm going to try for 2 months ha ha only joking

(6.74) meet you in corporation st outside gap … you can see how my mind is working!

Other instances of metacomment explicitly indicate texters' awareness of the medium. Texts which draw on references to the medium or to the failure of predictive texting include these.

(6.75) Your daily text from me – a favour this time

(6.76) if you text on your way to cup stop that should work. And that should be BUS

(6.77) Urgh, coach hot, smells of chip fat! Thanks again, especially for the duvet (not a predictive text word). xxx

(6.78) Thanks NAME270 and NAME56! Or bomb and date as my phone wanted to say! X

6.5.6 Summary of pattern re-forming structures

This investigation of pattern re-forming structures in texting draws attention to an aspect of language previously neglected in sociological and linguistic studies into texting, which tend to focus on creative respelling. Use of morphological creativity as in *beerage*, or idiom manipulation in: *Big brother has really scraped the barrel with this shower of social misfits* suggests that the desire to display identity, indicate stance and emotions, and signal intimacy and group membership is a strong motivation. However, texters do not indulge in language play of this nature *despite* technological constraints. Instead, metaphor or idiom extension, neologism, punning and metacomments are effective way to achieve interpersonal functions, *given* the tiny screen and lack of paralinguistic resources such as

gesture and of graphic resources such as italics. Use of pattern re-forming structures thus reconciles physical constraints with intimate, interpersonal functions, and creates language that is far from dry or formulaic.

6.6 Chapter Summary and Implications

The suggestion that everyday creativity found in spoken interaction can also be identified in texting is perhaps unsurprising, given its informality and previous observations (and concern) regarding texters' manipulation of language form through abbreviation and omission. The problem has been, perhaps, that an overriding concern with respelling obscures deeper investigation of other forms of creativity. This study shows that at least one group of texters[1] draw on creative resources of speech and adapt them to demands of texting. Although some functions of pattern-forming structures in speech are redundant in texting, texted repetition serves to evaluate and to connect utterances between text messages, often in particularly non-speech-like ways such as through repetition of nouns and grammatical structures as a cohesive device. Other uses of creativity characteristic of texting rather than speech include the forging of style and structure through imitation, itself a form of repetition. Exploitation of metaphor, idiom, neologism and pun, as well as the use of metacomments building on previous utterances with humorous or emphatic effect, similarly play crucial roles in texting, by allowing texters to signal highly evaluative and emotive stances, and as such reveal the importance of interpersonal relationships over information transferral or brevity. In fact, physical constraints of texting such as the lack of paralinguisitic or graphic resources, rather than necessarily encouraging abbreviation and omission, require repetition and figurative language in order for texters to fulfil interpersonal functions. As such, this study also extends and supports the argument (Hard af Segersteg 2002; Shortis 2007a,b; Crystal 2008) that texters respond creatively and appropriately to constraints and situational features of the medium, and with a high degree of metalinguistic awareness, by showing that texted creativity is not limited to respellings but found also in wider pattern forming and reforming choices.

Pedagogical implications of this study are practical ones which draw on conclusions reached in studies of spoken creativity.[2] As Carter and McCarthy (2004: 81) observe, the

[1] See Chapter 4 for the bibliographical composition of the corpus.
[2] See Tagg (forthcoming) for elaboration of these suggestions, in relation to both spelling and everyday creativity.

value of creativity in spoken language in the teaching of literature has already been acknowledged (Carter and McRae 1996; Cook 2000) as it allows learners to make connections between their language use and literary language. Implications suggested for language (as opposed to literature) teaching is two-fold. Firstly, studies of spoken creativity highlight the need for interactional alongside transactional classroom tasks such as 'information gaps' in helping learners to understand the interpersonal and contextualised role of creativity, with implications for their understanding of its use in literary language (Carter and McCarthy 2004: 81). Secondly, the understanding that first language users engaging in 'natural' communication focus on form as well as meaning suggests a return to repetitive, form-based instruction, alongside communicative methodologies (Cook 1994; 1997).

Benefits that texting can bring to classroom discussions are three-fold: firstly, the popularity of texting suggests it will be of interest to many learners, particularly children and teenagers. Texting highlights the need for written interactional communication to be motivating and purposeful if children or other learners are to fulfil their creative potential; and provides a tool in creating this kind of interaction. Secondly, as this research and the wider literature suggest, users may be more aware of their use of language form to create effect in texting (as a written medium) than in speech, or can at least more easily become aware of their practices through investigation of what is easily analysable data. In other words, if teachers exploit learners' existing creative resources and increase their awareness of the language resources they use when texting, the act of texting itself can become part of young people's literary learning experience. Thirdly, data can easily be collected, perhaps by learners themselves, and analysed without lengthy recording and transcribing processes. Investigation of their own text messages, either on phones, paper or in electronic format, can overcome to an extent limitations of learners as outside observers, analysing impersonal transcriptions of conversations spoken by others (Carter and McCarthy 2004: 82). Although the suggestion that children 'analyse' text messages on their own phones may have consequences for classroom discipline, the fact that texting offers a personal example of how attention to form can play a significant role in meaningful interpersonal communication is a strong motivation to overcome practical issues.

Finally, the approach adopted in the above study highlighted the evaluative impact of unexpected, creative manipulation of normal forms. The following chapter complements

this by focusing instead on the typical: the frequently-occurring, often formulaic phrases used in Txt.

CHAPTER SEVEN
WORDFORMS AND PHRASES IN TEXT MESSAGING

7.1 Introduction

Using wordform frequency to select linguistic features for further study allows the data to drive findings and limits the extent to which the researcher, although making use of previous research in interpreting findings, determines this initial selection. However, what also makes wordform frequency such a good starting-point for exploring how texting differs from other varieties is that the generally stable distribution of the most frequent wordforms is turned on its head. Unlike other corpora, not only is *a* more frequent than *the*, but *you* more frequent that *I*; *have* more than *is*; *me* more than *my*; and *just* more than *so*. Some variation in the order of these items occurs in other corpora: *I* and *you*, for example, move up to second and third position in spoken English (Leech et al 2001: 144); while *of* moves from its typical position as second most-frequent wordform down to ninth in romantic fiction (Tribble 2000: 82-84). However, as Sinclair (1991: 31) notes, so rarely does the top of the frequency list alter in a fundamental way, that noticeable change is likely to be significant.

Of greatest significance are the frequencies of *a* and *the*, not only the most frequent of the reversed items but those showing greatest discrepancy in comparison with other corpora. These items therefore form the basis of the following investigation. Explanation for their unusual frequencies lies not so much in omission of *the* nor in the speech-like nature of texting, but in the occurrence of both in phrases which recur throughout CorTxt. This explanation is compelling not only because of the perceived significance of formulaic language within the lexicon, but because particular chunks recurring in a corpus have been described as producing a linguistic fingerprint of that variety (O'Keeffe et al 2007). Frequently-occurring phrases appear to be used for more expressive functions in speech than in writing, and a similar use might be expected in texting, given its interactive and speech-like qualities as well as its formulaic nature. If we accept Wray's (2002) argument that formulaic language serves an interactive role in construing identity alongside widely-acknowledged cognitive roles in reducing processing loads and facilitating interaction, the

implications are that phrases in texting indicate group membership and individual texting identities.

This chapter outlines a corpus-driven attempt to identify recurrent phrases which characterise CorTxt. It starts by defining the *wordform* as the basic unit of analysis in the current investigation before outlining what is meant here by the term *phrase*, and describing the main functions phrases are perceived to have. Following discussion of the value of frequency wordlists, comparison of wordform frequency in CorTxt with the BNC highlights unusual frequencies of *a* and *the* in the former, and closer analysis suggests that explanations for these frequencies must acknowledge the frequent phrases in which *the* and, to a greater extent, *a* occur. What these chunks tell us about the nature of Txt is then explored.

7.2 Wordforms and Phrases

7.2.1 Words and units of meaning

Use of *wordforms* as the basic unit of text and investigation of their combination into *phrases* show how units of meaning can be both smaller and larger than the word, despite the traditional focus on words in lexicography (Stubbs 2001: 27). More accurately described as lemmas,[1] words necessarily involve arbitrary interpretation (Sinclair 1991: 29; Stubbs 2001: 25; Sinclair 2004: 24-28/131). Dictionaries, for example, vary as to whether they include parts-of-speech under one lemma: *ing* forms such as *confusing*, to take Stubbs (2001: 25) example, may be listed under base verbs or separately as adjectives. Furthermore, definitions based on semantic meaning fall down once, driven by corpus data, meaning is seen to emerge from usage patterns. *Consuming*, for example, occurs in phrases such as *consuming passion* and *time-consuming* in which *consume* and *consumed* do not appear (Stubbs 2001: 27); *decide* tends to collocate with *to* and *decided* with *that* in referring, respectively, to open and resolved decisions (Hunston 2003).[2] In contrast, wordforms are identified solely through form (Sinclair 1991; Stubbs 2001) as 'unbroken succession[s] of letters' (Sinclair 1991: 28). This has practical advantages in that wordforms can be automatically retrieved and counted more easily and reliably than lemmas (O'Keeffe et al

[1] That is, as 'the item which occurs at the beginning of a dictionary entry … essentially an abstract representation, subsuming all the lexical variations which may apply: the verb walk, for example, subsumes walking, walks and walked' (Crystal 2008b: 273).

[2] Although it should be noted that Stubb's example is an absolute one: *consume* does not occur in a phrase with *passion*; while Hunston's example is one of degree: the two wordforms tend to collocate differently.

2007: 33). For these reasons, while lemmas may be relevant units of *vocabulary*, wordforms are more useful, objective units of *text* (Stubbs 2001: 25).[1] In this chapter, *wordforms*, 'unbroken succession[s] of letters', therefore form the basis of the word frequency count.[2]

Units of meaning, however, are often larger than wordforms, in that meaning stretches linearly across strings of wordforms occurring successively in text.[3] The current investigation, which begins with wordforms, does not stop there but moves on to explore their combination into multi-wordform units (Sinclair 2004).[4] The formulaic nature of much language use, as Wray (2002: 7-8) points out, was noted from at least the early twentieth century by linguists such as Saussure (1916/1966), Firth (1937/1964), Hymes (1968), and Bolinger (1976) until attention was diverted from this view of language by the prominence of transformative-generative grammar (Chomsky 1957), which prioritises a slot-and-filler model of language and relegates idiomatic language to a marginal role in dictionaries. Attention re-focused onto formulaic language with the shift towards the study of language-in-use and through textual evidence of recurring lexicogrammatical patterns across corpora, both of which highlight that although it may be *possible* for people to generate infinite numbers of sentences by slotting words into grammatical structures, in practice they do not: each choice of word constrains and is constrained by those around it (Sinclair 2004). The central role of formulaic language in the lexicon is now widely recognised in psycholinguistics (Wray 2002; 2008), corpus linguistics (Sinclair 2004), lexicography (Moon 1998) and second language acquisition studies (Pawley and Syder 1983).

Studies of language variation (other than the current one) also draw on formulaic language, as they do on word frequency and the distribution of grammatical features. The significance of formulaic language to language description is shown in the number of works which draw

[1] Which can then be defined in turn as a 'succession of word-forms' (Sinclair, 1991:29).
[2] Although, as explained later, the wordforms are grouped together into the spelling groups identified in Chapter 5.
[3] This is not to deny the significance of single wordforms, as O'Keeffe et al (2007:58) point out, not only in the morphological information they carry, but in their role alongside multi-wordform strings and within those open to variation: even wordforms with very constrained contexts such as *kith*, generally only used in *kith and kin*, must be recognised as a wordform as shown in the attested example: 'no more expensive to call your kith in Sydney than your kin in Southampton'.
[4] The usual term is, of course, multi-word unit or string; but the obvious implication of starting from wordforms seems to be that they combine into multi-wordform units – as seen in *consuming passion*. In the following discussion of research into formulaic language, both multi-wordform string and multi-word string are used where relevant.

on the concept of multi-wordform strings: as mentioned above, O'Keeffe et al (2007) describe *chunks* found in CANCODE; Biber et al (1999) calculate the frequency of what they call *lexical bundles* in written and conversational genres; Oakey (2002) looks at *lexical phrases* across academic disciplines. What emerges from such studies is that the most frequent or salient chunks vary systematically according to communication type, participants involved and mode used, which suggests comparison of lists of multi-wordform strings can serve as a starting-point in describing language varieties (O'Keeffe et al 2007: 62). In speech, as discussed below, multi-wordform strings reveal an interactive 'speaker-listener world of I and you' (O'Keeffe et al 2007: 68); while texting is characterised not only by phatic interaction, with certain fixed sequences occurring with greater relative frequency than in speech, but concern for careful leave-taking and coordination of further interaction: the wider, often future world that texters share. However, despite recognition of the importance of strings longer than one word or wordform, in language description and elsewhere, there is less agreement as to what they actually are and how they can be identified or categorised.

7.2.2 Defining, identifying, retrieving formulaic language

The term *phrase* is used in this chapter to describe recurring contiguous strings of two or more wordforms defined primarily in terms of their use as pragmatic frames, a definition that draws on O'Keeffe et al's (2007: 64) description of spoken chunks.[1] The need for researchers to define carefully what they are investigating is evidenced in the plethora of terms and interpretations surrounding formulaic language, and hence the various terms used to discuss it. Wray (2002) lists over fifty, including the following.

Figure 7.1 Terminology used to discuss formulaic language (Wray 2002: 8-9)

amalgams, clichés, collocations, fixed expressions, preassembled speech, prefabricated routines and patterns, sentence builders, unanalysed multiword chunks, units. [2]

[1] The term *phrase* is used here as a 'neutral, non-technical term' (as in Stubbs, 2001: 63). Although O'Keeffe et al use the term *chunk* with a similar meaning, *phrase* is preferred here in that it highlights the fact that they are defined by meaning (albeit pragmatic) as well as by frequency of occurrence (as are O'Keeffe et al's). Other terms may appear in the following review of the relevant literature, following the terms used by different authors. *Multi-wordform strings* is also used as a general, cover-all term.

[2] Even so, Wray's list is not exhaustive: it omits, for example, her own use of *formulaic sequences*, as well as Biber et al's *lexical bundles*; *n-grams*, and Renouf and Sinclair's (1991) *frames* or *collocational frameworks*.

As Wray suggests, different terms do not map neatly onto different aspects of the same phenomenon, but instead mask overlapping or conflicting interpretations and theoretical priorities. Differences in interpretation, apparent in Wray (2002), include how specific and technical terms are: *cliché* and *stereotype* for example, are widely used (outside linguistics) as nontechnical terms with negative connotations; *collocation* is now established within corpus linguistics as referring specifically to statistically significant wordform co-occurrence (Sinclair 1991); while *unit* remains general and pre-theoretical (Crystal 2008b: 503). There are also assumptions concerning their use: *idioms* express concepts while *proverbs* state commonly believed truths or pieces of advice (Schmitt and Carter 2002: 9); and how fixed or fragmented they are: *frames* such as *a + ? + of* (Sinclair and Renouf 1991) explicitly recognise slots within otherwise fixed expressions which *n-grams* and *lexical bundles* do not.[1] The first distinction assumed in this study, however, is that multi-wordform strings are defined according to frequency of occurrence rather than how they are processed.

Wray's (2002: 9) influential psycholinguistic definition of a *formulaic sequence* describes word sequences which are 'stored and retrieved whole'. However, a definition which describes strings as being processed holistically excludes less fixed associations retrieved through corpus investigation, including Sinclair and Renouf's (1991) frames, Hoey's (2004) semantic associations, and Sinclair's (2004) lexical unit. As Read and Nation (2004: 26) point out, the focus on processing tends to discount substitutions or transformations such as *fat-chewing* from *chew the fat*, because these involve constituent analysis: leading Read and Nation to suggest the need to distinguish *semantic patterning* from *formulaic sequences*.[2] More importantly, holistic storage and retrieval of sequences, by Wray's (2002: 19) admission, is neither external nor observable and cannot therefore be measured or used to identify strings. Although, as Kuiper (2004: 869) points out, experiments such as cloze texts, and research into eye-movement and phonetic production, can and have been used to determine whether phrases are stored holistically, the validity of such tests can be questioned. As Read and Nation (2004: 25) suggest, for example, how language is stored

[1] Other differences include the opacity of strings (the terms *idioms* and *metaphors* presuppose a figurative meaning which may nor may not be retrievable from the individual words; but which is not implied by *chunks*); their provenance (*fossilised forms*, *frozen metaphors* or *phrases*) and whether they need to be meaningful 'wholes' (although *idioms* or *fixed expressions* are semantically complete, *chunks* and *lexical bundles*, as we shall return to later, need not be).

[2] Where semantic patterning not only allows for 'insertion, inflection, substitution, deletion, and transformation' (Read and National (2004: 26) but encompasses collocations, semantic prosody and other unfixed features found, for example, in Sinclair's (2004) lexical unit (see Chapter 3).

and retrieved varies between individuals and as communicative purpose and processing demands shift. Nor is it possible to use corpus data to determine how language is processed (Schmitt et al 2004: 147), and Wray's (2002: 25) assumption that if a string is used frequently it is likely to be stored holistically for ease of processing, and that this will in turn encourage greater use, remains to be tested (Schmitt et al 2004: 128). Formulaic sequences must therefore be extracted by recourse to *intuition*, and Wray (2008: 113-127) shows how criteria guiding intuitive judgements can be analysed and made explicit. However, the validity of this approach is limited by the subjectivity involved and by speakers' lack of awareness of frequent patterns of language use revealed through corpus data (Sinclair 1991; Stubbs 2001).[1] Also of relevance to corpus analysts is the fact that Wray's definition, although based on thoroughly reviewing the literature, does not draw on empirical data.[2] Empirical study, for example, shows that fixed expressions intuitively felt to be part of the language are not always particularly frequent (Moon 1998) and therefore not reflective of typical language use. In short, intuitive recognition of holistically-processed phrases cannot easily be reconciled with an approach which prioritises what is frequent and typical across large language samples.

Use of frequency to identify multi-wordform strings within corpus linguistics therefore emerges not only from prioritising attested data description, and an interest in what is typical and frequent in language, but because corpus analysis can identify recurrent strings neglected in lists determined through introspection (O'Keeffe et al 2007: 61). In other words, the process is *corpus-driven*, in so much as the computer identifies associations through statistical frequency without intervention from researchers and with certain implications for language description (Cheng et al 2006: 415).[3] Strings identified initially through frequency are distinguished from *formulaic sequences* by Schmitt et al (2004) as *recurrent clusters*,[4] and it is essentially these recurrent clusters that this study is interested in. This label would also include O'Keeffe et al's (2007) *chunks* as well as Biber et al's (1999: 990-1024) *lexical bundles* and *n-grams*, used widely as a probabilistic model in natural language processing (NLP) to predict subsequent items, as well as in corpus and

[1] Wray (2002; 2008) also makes the point that the fact native speakers do not break down longer units of meaning for analysis into their constituent parts can be used to explain their being unaware of frequent patterns of use emerging though corpus analysis.
[2] Although in Wray (2008: 129-185), 6 case studies of formulaic language 'at the boundaries' are outlined: one of the boundaries being the fact that the case studies are all fairly non-typical of everyday language use.
[3] See Chapter 3 for discussion of the lexical grammar model that has emerged from corpus-driven study.
[4] So that, along with *semantic patterning* (see footnote 2, p. 194), we now have three terms.

computational linguistics. Admittedly, one advantage of *concgrams* (Greaves 2005) over the above is their inclusion, alongside contiguous phrases, of phrases displaying constituency variation (where words are inserted into the phrase), as well as positional variation (where associated words occur in different positions to each other) (Cheng et al 2006: 414). As explained below, however, this study instead follows O'Keeffe et al (2007) in exploring contiguous forms which serve as pragmatically meaningful *frames*.

A limitation of the automatic searches needed to retrieve recurrent clusters is that they cannot make judgements as to whether frequent strings are semantically or pragmatically meaningful. Human intervention is needed to determine which if any to exclude: features of speech such as hesitations and repetitions, for example, or those extending across punctuation or speaker boundaries (Wray 2002: 27). O'Keeffe et al's (2007: 68) investigation of chunks in CANCODE excluded 'hesitant starts' such as *you, you, you* and 'reduplicated responses' as in *no, no, no* and 'non-lexical vocalisations' such as *er, er, er*. Nor can automatic searches distinguish formulaic from non-formulaic uses of the same word string (Wray 2002: 31).[1] Besides this, automatically identified multi-wordform strings will also include incomplete or fragmentary strings such as *are to my* and *this one for* alongside meaningful and productive (if syntactically incomplete) strings such as *it is true that* and *a bit of a* (O'Keeffe et al 2007: 61/64). Again, human intervention is needed. Although some researchers would include unproductive strings (Biber et al 1999: 990), the decision is thus made by O'Keeffe et al (2007: 64) to focus on those with syntactic or semantic incompleteness only where they had 'pragmatic integrity'.

Pragmatic integrity here refers to the 'interactive role' of strings such as *it is true that*, or *a bit of a* which may on their own be semantically empty but which function to preface, frame and evaluate information. Although grammatically incomplete and without independent meaning, *a bit of a* acts as a frame in fulfilling its pragmatic function as a downtoner.

[1] They cannot necessarily differentiate, for example, between cold feet in I'm getting cold feet about this idea of yours from You must have cold feet in those sandals.

Figure 7.2 a bit of a (O'Keeffe et al 2007: 71)

	mess
	problem
a bit of a	performance
	hassle
	nuisance
	bargain

As mentioned above, the term *phrase* is used similarly in this chapter to describe contiguous strings of two or more wordforms which recur, often as 'frames' to which 'new unpredictable content' can be attached (O'Keeffe et al 2007: 71) and with consideration of their pragmatic function, if not their syntactic or semantic completeness (O'Keeffe et al 2007: 64). Although this definition excludes aspects of formulaic language which other researchers are criticised for neglecting (including non-contiguous sequences), the decision to identify phrases in terms of frequency (and pragmatic integrity) rather than with reference to how they are processed is justified because of difficulties in identifying holistically-processed strings and because the current aim is towards data description rather than analysis of language processing. The decision to limit investigation to contiguous multi-wordform chunks rather than less fixed word associations is in part a practical one which recognises the need to be selective and also allows comparisons with spoken and written chunks identified by Carter and McCarthy (2006) and O'Keeffe et al (2007) and the functions they are seen to fulfil.

7.2.4 Functions of multi-wordform strings

The main function of very frequently-used phrases emerges, from across the literature, as that of oiling the wheels of interaction. As described above, data from CANCODE suggests that recurring semantically- and syntactically-incomplete chunks, at least in spoken language, often display pragmatic integrity (O'Keeffe 2007: 70-71);[1] similarly, formulaic sequences in the sense of holistically-processed phrases have *expressive* rather than *referential* functions (Wray 2002). Phrases facilitate interaction in three main ways: by

[1] My argument here draws on two very different interpretations of multi-wordform strings: as holistically-processed and memorised phrases (Wray 2002; 2008) and as strings of words that are seen to co-occur across corpora (O'Keeffe 2007); but hopes to show that the arguments and evidence of both support the idea that items longer than a word fulfil certain, largely pragmatic, functions.

organising speaker-listener relationships; by reducing processing loads; and, according to Wray (2002), through the construction of identity.

The function of frequently-occurring chunks in discourse marking includes those used in monitoring interaction and in facework and politeness, as well as vague language and approximations (O'Keeffe et al 2007: 73-75). The monitoring discourse marker *you know*, has, as O'Keeffe et al (2007: 72) point out, a well-documented role in indicating shared knowledge between speakers (Schiffrin 1988) and is the most frequent chunk in CANCODE. Conversation tends to be vague rather than precise and frequent chunks such as *and things like that* enable speakers to refer to numbers or quantities without sounding 'blunt and pedantic' but instead draw on shared real-world and cultural knowledge (O'Keeffe et al 2007: 74; Carter and McCarthy 2006). Other frequent chunks such as *I think* and *I don't know* function as hedges which soften assertions and protect from rebuttal (O'Keeffe et al 2007: 73). For similar reasons, observes Wray (2002: 95), formulaic expressions such as *I wonder if you'd mind …* tend to frame speech acts such as requests, demands and warnings (Austin 1962).[1] According to Wray (2002: 91), one reason for the fixedness of such formulae is to ensure hearers interpret utterances as polite, the effect of which is that such exchanges become rituals for which it is essential to learn the correct things to say in order to achieve communicative goals in smooth, non-face-threatening fashion.[2] This is interesting when applied to texting where, as we have seen, interpersonal awareness seems to override concerns with brevity.

The argument that formulaic language serves also to reduce processing loads rests on the premise that chunking reduces the items speakers retrieve from their information store. This can be compared with Wray's (2002: 72) examples of prayers or proverbs deliberately memorised so as to avoid reconstruction, and mnemonics such as *Richard of York gained battle in vain*[3] which trigger larger amounts of information; while Kuiper's (2004) in-depth studies of sports commentators and auctioneers suggest that speakers regularly performing under pressure rely on formulaic language to relieve demands on memory resources.

[1] Formulaic language can also have the opposite effect; that is, of distracting attention from what is being said, and of not drawing hearers into the interaction: 'When someone says *Excuse me* as they stand up or move away, the intention is an unobtrusive apology, which can be registered without any interruption to the main event' (Wray, 2002: 95).

[2] Hence parents' eagerness to teach children politeness formulae: *don't forget to say X, did you say Y?* (Gleason 1980: 252 in Wray 2002: 110).

[3] *Richard of York gained battle in vain* aids recall of the colours of the rainbow: red, orange, yellow, green, blue, indigo, violet.

Formulaic sequences are also believed to pad out what speakers say to give them time to think[1] and to contribute to the rhythm of sermons or rhetorical speeches (Wray 2002: 83-84). The easing of hearers' processing effort and their understanding of what speakers tell them is also achieved through formulaic language. According to Wray (2002: 95), formulaic phrases such as the aforementioned *I wonder if you'd mind …* draw hearers' attention to requests which may then be expressed in more novel ways.[2] This view, although widely accepted across the literature (Kuiper 2004; O'Keeffe et al 2007; Stubbs 2001: 60-61; Wray 2002), is largely unsupported by experimental data (Terkourafi 2003: 1722-23)[3] and corroborated only to an extent by corpus data, which can retrieve recurring clusters but cannot (as mentioned previously) make claims that they are necessarily processed holistically nor determine whether chunking has any effect on processing. Nonetheless, from Sinclair's (1991: 110) perspective, units of meaning can constitute 'single choices' and frequently-occurring chunks, according to O'Keeffe et al (2007: 61/76-79), contribute to fluency because they do not need 're-assembling' with each use; nor, presumably, deconstructing. Attempts by texters to relieve processing demands are widely-documented (Hard af Segersteg 2002), as is the formulaic nature of Txt, and so the potential role of ready-assembled chunks in easing cognitive loads should not be overlooked.

Expression of individual and group identity, Wray (2002: 72) argues, is another function of holistically-processed strings. Strings which communities deliberately remember and which distinguish them from others range from national anthems and pledges of allegiance to favoured quotations and rhymes such as *Remember remember the fifth of November* (Wray 2002: 72). Group rituals similarly range from politeness formulae mentioned above and *bless you* when somebody sneezes, to formulaic expressions such as *Amen!* in religious rituals. Driven by 'desire to sound like others in the speech community' (Wray 2002: 75), individuals hear, store, use and thus contribute to stocks of politeness formulae, as well as shared turns of phrase and collocations. Inevitably, this use of formulaic language excludes those who do *not* belong to the group such as non-native speakers (Wray 2002: 90), as

[1] Hence the option of the 'circumlocutory' formulaic *take a decision* as well as the shorter *decide*.
[2] Where speaker and hearer do not share the same formulaic phrases, those of the hearer's takes precedence: as can be seen in interaction between people from different speech communities, such as American and British, or between native and non-native speakers (Wray 2002: 99). In such situations, it is in the speaker's interests to adopt word strings which the hearer will understand and, particularly in the latter example above, will signal to the hearer that the speaker is 'fluent' and can cope with the community's formulaic language.
[3] However, Wray's conclusion, that the one unifying function of formulaic language is the furtherance of speakers' own interests or 'the promotion of self' is to some less convincing. Not only does Wray neglect to define what she means by 'self' (Terkourafi 2003: 1723-4) but her attempt to relate varied aspects of formulaicity to one human intention appears reductionist and simplistic (Cobb 2003).

illustrated in studies which focus on chunks and collocation in improving learner fluency (O'Keeffe et al 2007: 76-77); as well as by Kuiper's (1996) observation that formulaic sequences in auctions are restricted (despite the apparent need to reduce auctioneers' processing loads) by novice bidders present (Wray 2002: 77; see Kuiper 2004). As this suggests, *communities* may be academic, scientific or professional as well as national, in which formulaic phrases play a functional role as well as marking group membership (Wray 2002: 85), illustrated, for example, by empirical studies of phraseologies used by different academic disciplines (Groom 2006). Assuming that recurring clusters can be credited the same function as those holistically processed, the frequent occurrence of certain phrases in texting can be seen as a similar resource in texters' construal of identity and group membership: particularly given that, as described in earlier chapters, Txt style is evolving and grammatical and spelling choices contribute to meaning-making.

Conclusions reached below regarding the role of phrases in texting will, whilst taking the above factors into account, necessarily remain cautious and investigation will focus on data description, in particular comparison to written and spoken chunks.

7.2.3 Multi-wordforms strings in written and spoken language

Investigation of multi-wordform strings in spoken and written corpora (Carter and McCarthy 2006; O'Keeffe et al 2007: 58-75) are useful as a point of comparison for phrases found in texting. The most frequently-occurring two-, three-, four-, five- and six-wordform sequences[1] in spoken and written data are shown below.[2]

[1] As O'Keeffe et al (2007:65) point out, chunks longer than six wordforms are 'extremely rare', with only one seven-wordform chunk occurring with sufficient frequency to be included in the analysis (see Miller 1956, on the 'magic' number of seven as a psychological limit for the mind to process).

[2] The spoken chunks are taken from CANCODE; and the written from the 700-million Cambridge International Corpus (CIC), which informs Carter and McCarthy's (2006) reference work, and which comprises a range of texts, both spoken and written, including newspapers, advertising, letters, and literary texts.

Table 7.1a-d The most frequent clusters in spoken and written language (Carter and McCarthy 2006: 829-831; O'Keeffe et al 2007: 58-75)

a. Three-wordform clusters in spoken texts					**Three-wordform clusters in written texts**		
1	I don't know	11	you want to	1	one of the	11	it would be
2	a lot of	12	you know what	2	out of the	12	in front of
3	I mean I	13	do you know	3	it was a	13	it was the
4	I don't think	14	a bit of	4	there was a	14	some of the
5	do you think	15	I think it's	5	the end of	15	I don't know
6	do you want	16	but I mean	6	a lot of	16	on to the
7	one of the	17	and it was	7	there was no	17	part of the
8	you have to	18	a couple of	8	as well as	18	be able to
9	it was a	19	you know the	9	end of the	19	the rest of
10	you know I	20	what do you	10	to be a	20	the first time

b. Four-wordform clusters in spoken texts					**Four-wordform clusters in written texts**		
1	you know what I	11	I thought it was	1	the end of the	11	per cent of the
2	know what I mean	12	I don't want to	2	at the end of	12	one of the most
3	I don't know what	13	you know I mean	3	for the first time	13	the side of the
4	do you want to	14	that sort of thing	4	the rest of the	14	the edge of the
5	do you know what	15	I don't know how	5	in the middle of	15	the middle of the
6	I don't know if	16	if you want to	6	at the same time	16	in front of the
7	a bit of a	17	well I don't know	7	the back of the	17	I don't want to
8	I think it was	18	I was going to	8	at the top of	18	for a long time
9	I don't know whether	19	have a look at	9	the top of the	19	is one of the
10	what do you think	20	you don't have to	10	the bottom of the	20	on the other hand

c. Five-wordform clusters in spoken texts					Five-wordform clusters in written texts		
1	you know what I mean	11	and all that sort of	1	at the end of the	11	the other end of the
2	at the end of the	12	I was going to say	2	by the end of the	12	at the bottom of the
3	do you know what I	13	and all the rest of	3	for the first time in	13	the rest of the world
4	the end of the day	14	and that sort of thing	4	at the top of the	14	for the first time since
5	do you want me to	15	I don't know what it	5	at the back of the	15	had nothing to do with
6	in the middle of the	16	all that sort of thing	6	on the other side of	16	at the foot of the
7	I mean I don't know	17	do you want to go	7	in the centre of the	17	in and out of the
8	this that and the other	18	to be honest with you	8	the end of the day	18	in the direction of the
9	I know what you mean	19	an hour and a half	9	for the rest of the	19	is one of the most
10	all the rest of it	20	it's a bit of a	10	the middle of the night	20	the end of the year

	d. Six-wordform clusters in spoken English	
1	do you know what I mean	236
2	at the end of the day	222
3	and all the rest of it	64
4	and all that sort of thing	41
5	I don’t know what it is	38
6	but at the end of the	35
7	and this that and the other	33
8	from the point of view of	33
9	a hell of a lot of	29
10	in the middle of the night	29
11	do you want me to do	24
12	on the other side of the	24
13	I don’t know what to do	23
14	and all this sort of thing	22
15	and at the end of the	22
16	if you see what I mean	22
17	do you want to have a	21
18	if you know what I mean	21

Points to note regarding the extraction procedure followed in obtaining these lists, in order to facilitate their comparison with CorTxt, include their frequency cut-off point as well as yet finer distinctions as to what constitutes a wordform: contractions (as in *don't*) are not broken at the apostrophe but counted as one wordform (O'Keeffe et al 2007: 65).[1] Frequency cut-off points are needed for practical purposes, O'Keeffe et al acknowledge, although the number chosen is always arbitrary (Wray 2002: 27). The figure chosen for CANCODE was four times per million words (20 times in five-million words), somewhat more generous (as O'Keeffe et al 2007 acknowledge) than Biber et al's (1999: 992) choice of 10 times per million and Cortes's (2002) 20 occurrences per million. However, principles guiding this choice include consideration of purpose and data: namely, the low frequency of six-word chunks in CANCODE (only 18 occurring 20 or more times). Finally, as mentioned previously, results exclude hesitant starts, reduplicated responses and repeated non-lexical vocalisations (O'Keeffe et al 2007: 68).

Once lists are generated, the next step is to extract pragmatically-complete units, through 'manual inference and qualitative interpretation' (O'Keeffe et al 2007: 64). As evident above, although most spoken chunks are syntactically incomplete,[2] many fulfil aforementioned pragmatic functions (O'Keeffe et al 2007: 75).

Table 7.2 Functions of chunks in CANCODE

Function	Chunks used
Discourse marking	*you know* (the most frequent), *you know what I mean*, *at the end of the day*, and *if you see what I mean* (O'Keeffe et al 2007: 71)
Facework and politeness	*do you think*, *I don't know if/whether*, *what do you think*, and *I was going to say* (O'Keeffe et al 2007: 73)
Vague expressions	*a couple of*, *and things like that*, *or something like that*, *all the rest of it*, and *(and) all this/that sort of thing* (O'Keeffe et al 2007: 70-75)

These chunks, if not highlighting the interactive nature of speech (as implied by O'Keeffe 2007: 75), at least indicate that interaction between participants in speech is consistently

[1] Although arguments can also be made for breaking contractions at the apostrophe, as in the BNC word frequency count (Leech et al 2001).

[2] In terms of grammatical structure, commonly occurring chunks included subject [pronoun] + verb (*I was*, *you can*, and with negatives/interrogatives: *I don't know*, *you have to*, *what do you*); subject [pronoun] + verb + complement (*you know what*, *I don't know if*, *do you want me to*).

managed by certain recurring phrases, and that the most frequently recurring phrases fulfil expressive rather than referential functions. That these frequently-occurring phrases differ from those found in other varieties allow O'Keeffe et al to describe them as offering 'some sort of fingerprint of everyday conversation' (see also Carter and McCarthy 2006: 828-837).

The fingerprint of written data is, in O'Keeffe et al's words, 'a 'world-out-there' representation, created through impersonal constructions and prepositional relationships which contrast with 'the speaker-listener world' in spoken data (O'Keeffe et al 2007: 68).

Table 7.3 Functions of written chunks in the Cambridge International Corpus (Carter and McCarthy 2006: 832-837)

Function	Chunks used
Time and place	Prepositional phrases: *in the, on the, at the, out of the* Noun phrases + *of*: *the edge of the, the bottom of the, the side of the* Combinations: *in the middle of the, at the top of the*
Possession, agency, purpose, goal, and direction	*of a, of the, to the, with the, by the, for the*

These clusters relate together referents in the real world: they are referential rather than interpersonal. Although caution must be exercised in claiming that speech is therefore more interactive than writing,[1] this suggests that formulaic language in writing fulfils different functions than in speech, and that interaction in speech is managed through recurrent chunks to a greater extent than in writing. Finally, the overall frequencies of chunks in writing are lower than in speaking, at least for 3-word clusters and above,[2] suggesting that speech can be described as more formulaic (in the sense of comprising contiguous phrases) than writing.

Although the above descriptions cannot claim to be entirely comprehensive in their coverage of speech or writing (either in terms of data used or features explored), they indicate how phrases constitute one source of difference between language varieties. The following investigation, which uses wordform frequency as a way in to the data and

[1] Expressive functions could, after all, be expressed using non-contiguous phrases and other less fixed associations.

[2] There are 21,054 2-wordform clusters in CANCODE compared to 23,961 in the written corpus; but 13514 3-wordform and 2819 4-wordform spoken clusters compared to 9499 and 1328, respectively, in writing (Carter and McCarthy 2006).

explores the phrasal uses of frequent words, shows how phrases in texting compare, in form and function, from those above.

7.3 Wordform frequency in texting

7.3.1 A corpus-driven way into text

The value of wordform frequency lies in the observation that comparisons between corpora or subcorpora reveal differences in their distribution across language varieties. This idea emerged, as Leech et al (2001: ix-xi) explain, with the compilation of lists for the first large computer-stored corpora of written American and British English (Kucera and Francis 1967; Hofland and Johansson 1982), and can be seen in descriptions of particular varieties represented by small corpora: computer science (James et al 1994); psychiatric interviews (Dahl 1979), romance fiction (Tribble 2000) and short stories (Stubbs 2001); as well as in large-scale comparisons of spoken and written varieties (Leech et al 2001; Biber et al 1999; Carter and McCarthy 2006). Although there is some truth in the assumption[1] that the most frequent items do not vary (*the*, for example, generally remains the most frequent), these studies highlight important differences: even those most frequent in writing (*the, of, to, a, and*) are not the same as those in speech (*the, I, and, you, it*), as Stubbs (2001: 75) shows. Wordform-frequency lists for the BNC have, Leech et al (2001) suggest, uses for NLP and language teaching as well as offering useful points of comparison for other corpora, and form the starting-point for the current investigation into texting.[2]

Limitations to wordform-frequency lists, however, include the fact that the significance of difference between frequencies cannot usually be assumed without statistical calculation of the probability that the observed differences are due not simply to chance but to a 'substantive divergence' between varieties (Leech et al 2001: 16). This calculation can be done automatically using corpus analysis tools such as WordSmith (Scott 1996), which uses Dunning's (1993) log-likelihood calculation to determine significantly frequent, or infrequent, items in the frequency wordlist of one corpus compared to that of a reference corpus.[3] These words are *keywords*.[1] The difference between wordform frequency and

[1] Stubbs (2001: 75) suggests that '[i]t is often assumed that frequency lists of the most common words in the language do not differ much between corpora'.
[2] Although the alphabetised lists compiled by Leech et al (2001) are lemmatised, the rank frequency lists are not: they are wordforms.
[3] See Leech et al (2001:16-17) for a brief summary of the advantages of log-likelihood over a chi-square test and an outline of the calculation.

keyword lists is, therefore, that while the former highlights the most frequent wordforms in a corpus, generally closed-class words, the latter leans more towards lexical words, the frequency of which, whilst not necessarily high in relation to other wordforms, are significantly different from that in another corpus.[2] Keyword lists can therefore indicate the 'aboutness' and 'style' of corpora (Scott 1997) and, as such, form the basis of investigation into language varieties (e.g. Tribble 2000). However, as mentioned in Chapter 3, caveats regarding keyword lists include the possibility of uneven distributions of keywords across a corpus, as well as their overemphasis on lexical *differences* between corpora at the expense of similarities (Baker 2004: 349-357). These are largely avoided with wordform-frequency lists, not only because particularly frequent wordforms tend to be evenly distributed but because wordform frequency lists highlight similarities as much as differences. Furthermore, in the current investigation of CorTxt the reversed order of many of the most frequent wordforms in the frequency list is sufficiently marked to render keyword analysis unnecessary.

Finally, it should be remembered that wordform-frequency lists only ever serve as a way into the data and as an indication of what to explore (Sinclair 1991: 31). The following investigation therefore begins with the frequency list and proceeds to explore clusters and associations of distinct wordforms. As such, the aim is to let the nature of CorTxt drive the course of investigation and shape findings. However, as well as the above caveats, text message wordform-frequency raises practical issues which must first be addressed.

7.3.2 Practical issues

Generating a wordform list for CorTxt using Wordsmith Tools (Scott 1996) is complicated by respellings and abbreviations which characterise texting.[3] These are illustrated by the following text message from CorTxt which includes number homophones (*2*, *4*) and other phonetic spellings (*u*, *wot*, *sumfing*), as well as clippings (*jus*, *prob*, *bday*) and consonant writing (*frm*).

---

[1] Or, to follow the argument made in this chapter, keywordforms! Given that keyness is not central to this thesis, this is not pursued here.

[2] Although, as Baker (2004: 349-357) points out, the focus in keyword analysis on lexical difference has its own limitations, in so much as it neglects grammatical functions (which could be picked out by word-frequency lists) and semantic meaning.

[3] See Chapter 5: Respellings in Text Messages.

(7.1) Hi NAME219 did u decide wot 2 get NAME85 4 his bday if not ill prob jus get him a voucher frm virgin or sumfing xxx[1]

The question this raises is whether variants should be counted as individual wordforms (so that *u* is distinct from *you*), or whether spelling groups created in Chapter 5 should form the basis of current investigation. Arguments can be made for both approaches. As argued in Chapter 5, respellings carry the potential for making social meaning, so that choice of *u* makes statements regarding group identity and deviance, and *ye* is more informal and intimate than *you*. At the same time, however, the referential meaning of these three items is the same, as seen in the concordances below, and treating them as separate items lowers the apparent frequency with which interlocutors are addressed.

Figure 7.3 Concordance lines for *you* / *u* / *ye*

1	rts again! What time do	**you**	get in? If you can s
2	ello NAME39, what time r	**u**	thinking about leaving?
3	ld be fine. What time do	**you**	need picking up? 7
4	ust be a terrible time for	**you**,	and everyone's bein
5	hats cool what time do	**you**	get back cos we m
6	real word? wot time sun	**u**	planning leaving or did
7	ey NAME242, what time r	**u**	guys heading 2 spoon
8	ey honey! What time do	**u**	want me n NAME18 to
9	t, 2029. What time have	**you**	booked? Hope you
10	ey buddy, what time do	**you**	finnish work? just ro
11	e tights. What time are	**you**	thinking? NAME64 Chee
12	cornwall. What time do	**u**	finish work? NAME2 and
13	yet Boo. What time do	**you**	finish work? Hey jus
14	ress? Hi, what time are	**you**	coming in? Musgrov
15	ust finished wot time do	**u**	think you'll be done?
16	ill alright? What time do	**u**	want to go? Ready w

The decision to use spelling groups as the basis of this wordform-frequency count takes limitations into consideration, while also retaining information as to the forms comprising the groups. The decision is not unlike that faced by analysts of Chaucer's *Canterbury Tales*, similarly characterised by numerous spelling variants (Barnbrook 1990). Finally, it should

[1] Rather than *Hi NAME219 did you decide what to get NAME85 for his birthday, if not I'll probably just get him a voucher from Virgin or something xxx.*

be noted that variants have been carefully checked using concordance lines, in order to confirm their status as respellings of the headword.[1]

Another issue is that of alphanumeric sequences such as *lookin4ward 2seein u soon*, which comprise strings of letters and numbers without spacing. Individual wordforms are therefore not recognised by Wordsmith and the strings (*lookin4ward* and *2seein*) treated as one wordform in frequency counts, thus reducing the apparent frequency with which wordforms like *for* and *to* occur.

A final issue is that shared by researchers of spoken or informal written corpora. Texting has not only been described as an intimate and interpersonal medium more similar to conversation than typical writing situations in terms of functions and participant relationships,[2] but expressions and discourse markers more typically associated with the spoken mode also occur in texting.

(7.2) Well done, blimey, exercise, yeah, i kinda remember wot that is, hmm. Xx

(7.3) Oi when you gonna ring

(7.4) I'm coming back on Thursday. Yay. Is it gonna be ok to get the money. Cheers. Oh yeah and how are you. Everything alright. Hows school. Or do you call it work now

As well as the issue of spelling variants such as *kinda* and *gonna*, another issue raised here concerns standard contracted forms, seen in (7.4) with *I'm*, as well as frequent contractions missing apostrophes: *Hows*, in (4). The question is whether to set Wordsmith to recognise strings which include apostrophes as one wordform (lowering the overall frequency of words like *I*, as *I'm* and *I'll* for example are treated as distinct items) or two (affecting investigation of words like *don't*). The current decision to treat contracted forms as one word not only avoids ambiguous wordforms such as *'s* and *'t* but includes spelling variants of contractions, as seen in Chapter 5 (largely the omission of apostrophes so that *I'm* is respelt as *im*). It also enables comparisons between texted wordforms and CANCODE (O'Keeffe 2007: 65), although contrasting with the practice used with the BNC (Leech et al

[1] See Chapter 5: Respellings in Text Messaging.
[2] See Chapter 2: Research intoText Messaging.

2001). Finally, it must be borne in mind when exploring the CorTxt wordform frequency list that *I*, for example, would occur more frequently if we included its occurrence in *I'm*, *I've*, *I'd* and *I'll* (not to mention *im*, *ive*, *id* and *ill*).

As argued in Chapter 5, the choice this implies for texters when composing text messages, between standard and non-standard spelling variants or between abbreviated and speech-like modes of expression, suggests texters are actively involved in construing language particular to texting, and that in doing so they perform illusions of brevity, speech-like informality, and deviance from conventional norms. As outlined below, wordform frequency and the phrases in which frequent wordforms occur similarly contribute to the construal of texted identity.

7.4 *the* and *a* in texting

7.4.1 Most frequent wordforms in texting

Comparison of the twenty-five most frequent wordforms in the BNC (Leech et al 2001: 120) and in CorTxt (generated with WordSmith's Wordlist) highlights significant differences in the distribution of wordforms between the two corpora.

Table 7.4 Wordform frequencies for the BNC and CorTxt[1]

BNC wordform frequency		
Word	**PoS**	**Frequency (million wds)**
the	det	61847
of	prep	29391
and	conj	26817
a	det	21626
in	prep	18214
to	inf	16284
it	pron	10875
is	verb	9982
to	prep	9343
was	verb	9236
I	pron	8875
for	prep	8412
that	conj	7308
you	pron	6954
he	pron	6810
be	verb	6644
with	prep	6575
on	prep	6475
by	prep	5096
at	prep	4790
have	verb	4735
are	verb	4707
not	neg	4626
this	detp	4623
‘s	gen	4599

CorTxt wordform frequency			
Word	**PoS**	**Raw Frequency 190,516 wds**	**Rounded Frequency million wds**
you	pron	7,884	41382
to	inf/pr	4,976	26119
i	pron	4,257	22345
x	sym	3,689	19363
a	det	3,580	18791
the	det	3,553	18649
and	conj	3,171	16644
in	prep	2,387	12529
for	prep	2,057	10797
it	pron	2,020	10603
have	verb	1,993	10461
is	verb	1,577	8278
be	verb	1,567	8225
me	pron	1,555	8162
on	prep	1,523	7994
are	verb	1,478	7758
of	prep	1,420	7453
at	prep	1,393	7312
my	pron	1,285	6745
good	adf	1,265	6640
but	conj	1,257	6598
see	verb	1,255	6587
just	adv	1,240	6507
i'm	p/v	1,216	6383
so	adv	1,159	6084

* *u* accounts for 3043 of the total occurrences of *you* (39 percent)

** Unlike the BNC wordlist, *to* includes infinitive and preposition (i.e., it is treated more consistently as a wordform.

*** Unlike the BNC wordlist, *I* in Cortxt does not include its use in contractions such as *I’m*.

† *b* accounts for 375 of the total occurrences of *be/b* (24 percent)

†† *of* accounts for 0.75 percent of total word count, compared to over 2 percent in most corpora

[1] See Tagg (2007) for an account of a word frequency analysis using a smaller version of CorTxt (550 messages).

Observations emerging from comparison of the frequency lists are, firstly, that in texting *the* is not at the top of the list but is the third most frequent, while the most frequently occurring wordform is *you*. [1] *Of* is also markedly less frequent in texting (moving from second to seventeenth place); while several of the twenty-five most frequently occurring wordforms in the BNC do not appear among the top twenty-five in texting, including: *was*, *that*, *he*, *with*, *by*, and *not*. On the other hand, wordforms appearing among the twenty most frequent in texting but not in the BNC include: *me*, *my*, *good*, *but*, *see*, and *just*.[2] Another observation is that there appears to be less variation in lexical frequency in texting: that is, the most frequent wordforms occur relatively more often in texting than in the BNC. In the BNC, although *the* occurs over 60,000 times per million words, it is twice as frequent as the next item *of* with 29,391 occurrences; in texting, *you* would occur only around 40,000 times per million wordforms but *to*, second most frequent, would occur almost as frequently as does *of* in the BNC (just over 26,000 occurrences), and *a*, as the fifth most frequent wordform in texting, occurs relatively more frequently than *in*, the fifth most frequent in the BNC (18,791 and 18,214 per million words, respectively). This continues down the list[3] with *it* occurring 10,603 times in tenth position in texting compared to *was* with 9236 occurrences in the BNC; and, in twentieth position, *good* occurs 6640 times per million words in texting while *at* occurs 4790 times in the general corpus. In other words, although no wordform in CorTxt is as frequent as *the* in the BNC, the most frequent are generally *more* frequent in texting, suggesting less variety of wordforms and the likelihood, as described later, of a more formulaic language with greater repetition of fixed phrases.[4]

Perhaps the most striking observation that can be made from this list, as well as the fact that *the* is not the most frequent wordform in texting, is that the order of many frequently texted wordforms are the reverse of those in the BNC. In other words, the wordforms swap

[1] Incidentally, the fact that the entries in the texting list are spelling groups affects to only a limited extent their frequencies and order *x* would not occur in the top twenty-five, and nor would *are* (which moves from twenty-eighth position to sixteenth), or *see* (moving from thirty-first to twentieth) or *i'm* (from thirty-fourth to twenty-fourth). Meanwhile, *for* moves up from twelfth to ninth; and *be* up from twenty-second to thirteenth; as a result of this, other wordforms move down, and *will* and *if* are not included in the list used. However, as we shall see below, the inclusion of the respellings does not affect the observations made and conclusions reached in this chapter. See Appendix 7.1 for the ungrouped wordlist.

[2] Care must be taken in assessing these differences, as no indication is given as to where in the lists these items occur; they are listed here simply as an indication of differences between the two. *i'm* (in CorTxt) and *'s* (BNC) are not considered here, as they reflect distinct practices in how words are broken up.

[3] With the exception of *is* in texting, which is slightly less frequent than its counterpart in the BNC, *for*.

[4] Interestingly, however, this distribution pattern is even stronger in speech; that is, the frequencies of the most frequent texted words are not as high as those in spoken corpora (see Leech et al 2001: 144), the tenth most frequently spoken word in the BNC occurring 14,252 times.

positions in the frequency list. Although, for example, personal pronouns are also more frequent across informal and interactive spoken corpora (Leech et al 2001), texting differs from conversation in that *you* is more frequent than *I*.[1] Similarly, *the*, replaced as the most frequently-occurring wordform by *you*, can be found just below *a* rather than being three times as frequent (as *the* is in most corpora). Although these are not the only examples: *have* and *is* are also reversed, as are *me* and *my*, and *just* and *so*, the reversal of *a* and *the* seems more startling not only in their upset of the established order but because of the generally high frequency of *the*, and the implications such a change has for Txt.

The fact that *the* is far less frequent in texting than general language, while *a* retains a similar frequency, is to a certain extent unsurprising: Biber et al (1999: 267) note that *the* varies more across registers than *a*, while Leech et al (2009) document a gradual decline of *the* between the 1960s and 90s. In other words, while *a* remains relatively stable across varieties, frequencies with which *the* occurs vary according to language situations in which it is used and over time. However, in no other corpus is the frequency of *the* quite so low in comparison with *a*. The question as to why their frequencies are so similar, and what this reveals about the nature of Txt, is explored below.[2]

6.4.2 Explaining the frequencies of *a* and *the*

7.4.2.1 Speech-like features of texting

One immediate answer to the above question lies in the speech-like informality construed by texters. As Biber et al (1999: 267) point out, frequencies of *a* and *the* are more similar in conversation (as they are in texting), with 27,351 occurrences of *the* per million tokens in the conversational parts of the spoken BNC, for example, compared to 17,056 of *a* (Leech et al 2001: 223-239). The suggestion that the low frequency of *the* can be explained by the similarity of texting to spoken language is also supported by the absence of *of* in the top wordforms. This is highly unusual for, as Sinclair (1991: 84) notes, *of* usually represents over 2% of all tokens regardless of text-type. In CorTxt, *of* comprises 0.75%, lower even than the figure of 1.72 percent found in romantic fiction (Tribble 2000). The relevance of

[1] This is bearing in mind that we have not included *I* in contractions. If the occurrences of *I* as part of contracted forms (*I'm*, *I've*) are included in the *I* count, the total is still exceeded by the *you*-group: 7884 compared to 6443.

[2] A note on *an*: an occurs less frequently in texting than it does in speech, at 1422 occurrences per million words, compared to 3430 in the BNC. Concordance lines show that it occurs frequently to the left of *hour*, and also with *email* and *amazing time*. Except for noting similarities with the phrases in which *a* will be seen to occur, there do not appear to be any implications of this for the frequencies of *a* and *the*, and *an* is not included in the analysis.

this for the infrequency of *the* is that, according to Sinclair's study, 80% of occurrences of *of* are involved in elaborating nominal groups often pre-modified by *the*, and fewer nominal groups is also a feature of speech. *of* is therefore also less frequent in spoken language than it is in written corpora: occurring 14,550 times per million tokens in the spoken subcorpus of the BNC compared to 31,109 in the written (Leech et al 2001: 144). Other similarities between wordform frequency in spoken language and texting include the aforementioned increased frequencies of *you* and *I*, suggesting interactive, interpersonal language.

Differences between spoken and texted wordform frequencies, however, suggest other features, particular to texting, distinguish it from spoken language and may account for the low frequency of *the* (and relatively high frequency of *a*). These differences include the fact that *you*, as mentioned above, is more frequent than *I* in texting; and, most importantly, in general speech, *the* remains nearly twice as frequent as *a* whereas, in texting, *a* is (slightly) more frequent than *the*. Another factor is the omission of *the* from (possibly speech-like) structures where we would normally, in other text-types, expect to find it.

7.4.2.2 Omission of a *and* the

To explore the extent to which omission plays a role in reducing occurrences of *the*, *a* and *the* were counted across a random sample of text messages. *the* occurred 75 times, and *a* 86.[1] Omissions of both articles were then counted. Assessing where articles are missing was intuitive, and in most cases seemed uncontroversial.[2]

(7.5) Hiya, probably coming home * weekend after next

(7.6) You all ready for * big day tomorrow?

(7.7) Reckon need to be in town by eightish to walk from * carpark.

(7.8) Cant think of anyone with * spare room off * top of my head

[1] This reflects the general word frequency of CorTxt in that *a* is more frequent but differs slightly from the general list in the size of the gap between them: *a* is slightly more frequent in the sample.
[2] Omitted articles are marked throughout the examples with an asterisk.

(7.9) have got * few things to do. may be in * pub later.

(7.10) have * good weekend.

Uncertain cases were disregarded, either where it was not clear whether *the*, *a* or another item was needed, or where it was debatable whether anything was missing at all.

(7.11) Car passed [the/its] mot

(7.12) Currently test driving [the/an] electric car

(7.13) i want her to be able to drink cos she hasn't since having [the/a] car

(7.14) Just got back from [the/?] Exit festival in serbia

(7.15) [the/?] prince of wales beer garden is best

(7.16) will call u when at [the/?] uni

(7.17) am in [the/?] reception

(7.18) am off for [a/the/?] games night

Table 7.5 Realisations and omissions of *the* and *a*

	Realisations	Omissions	Total potential occurrences
the	75	34	109
a	86	16	102

As the table shows, *the* is omitted more than twice as often as *a*, which supports the hypothesis that omission accounts for its infrequency. However, if the omitted occurrences are added to those realised, *a* and *the* still 'occur' only to a similar extent. While this suggests omission of *the* goes some way towards explaining its relative infrequency, it does

not fully explain the discrepancy between its frequency in texting and in general corpora. The question also remains as to why, if the low frequency of *the* is explained in part by structural abbreviation, *a* is not similarly omitted. Why is *a* (at 18,863 occurrences per million tokens) slightly *more* frequent in texting than in conversation? This suggests that the low frequency of *the* cannot be explained solely through abbreviation, just as it cannot simply through linguistic similarities with speech. The distinctive ellipsis which characterises Txt is explored across linguistic elements in Chapter 8: Spoken Grammar, Texted Grammar.

7.4.2.3 Densification of language

A third explanation for the infrequency of *the*, alongside its omission and comparison to spoken data, is that of the more general shift, seen across written texts, towards structures which do not include *the*. Leech et al's (2009) comparison of corpora from the 1960s and 1990s points to decline in the occurrence of *the* of 5.4% and 11.1%, respectively, in British and American written English. This decline is not, however, linked to a corresponding decline in nouns, nor to moves towards more speech-like styles, but can instead be linked to changes within the structure of the noun phrase (Leech et al 2009). The changes, including a rise in the open-class categories of nouns and adjectives and a corresponding decline in the closed classes of determiners, articles and prepositions, point to an increase in lexical density and hence a more densely-paced noun phrase. In particular, decline of *the* can be explained by reductions in the structure *N of N* (as in *the fruit of the coconut palm* and *the behaviour of the patient*) and corresponding rises in the structures *N2 N1* (*coconut palm fruit*) and *N2's N1* (*the patient's behaviour*), as well as an increased use of proper nouns. Also linked to this process is the decline in *of* by 11.3% in AmE and by 4.7% in BrE, in the postmodification of noun phrases.

The explanation posited by Leech et al (2009) for the greater lexical density is that of compression of information for reasons, especially in the the press, of *economy*, or the pressure to condense as much information as possible into the space available (Leech et al 2009: 17). The rise in the number of proper nouns, for example, which includes a trebling of acronyms such as UNESCO, FIFA and WHO across the American and British corpora[1] and a rise in personal names, are convincingly described by Leech as increasingly useful tools for condensing ever-expanding information into brief references, so that *the World*

[1] Compare to Weber's (1986) mention of the rise in acronyms (see Chapter 5).

Health Organization becomes *WHO*; encouraged by a current, Western 'trend towards individualisation'. A related phenomenon is the rise of noun-phrase name appositions, as illustrated by Leech's (2009:15) example of *midnight bather Brian Best* (rather than *Brian Best, the midnight bather*).[1]

Increasing N + N sequences are most obvious in Press headlines such as PIG TRANSPORT OFFENCES MAN FINED (Leech et al 2009: 13), as well as in scientific or technological discourse, although also found across the subcorpora investigated by Leech et al (2009).

Figure 7.4 N + N sequences (Leech et al 2009)

> *Black Country car sales group West Midland motors* [F-LOB A38]
>
> At a *county tour* held during the fall of 1989, *county commissioners*, *conservation district supervisors*, and other local *agency personnel* saw first hand what it takes to successfully establish *CRP cover*. [Frown J70]

The increase also includes a rise in sequences such as *sports car*, *Ramblers Association*, *masters degree* and *weapons purchases* in which attributive nouns retain plural inflections, perhaps (Leech et al 2009: 21 suggests) to aid in interpreting the noun phrase. These too, argues Leech et al (2009), act toward the process of 'densification', including declines in *the* and *of*. Although this explanation remains to an extent speculative and factors other than information compression may prove influential,[2] it is convincing and finds support elsewhere (Biber 2003; Biber and Clark 2002).

This shift has relevance for texting, even assuming Txt reflects spoken rather than written language both in its interactive (rather than informative) function and its verbal (rather than nominal) style. Firstly, texting is characterised not only by speech-like features (including respellings such as *kinda* and *goin*) but by abbreviation. These potentially competing motivations mirror demands which, according to Biber (2003), face the Press (and, according to Biber and Clark 2002, other written discourses) and include not only those of 'economy' but 'popularisation': that is, the need to adopt a colloquial style as well as the pressure to condense information into the space available (Leech et al 2009: 17). Secondly,

[1] And see Jucker (1992: 207-249)
[2] The selection of the genitive, for example, does not always replace that of a N of N structure but, as Leech et al (2009: 26) acknowledge, is determined by 'syntactic, semantic, pragmatic, even phonological' factors.

Leech et al (2009: 6) raise the question as to whether similar trends to those documented in written corpora can be seen in speech. Although nouns are less represented in spoken than written corpora, Leech et al note a significant rise in nouns between spoken corpora from the 1960s and 1990s. As in writing, the increase in spoken corpora can be explained in part by corresponding increases in N + N combinations, which in the spoken corpora include *blood sports*, *career women*, and *brass hats* (Leech et al 2009: 18-19).[1] Problems, however, with exploring this in texting is that it is not possible to retrieve, for example, N + N sequences from the untagged corpus, and the risk that imposing trends seen in other discourses may obscure other patterns in texting.

The actual relevance of this observed compression of information for the present study is therefore not that texting may represent an extension of the shift from N of N sequences towards N2 N1 sequences and the genitive, thus explaining differences in wordform frequency between CorTxt and the BNC (compiled in the 1990s), but that it highlights the possibility that texting comprises structures, or phrases, which do not occur so frequently in the BNC and which do not include *the*. In other words, the argument here is that underuse of *the* in CorTxt and its similarity in frequency to *a* can be explained in part by recurring phrases in which both (but particularly *a*) occur.

7.4.3 Significance of frequent phrases in texting

General differences between texting, and spoken and written corpora, is apparent in lists of the most frequent 3-, 4-, 5- and 6-wordform phrases from CorTxt which can be compared with the lists in section 7.2.3 above. It should be noted that the following were extracted directly from the corpus and wordforms do not represent spelling groups.

[1] What is surprising here, suggest Leech et al (2009: 40), is that spoken language appears in this regard to be following trends in written language, rather than influencing changes in writing.

Table 7.6 Phrases in CorTxt[1]

3-wordform phrases in CorTxt		
1	have a good	231
2	let me know	163
3	see you soon	74
4	a good time	71
5	do you want	68
6	how are you	65
7	let you know	63
8	a good day	62
9	happy new year	60
10	give me a	53
11	looking forward to	49
12	see you later	48
13	and see you	47
14	to see you	46
15	see you in	44
16	on my way	44
17	a good one	43
18	in a bit	42
19	see you then	41
20	be able to	40

4-wordform phrases in CorTxt[2]		
1	do you want to	40
2	have a good day	35
3	let me know when	34
4	hope you had a	31
5	let me know if	29
6	have a good one	27
7	had a good time	24
8	to let you know	23
9	have a good week	23
10	what you up to	21
11	have a good night	21
12	have a good weekend	20
13	happy birthday to you	19
14	had a good day	19
15	just to let you	18
16	have a good evening	18
17	hope all is well	17

[1] Excluded from the list of 3-word phrases are *x x x* (118) (as a non-lexical symbol) and pragmatically incomplete phrases: *you want to* (81), *had a good* (74), *you have a* (61), *you had a* (48) and *to go to* (41).
[2] Excluded from the list of 4-word phrases are *x x x x* (37) (as a non-lexical symbol) and pragmatically incomplete phrases: *you had a good* (31), *you have a good* (28), and *you in a bit* (17).

18	good to see you	16
19	me know when you	15
20	get back to you	14

5-wordform phrases in CorTxt[1]		
1	hope you had a good	21
2	just to let you know	17
3	see you in a bit	15
4	let me know when you	15
5	did you have a good	13
6	let me know if you	11
7	looking forward to seeing you	10
8	happy new year to you	9
9	to let you know that	8
10	hope you have a good	8
11	do you want to meet	8
12	i love u so much	7
13	what are you up to	6
14	trying to get hold of	6
15	or do you want to	6
16	do you want to come	6
17	do you want me to	6
18	do you want a lift	6
19	this is my new number	5
20	the end of the month	5
6-wordform phrases in CorTxt (all)		
1	did you have a good time	8
2	hope you had a good time	7
3	just to let you know that	6
4	do you want a lift tomorrow	5

As in speech, texted phrases appear highly interpersonal, although the most frequent phrase used to organise the immediate interaction, *just to let you know that*, is more reminiscent of

[1] Excluded from the 5-word phrase list are pragmatically incomplete phrases: *you had a good time* (13), *you have a good time* (11), *u had a good day* (8), *you have a good night* (5) and *you had a good weekend* (5), as well as *have a good week x* (7).

written notes than speech. Generally, however, compared to the immediate 'I and you' world suggested through spoken chunks or the 'world-out-there' represented by written clusters, frequent texted phrases are used to arrange a shared, often future, world pursued outside the current texted interaction.

Table 7.7 Function of chunks in CorTxt

Function	Chunks used
Expressing wishes and sentiments	*hope you had a good* (*time*), *did you have a good* (*time*), *happy new year to you*, *happy birthday to you*, *hope you have a good*, *have a good* (*day*, *week*, *one*, *weekend*), *hope all is well*, *i love you so much*
Framing current interaction	*just to let you know* (*that*)
Future (or other) shared interaction	*see you* (*soon*, *then*, *in a bit*), *looking forward to seeing you*, *do you want to meet*, *or do you want to*, *do you want to come*, *do you want me to*, *do you want a lift* (*tomorrow*), *good to see you*, *on my way*
Future (texted) communication	*let me know when you*, *let me know if you*

The following overview of *a* and *the* in CorTxt focuses on how their frequencies can be explained in terms of the phrases in which they occur.

7.4.4 Patterns of *a* in texting

7.4.3.1 Overview

Investigation of *a* in texting reveals a phrasal use seemingly specific to texting in that, unlike spoken and written corpora, *a* occurs most frequently in formulaic and phatic phrases such as *have a good day* and *give me a ring*, as well as in interpersonal hedges such as *a bit*. Before looking more specifically at collocates of *a* and the phrases in which it occurs, randomly-selected occurrences of *a* in context indicate the proportion of occurrences of *a* in frequently-occurring phrases.

Figure 7.5 Random sample of concordance lines for *a*

```
1              t outside Hey buddy, us a bell when you finnish
2            st arrived in and there's a couple of things we n
3                oman and sounded like a pile of snot and tears.
4             er Katie here and i have a new number. This is it
5             at kav so there might be a bit of a selly oak poss
6                 me posted xx Depends a little whether nik joins
```

7 mum company … **Have a good time** xx Mystery
8 nd of person who needs a smile to brighten his
9 in. We bought **a couple a** Xmas pressies. Wot u
10 da throw the night. **have a good one** x Cool cool.
11 ing.she's better now stil **a bit** dazed i think.i had
12 sumed we were getting a lift… Want me to boo
13 kered so better not. **Hve a gud time** tho Yipee i'v
14 ob than dave! Fab. **Have a good week,** don't work
15 ow tomorrow. Did u hav a successful shoppin tri
16 know I have just bought a fish and chip and mus
17 ow the address **give me a txt**. steve Hey hav se
18 nything about it.x **Have a good day** love to all Y
19 orean girl moving down a class tomo, poss new
20 u know wot i mean! **Wot a slapper** eh?! X Happy

The phrases above can be divided into four groups: firstly, *give us a bell* and *give me a txt*; secondly, *a couple of, a pile of, a bit of a, a little, a couple a,* and *a bit*; then *Have a good time, have a good one, Hve a gud time, Have a good week,* and *Have a good day*; and finally *wot a* (*slapper*). These phrases, most of which occur frequently throughout CorTxt, account for fourteen of the twenty concordance lines (70 percent). Although this is only one small sample, it suggests that any attempt to explain the frequency of *a* in texting must make reference to the phrases in which it occurs.

3.4.3.2 Collocates of a

Collocates can be an interesting starting-point in exploring how a wordform is used in particular contexts, although care must be taken not to extrapolate too much from the results without recourse to concordance lines.[1] Collocates of *a* in spoken English (as represented by 20,000 sample lines of the BoE's *brspok* subcorpus) are as follows.[2]

Figure 7.6 Collocates of *a* in brspok (BoE)

Quantifiers:	lot, bit, little, few, couple, hundred
Prepositions:	of, as, for

[1] It can only be assumed by looking at a list of collocates, for example, that the co-occurrence of *a* with *job* indicates the sequence *a job* (concordance lines show that although job overwhelming occurs to the right of a, as part of a noun phrase of varying lengths, the structure *job as a* also occurs: *after you left that job as a projectionis where did you move to?*.

[2] The collocates were calculated using t-score (on the basis that co-occurrence is higher than would be expected, statistically) rather than an MI measurement, as the latter does not take into account absolute frequency and therefore produces a list of very infrequent words occurring in the span of *a* and headed by *sprain, crotchet* and *scorpion*.

Noun modifiers:	a, good, big
Nouns:	job, week, problem
Verbs:	had, have, was, got, is
Adverbs:	quite
Other:	there

Collocates include quantifiers as well as the adverb *quite*, which could presumably be fulfilling a similar quantifying function, and some prepositions, noun modifiers, nouns and verbs. A common structure suggested by the collocates is *There is a* ….. .

Collocates in texting, generated through WordSmith's Concord, reveals differences in use which begin to highlight specific ways in which texting is construed.[1]

Figure 7.7 Collocates of *a* in my text message corpus

Quantifiers:	bit
Prepositions:	for, with
Noun modifiers:	the, good
Nouns:	day, time, night
Verbs:	have, had, was, hope, got, get, see
Adverbs:	just, not
Pronouns:	you, I'm
Conjunctions:	and, but, that

Firstly, the only quantifier to collocate with *a* is *bit*; and nor does *a* collocate with preposition *of* which suggests fewer elaborated nominal groups. The nouns which collocate with *a* are distinct from those in spoken language, and are references to periods of time (compared to *job*, *week* and *problem* in the spoken corpus). There are more verbs in the list than among the spoken collocates, of which the most frequent is *have* while others include *hope* and *see*;[2] but not *is*, nor *there.* There are also several conjunctions, all of which suggests a degree of the clausal complexity identified by Halliday (1985) in spoken language, rather than the nominalisation described as more typical of writing. *Just* and *not* are collocates, but not *quite* (which, as we shall see with *a bit*, may reflect a tendency to

[1] Unlike the access tool to the Bank of English, Wordsmith calculates collocates according to their raw frequency. These were calculated using a 5-word span (as were those below for *the*).
[2] These are particularly frequent verbs in texting, as I commented on previously in Tagg (2007).

hedge and soften rather than intensify statements), and personal pronouns: *you*, in fact, is the most frequent collocate. As described below, most of these collocates combine with *a* in the phrases which characterise texting.

3.4.3.3 Phrases with a *in texting*

Phrases in which *a* most frequently occurs were identified by generating and sorting concordance lines in Concord (one place to the right of the node), and then manually counting the most frequent recurrences.[1] The most frequent 3-wordform phrase *have a good* occurred 236 times, followed by various references to time.

Figure 7.8 *Have a good* … phrases in texting

Have a good	(236)	day	(35)
		one	(30)
		time	(30)
		week/wk	(24)
		night/nite	(22)
		weekend/wkend	(21)
		evening	(18)
		trip	(8)

There are also seven occurrences of *have a gud*, usually followed by *nite*.

Figure 7.9 Concordance lines for *have a gud*

1	ean a nurse's uniform!" "**have a gud** nite. Will ri
2	time at tracy's wedding. **Have a gud** nite 2nite."
3	d a half pounds tonight. **Have a gud** time at your
4	ve ya more than i tell ya **have a gud** nite xxx" "
5	bably won't happen! Lol. **Have a gud** nite at work
6	E268!How funnys that! **Have a gud** Nite!Hope th
7	ber." "Got new phone. **Have a gud** nite" "I tak

The pattern is at once formulaic and phatic, and also one with some flexibility and creativity (besides that of respellings), as shown by more unusual instances in the

[1] This is possible given the size of the corpus; with a bigger corpus, a sample of lines could be used or, alternatively, the Cluster Tool in Wordsmith (the drawbacks being that it does not recognise any variation and includes all 3-word clusters within the node's span: ie, they may not include the node word itself).

concordance lines below: *Have a good strike!* (in reference to union action, line 8 below) and *Have a good debriefing!* (line 11).

Figure 7.10 Selected concordance lines for *have a good*

```
1         em is, i wasn't alone ..! Have a good weekend a
2              Lunch?steve Hope you have a good night! see
3            t was you. Speak soon. Have a good week. Xxxx
4         ike it not. Haha. Try and have a good day, i kno
5         n sort it as we go along! Have a good evening x
6           it. See you later mate. Have a good 1. i use m
7         yeah its all good so far. have a good day aswell
8               . Going to bed now. Have a good strike!  Sorry,
9        The jury is still out! Ok. Have a good journey. I
10           re crossword clues and have a good week x Do
11            decent weather for me Have a good debriefing!
```

There is also variation in the adjective used: *nice* occurs thirty times to describe *lunch* and a *meal* as well as periods of time; *great* occurs twenty-eight times, most frequently in *have a great time*; *lovely* fifteen times, most often alongside birthday or Christmas wishes, as in *have a lovely birthday*; or *Happy bday 2 u! Hope u have a lovely day*. *Fab* occurs fourteen times, predominantly in *Have a fab day.* Other less frequent phrases include: *hope you have a spiffing good night*; *a smashing day*; *a wicked time*; and *a wonderful afternoon*. In other words, the 'prototypical' formulaic phrase *have a good day* forms the basis of variation and creativity throughout CorTxt.

Another frequent phrasing is *give me a* (53 occurrences) or *give you a* (21), which refer to future phone calls, text messages or other acts of communication. As such, these serve to bolster participants' ongoing relationships and co-ordinate their wider interaction (including communication beyond texting).

Figure 7.11 *Give you/me a …* in texting

give me a (53)	shout	(11)
	call	(10)
	text/txt	(8)
	ring	(6)
	bell	(5)

give you a (21)	ring	(9)
	text	(4)
	call	(3)
	shout	(2)

Concordance lines for *give me a* and *give you a* show the former are predominantly imperative (*Ok give me a text*) although also occurring in requests with *could you*, while the latter are prefaced with *will* (or *I'll*).

Figure 7.12 Selected concordance lines for *give you/me a*

1	ound tomo afternoon. i'll	**give you a**	ring x All don
2	ell if you choose to then	**give me a**	shout and i'll
3	Ok just walking up x Ok	**give me a**	text when you
4	t me know when you I'll	**give you a**	text later cos
5	ed to know the address	**give me a**	txt. steve He
6	w? Hey tagg. Could you	**give me a**	call when you
7	e shopping successful.	**Give me a**	warning when
8	of gays I mean games!	**Give me a**	bell when ya
9	, i've finished work now,	**give me a**	shout when you
10	saturday. Not sure yet	**give me a**	second Loads
11	ill meeting later and will	**give you a**	ring after you
12	xx What's happening?!	**Give me a**	missed call

The exception in the above concordances is *give me a second* (line 10).[1] *Give me a missed call* (line 12) is interesting as a phrasing which has surely come about with the use of mobile phones, to refer to the act of ringing a phone without it being answered in order to pass on your number.[2]

While *have a good day* and *give you/me a call* highlight quite explicitly the interpersonal purposes of texting, of maintaining and managing ongoing relationships, the hedge *a bit* shows the interpersonal sensitivity which shapes how texters construe text messages. In spoken communication, *a bit* does not simply serve to specify amount but instead, as Carter and McCarthy (2006: 64) state, it is 'deliberately vague and informal' and serves to soften language otherwise deemed by the speaker as direct or overly authoritative, as well as

[1] Which interestingly refers to time, a common concern in texting, but does not represent a frequent pattern in terms of the present phrase

[2] And also leading to the convoluted but now frequently used phrase: *I've had a missed call from you* (rather than *I missed your call*).

hyperbolic or metaphoric (Carter 2004b). *A bit* can also be described as playing an evaluative role, in that (when used with non-comparative adjectives or determiners) it marks statements as 'negative or critical', as pointed out by Carter and McCarthy (2006: 65), who give the examples: *It's a bit extravagant, isn't it?* and *He's a bit old to be driving*.

Concordance lines below from CorTxt suggest that *a bit* (202 occurrences), when used with non-comparative adjectives, occurs predominantly in texting with negatively-oriented adjectives such as *dazed*, *lame* (meaning *weak* or *pathetic*, as in *a lame excuse*), *miserable*, *flat*, *bizarre*, *useless*, *gay* (in its derisive sense, similar to *lame*), and *mind boggling*.

Figure 7.13 Selected concordance lines for *a bit*

1	ing.she's better now stil	**a bit**	dazed i think.i had
2	hat she did, which was	**a bit**	lame, i thought tha
3	e fun! Am ok tho feeling	**a bit**	miserable in my bi
4	ws, am ok so will try for	**a bit of**	independence an
5	y local unfortunately it's	**a bit**	flat, infact it has n
6	ank you and see you **in**	**a bit**.	10mins is your fri
7	ng anything. See you **in**	**a bit**.	How's it going? G
8	it? Well we are running	**a bit**	late… it turns out t
9	rial. But she talks **quite**	**a bit**.	So its ok. Any ad
10	i'll come up seaview **for**	**a bit**.	Xxx Just finished
11	rrow when I hope to feel	**a bit**	better Just got a te
12	eed2 ask u something,	**a bit**	bizarre just say se
13	en u would hear by? Im	**a bit**	use less 2 u really
14	te! Ok maybe 5.30 was	**a bit**	hopeful...we'v only j
15	ink Jeremy Clarkson is	**a bit**	sexy now! hope y
16	u guys later. See you **in**	**a bit**.	NAME298 Lol! Nah was
17	NAME46 for shopping **in**	**a bit**.	You lot? I reckon
18	o play with x Well thats	**a bit**	gay.. Make a tin fo
19	offered it to me already-all	**a bit**	mind boggling really...
20	avour ... Yep see you **in**	**a bit**.	Ex Might not have
21	an comes tomoz up for	**a bit of**	motown madness?

Negative connotations can also be perceived in other, less obviously derogatory text messages such as line 14 above: *Ok so maybe 5.30 was a bit hopeful* (that is, they did not make it on time), while the seemingly positive evaluation (line 15) that: *Jeremy Clarkson is a bit sexy now!* is marked as grudging or cautious with reference to the full text message: *Did you watch Top Gear? It was so funny, even I think Jeremy Clarkson is a bit sexy now!*. As well as negatively evaluating, the above uses of *a bit* serve to downplay assertions. In

contrast, *But she talks quite a bit* (see above, line 9) can be interpreted as *But she talks a lot* (see Carter and McCarthy 2006: 128). *Quite* is much less frequent than the hedge *a bit*, with only one other occurrence of *quite a bit*: *I've got quite a bit of work 2 do, but the does sound tempting*. This suggests, along with the generally low frequencies of *quite* and of quantifiers such as *a lot*, that texters tend towards softening rather than intensifying their language.

A bit of (occurring 34 times) serves as a hedge but, as evident in lines 4 and 21 above, is less evaluative than *a bit*. It softens the strength of the texter's ambition in: *so will try for a bit of independence* (along with the verb *try*); and plays down the extravagance of the entertainment promised by *up for a bit of motown madness?* (rather than the more unmarked *some motown madness*). Similarly, *in a bit* (occurring 41 times) and *for a bit* (18 occurrences) modify the length of time involved and thus construe more intimate language.

Interestingly, however, within the above examples, *a bit of a* (20) reverts to the negative connotation also implied by *a bit*, as seen in the following concordance lines.

Figure 7.14 Selected concordance lines for *a bit of a*

1	Well done. Often	**a bit of a** challenge to do it
2	rd. i know this is	**a bit of a** cheek especially
3	eems to be. I had	**a bit of a** cold but gettin bet
4	y touch! feel like	**a bit of a** cow. But ta for th
5	. I know I've been	**a bit of a** drip but I will start
6	dy wife are having	**a bit of a** get together satur
7	at might have been	**a bit of a** mean message to
8	thiing has proved	**a bit of a** mistake i think- lo
9	il out tonight for	**a bit of a** piss up, he's not
10	cancel lunch. Have	**a bit of a** problem. Call you

Use of *a bit* in texting, then (rather than *a lot* or *a few*), weaves into the construal of texting a concern for interpersonal considerations, both in softening expressions and in evaluating them, just as the phatic (but potentially creative) phrases: *have a good time (*or *day, week, weekend)* and *give me/you a shout* (*ring* or *text*) illustrate the role of texting in building and maintaining relationships. These interpersonal phrases are very frequent in texting. As illustrated in concordance lines at the beginning of this section, as many as 70 percent of total occurrences of *a* in CorTxt can be described as phrases, and this phrasal use of *a*

(along with abbreviation and the relative infrequency of elaborated noun groups) accounts for its high frequency in comparison with *the*. As outlined below, although the use of *the* in texting in many ways parallels that of *a*, the phrases in which it occurs are less frequent than those with *a*.

7.4.4 Patterns of *the* in the construal of texting

7.4.4.1 Overview

A phrasal approach to the description of *the* in texting (that is, classifying uses of *the* according to lexical patterns in which it occurs) can be justified not only because it offers a useful comparison with the more strongly phrasal use of *a* in CorTxt, but also by the limitations involved in classifying uses of *the* on the basis of its functions (Biber et al 1999: 263-5; Kachru 2003; Poesio and Vieira 1998; Quirk et al 1985), both textual (anaphoric and cataphoric reference) and contextual (situational reference). Firstly, it becomes necessary to add other categories: Kachru, for example, also includes what he vaguely calls 'unfamiliar noun phrases' such as *the fact that* or *the OP ED page of the NY Times* (which Biber et al classify as cataphoric); as well as use of *the* in idioms; while the superlatives *the last bus to Denver* and *catch the last bus*, which Quirk et al call 'the logical use of *the*', Kachru labels 'fixed collocations'. Non-native learners of English, meanwhile, appear to use *the* in a sense that differs to conventional descriptions (and native use): according to Dewey (2007), *the* is used in ELF (English as a lingua franca) to emphasise the keyness or significance of words or topics in the current interaction rather than in structures where *the* could be deemed superfluous, such as *the first* (Dewey 2007). Although Dewey does not consider article use in the first languages of the speakers contributing to his data and the potential effect of this on their use of *the*, his findings find corroboration in other research (Seidlhofer 2004, also cited by Dewey 2007). A more effective description of *the* may therefore start from analysis of the lexical patterns or phrases in which it occurs, reflecting Sinclair's (1991) approach in his investigation of *of*.

Analysis of phrases in which *the* occurs in texting reveals prepositional phrases referring to time and place, rather than elaborated nominal groups in subject or object position, while the frequency of *by the way* reflects again the importance of interpersonal considerations in texting. None, however, of these uses of *the* is as frequent as phrases in which *a* has been seen to occur. The relative lack of patterning can be seen in the following randomly selected concordance lines.

Figure 7.15 Selected concordance lines for *the*

1	t looks like you are ill **in the morning** ? X Neath
2	like cluedo but I'm sure the bishops won't mind.
3	imbo waiting to cross **to the other side**. Hence th
4	he lion says when I roar the whole jungle shakes
5	at list ..? Can you bring the suitcase to work. X
6	d little friends in town **in the afternoon**/evening.. J
7	little thing maybe? X Is the cd somewhere i can
8	u c d little thing i left in the lounge?it was free b
9	r lives together call u **in the morning** xxx am ho
10	! And lo, it is printed, n the computer switched o
11	loads of exercise shit **in the summer** when are yo
12	ldlife expert. What job? The one in Belgium? Ok
13	ondon office for work for the next week.PHONENUMBER
14	london, and u shall hav the grand slam! X I am
15	r long, me! X Why dont the txts deliver? Maybe r
16	g. Would u be watching the rugby? Dunno if its
17	ow's the house without the chicks. quiet and tid
18	ook soonish and tell me the date i received my
19	will look out for NAME146. **All the best**! No definately-
20	after screaming child **at the mo**! u two gals run

The above concordances represent a common theme throughout CorTxt in terms of the frequency of references to time: *in the afternoon*, *in the morning*, *in the summer*, and *at the mo*; and place: *to the other side*; as well as social formulae *All the best!* The lines are also representative in that a smaller proportion of occurrences can be described as phrases than for *a*: in the above set, seven of the twenty lines are labelled phrases (35%), compared to fourteen with *a* (70%).

7.4.4.2 Collocates of the

If we start, as for *a*, by comparing collocates of *the* in spoken English (again, as represented by *brspok*) we can begin to highlight specific ways in which *the* is used in texting.

Figure 7.16 Collocates of *the* in brspok (BoE)

Prepositions:	of, in, on, at, from, by;
Nouns:	end, way, thing, time, moment, side, world, one
Pre-modifiers:	same, first, other, whole, main, most
Verbs:	is

The collocates with several prepositions, most frequently *of*; several nouns such as *end*, *way* and *thing*; and pre-modifiers including adjectives, determiners and superlatives. The only verb is *is*. All this suggests frequent occurrences of *the* in elaborate nominal groups such as *the end of the world* and in lengthy adjuncts such as *at the time*, as well as certain phrases associated with the spoken mode: *by the way*, perhaps, and *the thing is*.[1] Collocates of *the* in texting, however, indicate a weaker tendency for *the* to occur in elaborated nominal groups.

Figure 7.17 Collocates of *the* in texting

Prepositions:	for, with
Pre-modifiers:	the, all, good
Pronouns:	you, I'm
Conjunctions:	and, that, but
Verbs:	have, was, are, will, get, see, going, can
Adverbs:	just, about

Firstly, there are few prepositions and those that feature (*for*, *with*, and *about*) are different from those in *brspok* (*of*, *in*, *on*, *at*, *from*, *by*). Most noticeably, *the* does not collocate in texting with *of*, nor with nouns or many noun modifiers, suggesting a lack of elaborated nominal groups. The collocates instead comprise several verbs (*have*, *was*, *are*, *will*, *see*, *going*, *can*), pronouns (*you, I'm*), adverbs (*just* and *all*) and conjunctions (*and*, *but*), rather than the prepositions, nouns and adjectives seen in *brspok*, suggesting greater use of definite noun phrases in speech compared to texting and, again, clausal complexity in texting normally associated with spoken language. The other observation that analysis of the collocates supports is that patterns in which *the* occurs are less frequent than those in which *a* occurs: neither *in* nor *morning*, for example, are collocates and yet (as described below) *in the morning* is the most frequent phrase in which *the* occurs.

7.4.4.3 Phrases with the *in texting*

The most frequent phrases in which *the* occurs are prepositional phrases associated with time and place, as well as certain spoken discourse markers, and formulaic phrases. This again highlights social and interactive functions of texting, and marks the speech-like

[1] Bearing in mind that these observations are speculative, and that the use of the collocates would need to be checked with concordance lines.

interpersonal sensitivity of texters. However, no one phrase seems to occur particularly frequently, and not as frequently as those with *a*, and it is this which helps explain the unusual relative frequencies of the two. For example, the most frequently occurring phrase, *in the morning*, occurs only 35 times, compared to 236 occurrences of *have a good* (the most frequent phrase with *a*).

Figure 7.18 Concordance lines for *in the morning*

1	me up before u leave	**in the morning**	i need to get up
2	If not give me a call	**in the morning**	to make sure I wake
3	y, prob you'll see me	**in the morning**.	F x Yep will do
4	. How about meeting	**in the morning**	for s&m, i go
5	lives together call u	**in the morning**	xxx "its yours
6	a picture of it to you	**in the morning**	x OK seeya soon"
7	possibly pick me up	**in the morning**.	My car is not fi
8	u r re lesson! See u	**in the morning**.	F x "I'm saving my
9	mselves Yes ok will do	**in the morning**	"Only a couple of
10	They want you to ring	**in the morning**	if poss" "The bank is

This seemingly referential phrase in fact plays a role in co-ordinating participants' interaction. Other references to time include *at the moment* (with 27 occurrences) and the slightly more frequent *at the mo* (33 occurrences).

Figure 7.19 Selected concordance lines for *at the moment/at the mo*

1	Dying of pneumonia	**at the moment**	well got alittle cold
2	fine, on holiday	**at the moment**	just relaxing. I have
3	having a shit time	**at the moment**	but we will have it
4	able to do much	**at the moment**	as my CRB check has
5	get hold of anyone	**at the moment**.	I'll have to let
6	Doing my career plan	**at the mo**.	Thats an interesti
7	hard at college work	**at the mo**.	I've got so much to
8	in hospital. She is ok	**at the mo**	r u home alone
9	orague? I'm in ireland	**at the mo**	with the in laws! Go
10	is beta. I'm at home	**at the mo**,	that's why i wanted

The number of times *the* is omitted from these phrases is also interesting. While *at moment* occurs only three times, *at mo* occurs twenty-three times throughout CorTxt: over two thirds again the occurrences of *at the mo*.

Figure 7.20 Selected concordance lines for *at mo*

1 ng to gym and I'm in town **at mo**.So can we make it for
2 run here with my bro but **at mo** got stinking cold. Have
3 Can't return the camera **at mo**. Prob. Be able to make
4 chine! And sleeping badly **at mo**. Observation nxt wk -
5 Naughty gal! Yes, on bus **at mo**. Are you still in town?

This phrase does not occur in the BoE[1] and so its occurrence in texting would appear to illustrate the role of abbreviation in the low frequency of *the*.

Other most frequent phrases are *the end of* (19 occurrences) and *the rest of* (17) which, as these concordance lines show, refer predominantly to time. Again, the phrases are often involved either in coordinating shared activities (*it should be back by the end of week*) or as part of phrases similar to *have a good* (*enjoy the rest of your day*).

Figure 7.21 Selected concordance lines for *the end of* and *the rest of*

1 it should b back **by the end of week** - not happy
2 issue delivered **by the end of the month**. Deliver
3 nt to the run up **to the end of term**! Nothing els
4 near coming **near the end of Cornwall** now, so

1 looking forward **to the rest of the month**! See
2 to do? What **about the rest of me**?! What day
3 e thats cool, enjoy **the rest of your day**. Hey y
4 ice weekend, enjoy **the rest of your night** and
5 Should be in for **the rest of the evening** t
6 should be nice for **the rest of the week**.

Looking first at *the end of*, those that do not refer to time often refer to physical locations: such as *near the end of Cornwall*; while exceptions with *the rest of* include the presumably light-hearted *the rest of me* above. Again, as with *at the mo*, it is interesting to note the frequency with which *the* is omitted from these phrases. There are eighteen occurrences of *end of* in CorTxt, and ten of *rest of*. Although the phrases do occur in the BoE they are by no means as frequent as in texting.

[1] Except in reference to staying with Mo Mowlem!

Figure 7.22 Concordance lines for *end of* and *rest of*

```
1          nds good! Txt u towards end of day re exercise.
2          spuds r gud! Only until end of today. This is fro
3          ess. Won't find out til end of july - another co

4          'ham. Have a jolly good rest of week xx Hello! T
5          but Sat fine too. Enjoy rest of hols x We're her
6           orting tomo now. Enjoy rest of wk! E x Hope w
```

Other frequent phrases with *the* in texting include prepositional phrases referring to place or to travel: *on the way* (17), *on the bus* (6), *in the car* (10), *to the cinema* (13), *at the pub* (12), *in the pub* (18), and *on the train* (12), among others. As with time phrases, these highlight the role of texting in coordinating social and mundane arrangements (as described by Ling and Yttri 2002), and that texting is about doing things, through verbs, rather than elaborate nominal groups.

The twenty-nine occurrences of *by the way*, a discourse marker involved in organising text segments, are interesting because, while the time and place adverbials explicitly reflect the purpose of texting, *by the way* shows how its interpersonal nature shapes the language in a similar way as with *a bit*. As Carter and McCarthy (2006: 262) point out, the phrase is 'particularly frequent in informal spoken contexts'. Assuming that the phrase is used in a similar way to that identified in speech, *by the way* also indicates interpersonal sensitivity on the part of texters. Specifically, *by the way* shows how text segments it introduces relates to previous sections: namely, that it 'represents a temporary digression from the previous segment or a shift in topic' (Carter and McCarthy 2006: 262)

In 17 text messages, *by the way* precedes the 'text segment' it introduces, as below.

(7.19) Not yet. Just i'd like to keep in touch and it will be the easiest way to do that from barcelona. By the way how ru and how is the house?

(7.20) That way transport is less problematic than on sat night. By the way, if u want to ask NAME37 n NAME561 to join my bday, feel free. But need to know definite nos as booking on fri. X

(7.21) Jolly good! By the way, NAME57 will give u tickets for sat eve 7.30. Speak before then x

In the remaining 12, *by the way* follows the segment.

(7.22) At home by the way

(7.23) Have you bookedthe hut? And also your time off? How are you by the way?

(7.24) Am on a train back from northampton so i'm afraid not! I'm staying skyving off today ho ho! Will be around wednesday though. Do you fancy the comedy club this week by the way?

It is preceded in 3 cases by *and*.

(7.25) Thank you. And by the way, I just lost.

(7.26) Oh and by the way you do have more food in your fridge! Want to go out for a meal tonight? X

(7.27) Indeed and by the way it was either or - not both !

In 3 cases, *by the way* opens the text message, suggesting a link to previous text messages or to communication through other means.

(7.28) By the way, 'rencontre' is to meet again. Mountains dont....

(7.29) By the way, i've put a skip right outside the front of the house so you can see which house it is. Just pull up before it.

(7.30) By the way, make sure u get train to worc foregate street not shrub hill. Have fun night x

Five text messages relate to texter identification, and this seems a more text-specific use of *by the way*.

(7.31) Not tonight mate. Catching up on some sleep. This is my new number by the way. NAME318

(7.32) Hello hun how ru? Its NAME428 here by the way. Im good. Been on 2 dates with that guy i met in walkabout so far. We have to meet up soon. Hows everyone else?

(7.33) yeah sure thing mate haunt got all my stuff sorted but im going sound anyway promoting hex for NAME186.by the way who is this? dont know number. Joke

Again, it is interesting to note that this frequent phrase also occurs in a form without *the*: the abbreviation *btw* (8), which may help to explain the low frequency of *the*.

Figure 7.23 Selected concordance lines for *btw* in texting

```
1          iff and its cold! What does btw stand for? I'm just
2          ian network! All fine here. Btw, am back friday morn
3            tty fab thanks 2 me! Hehe btw NAME87 where were
4          Tee hee. But ta for invite. BTW, running late and mi
5            elly, how r u? Hows work, btw how r they 4 staff
6           ws on what ur doin 2nite?! Btw do i nt get x x x's
7            monday, probably all day, btw did you know this we
8            wot the details  Yeh yeh. btw NAME13s access have
```

The implications for Txt are that it comprises a language characterised not by subjects or objects realised by elaborated nominal groups (which include *of*), but by adjuncts (time and place adverbials such as *in the morning* and *at the mo*); and by discourse markers typically associated with the spoken mode, chiefly *by the way*, alongside variants without *the*: *at mo* and *btw*. As with phrases comprising *a*, we see how interactive functions of texting produce phatic and formulaic phrases, while the occurrence of phrases such as *and by the way* reveal a degree of interpersonal sensitivity on the side of the texters. And crucially, in explaining wordform frequency, we see a less strongly phrasal use of *the* than of *a*.

7.4.5 Summary

The starting-point for this investigation was the highly unusual, if not unique, way in which the typically most frequent wordforms are distributed in texting. The example explored was *a* and *the*, the frequencies of which are not only similar, but reversed: *a* is more frequent than *the*. This can be accounted for only in part by features of texting more typically associated with the spoken mode (such as the infrequency of elaborated nominal groups suggested by the fewer occurrences of *of*), as well as the frequent omission of *the* or abbreviations such as *btw*. Of greater relevance is the distinctly text-specific phrasal use of *the* and, to a greater extent, *a*. This reflects findings by Leech et al (2009) regarding the decline in *the* across written corpora, which they explain with reference to shifts in commonly-used structures, from those including *the* (*the N of the N*) to those which do not (*N2 N1*, for example). The phenomenon is, however, somewhat different in texting. Texting comprises a high frequency of interpersonal phrases, which account for a third of occurrences of *the* and 70% of *a*. These include phatic (yet creative) formulae such as *Have a good day* and *give us a ring*, repeated reference to time and place in *at the moment* and *on the way*, all of which play roles in managing participants' ongoing relationship and coordination of activities, and apparent signs of interpersonal sensitivity in *a bit of* and *by the way* by which participants manage the present interaction. This role of multi-wordform strings in oiling the wheels of texted interaction, returned to below, reflects observations made by other researchers into spoken and written data (Wray 2002; O'Keeffe et al 2007). Crucially for understanding the frequencies of *a* and *the*, however, phrases with *a* are much more frequent than those with *the* and this, along with spoken features and abbreviation, explains their frequencies.

This chapter thus highlights that abbreviation is but one feature in the construal of texting. As described above, article omission is neither comprehensive nor consistent, and appears alongside features characteristic of texting such as phatic social formulae, hedging, and an almost speech-like performance of discourse markers. What we see being construed through texting is the performance of brevity and spoken informality through communicative features more typically associated with the spoken mode, as well as features specific to texting. The combination of these features reverses the frequencies of *the* and *a*.

7.5 Chapter Summary

Methods followed in this chapter aim to be corpus-driven, in that they involve computer-generated wordform frequency lists as a way in to the data, thereby delaying human intervention in the decision as to which features of the language variety to explore. The extent to which this study proves a framework for analysis of other varieties is limited, given the fact that its direction was determined by wordform frequencies apparently specific to CorTxt.[1] In this case, the similar frequencies of *a* and *the*, with *a* slightly more frequent than *the*, formed the basis of investigation. This investigation does not claim to be comprehensive, and further study could be conducted into other wordforms with 'reversed' frequencies: *you* and *I*; *have* and *is*; *me* and *my*. What is does intend to show, however, is insights which approaches based on wordform frequency can provide; and the fact that frequencies can be explained by phrases in which wordforms occur.

Observations regarding the recurrence of certain wordform sequences across corpora has focused attention on the significance of phrases, or multi-wordform strings, in describing language; and their distribution across corpora, it is suggested, creates linguistic fingerprints of language varieties: frequent spoken phrases, for example, are concerned with managing interaction and relationships between participants, and written ones with representing external reality (O'Keeffe et al 2007: 68). The fact that strings recur at a higher frequency in spoken data than in written can be explained by the observation that multi-wordform strings are primarily suited for use in the oiling of social interaction (that is, they have an expressive function), which is a greater concern for speakers. It should perhaps come as no surprise that phrases in texting are particularly frequent, and that they are chiefly involved in the management of texters' interaction. Where texting differs from speech lies in the focus on formulaic wishes such as *have a good day* and *enjoy the rest of your week* and the concern with managing participants' wider, often future interaction and careful leave-taking, seen in expressions such as *give me a ring* and *see you in the morning*. Use of these phrases, alongside performances of brevity and spoken informality, adds to the construal of texting identities.

The previous chapter on everyday creativity extended the widely-accepted notion that creativity in Txt is limited to respellings and showed how texters manipulate wider

[1] See Chapter 3 for discussion of the extent to which truly corpus-driven investigations can be extended across varieties.

language patterns to create highly evaluative effects. This chapter complements the focus on respellings and creativity by exploring the typical, 'formulaic' lexical patterns that construe Txt. Thus, the combination of the three studies of CorTxt encompasses both the typical and the unexpected: the typical, perhaps, laying out the backdrop from which creativity emerges. The following chapter, by applying a model of spoken grammar to the investigation of Txt, both adds to the description of what is typical whilst reminding us that the nature of the approach adopted determines to an extent what is found.

CHAPTER EIGHT
SPOKEN GRAMMAR, TEXTED GRAMMAR

8.1 Introduction

This chapter explores texted grammatical features in CorTxt through application of the description of spoken grammar outlined by Carter and McCarthy (2006), some of whose observations find parallels in Biber et al (1999). The motivation behind the use of their framework in this study is twofold. Firstly, initial observation suggests texting (like speech) cannot fully be described using current grammatical models which largely remain based on descriptions of written language, but instead requires reference to an integrated model of spoken and written grammar which accounts for the interpersonal nature of speech, its linear structure and non-clausal elements. Secondly, Carter and McCarthy's description is, to a greater extent than other models used to describe spoken grammar (e.g. Brazil 1995), grounded in the data it aims to describe. Investigation in texting of features characteristic of spoken language may seem to contradict the need to let the data in question drive language description, but Carter and McCarthy show through example how new or different features can be included within, and build upon, existing frameworks. The current study therefore places texted grammar within an integrated model of written and spoken varieties. In particular, as well as highlighting how texters draw on spoken features to perform speech-like informality, use of the spoken model shows how texted grammar differs not only from written but spoken grammar and highlights other factors or motivations involved in shaping Txt.

This chapter therefore explores texted grammar in particular comparison with speech. It starts by outlining the need for an integrated model of speech and writing, and exploring models attempting to incorporate spoken features (Halliday 1994; Brazil 1995; Sinclair and Mauranen 2006;[1] Carter and McCarthy 2006). Investigation of CorTxt using Carter and McCarthy's framework reveals spoken features alongside others not typically found in spoken language. Conclusions are reached as to the factors, alongside texters' exploitation

[1] Although Sinclair and Mauranen's linear unit grammar should perhaps be described as a 'pre-grammar' in that it processes the text for subsequent grammatical analysis (Mason 2007): see section 8.3.2 below.

of spoken language to construe informality and intimacy, which motivate grammatical choices.

8.2 Written Grammar, Spoken Grammar

Reliance on written data (or, more accurately, on intuitions based on written data) has led linguists to describe grammar in terms of constituent hierarchies, a description which, it is now argued, is particularly unsuited to the linear and online nature of speech (Brazil 1995) and which cannot account for all spoken features (Carter and McCarthy 2006). These grammars, ranging from eighteenth-century works to those of twentieth-century formal linguists such as Bloomfield, generative grammar analysts influenced by Chomsky, and constituent-structure grammarians such as Quirk et al (1985), focus on how words combine to form *sentences* defined as one or more independent clauses[1], and analysed into *immediate constituents* (Quirk et al 1985: 40): the verb phrase *have gone* is an immediate constituent of the clause, and the verb *go* and auxiliary *have* immediate constituents of the verb phrase. The hierarchical complexity of this can be illustrated through tree diagrams, or what Halliday calls *maximal bracketing*.[2]

Recognition that constituent analysis is suited primarily for written language emerged during the mid-twentieth century, with exploration of differences between written and spoken language such as the higher frequency of subordinate clauses in writing (Harrell 1957; Kroll 1977) and greater use of quantifiers and hedges in speech (Devito 1966).[3] By the 1980s and 1990s, linguists were more systematically linking features of spoken and written language to the *production constraints* inherent to the two modes (Chafe 1982) as well as their respective *social functions* (Halliday 1985/89; 1993; 1994).[4] These linguists question whether a model of elaborately constructed, embedded constituents is adequate to describe language produced under online pressures and designed to fulfil interpersonal rather than transactional functions. These distinctions, between writing and speech

[1] An association with written language is already suggested by the parallel, orthographic definition of a sentence as beginning with a capital letter and ending with a full-stop (Halliday 1994: 3-6).
[2] In comparison, functional analyses such as Halliday's which are concerned more with grammatical function than formal grammars thus rely less on the constituent structure associated with written language. As Halliday (1994: 20-23) explains, his use of *minimal bracketing* means that separate constituents are only labelled as such where they are seen to fulfil a distinct grammatical function (such as that of *subject*). As we shall see, Halliday's work is of further significance to analysis of casual conversation (Eggins and Slade 1997) and is explored below.
[3] For an overview of the development of this field, see Chafe and Tannen (1987) and Biber (1988).
[4] See Chapter 3: Language Variation Research and the Role of Corpora.

situations and between written and spoken language, are now widely accepted by linguists and discourse analysts, and the implication for grammarians is that grammar based solely on written data cannot be considered a grammar of the language as a whole.

The need to incorporate written and spoken varieties within one model is acknowledged in most attempts to describe spoken language (Brazil 1995: 12; Halliday 1994: xxvii; Leech 1998: 5-8) and builds on the notion, not that the two modes have their own underlying systems of grammar, but that spoken and written language are different expressions of the same system (Halliday 1994: xxvii). This is due in part to recognition that speech and writing cannot be conceptualised as a dichotomous or binary system,[1] and that separating varieties according to their mode of communication creates a non-productive and artificial distinction, which on a practical level also complicates comparison between spoken, written and mixed-mode varieties (Halliday 1994: xxvii). The question remains, however, as to whether a new model such as Brazil's (1995) is needed or whether spoken features can be incorporated into existing frameworks (Biber et al 1999; Carter and McCarthy 2006). The following overview shows that although Brazil's and Halliday's models highlight the linear nature of speech and the role of social function in shaping language, their failure to account for observed features of speech suggests that closer investigation of spoken data is needed before fully operational analytical models can be used. Descriptive, corpus-based frameworks outlined by Biber et al (1999) and Carter and McCarthy (2006) are more fully grounded in the data they aim to describe and as such account for non-clausal, 'incomplete' elements.

8.3 Models incorporating Spoken Grammar

8.3.1 Systemic Functional Approach to Casual Conversation

The significance of Halliday's systemic functional approach (SFL)[2] to analysis of spoken grammar lies in its concern with the interpersonal, alongside other language functions. This is evident firstly in its functional approach to clause analysis, by which only clause constituents fulfilling distinct grammatical functions are labelled; and, secondly, its interpretation of clauses along simultaneous strands of meaning, ideational, textual and interpersonal, by which grammatical structure is related to function and social context. In

[1] Again, see Chapter 3: Language Variation Research and the Role of Corpora.
[2] See, for example, Halliday (1994).

other words, SFL relates grammatical structure to social function by describing grammar as a resource for making interpersonal ties as well as conveying information and reflecting experience.

Eggins and Slade's (1997) analysis of casual conversation shows that this theorised link between clause functions and what participants try to achieve socially can be put to practical application. In particular, their exploration of the interpersonal strand[1] shows that linguistic features can be related to social role and conversational strategies. For example, grammatical patterns of mood choice emerge as a key resource for enacting and constructing *status differences*, one dimension of the *tenor* of a register (Eggins and Slade 1997: 52-53). In interaction between Brad and his parents, social roles are indicated by the fact that while the latter frequently use interrogatives, Brad makes statements rather than asking questions. While this suggests he does not have to compete for the floor, his mother's use of elliptical declaratives and minor clauses shows her supportive, non-initiating role.

Eggins and Slade's (1997) investigation draws on naturally-occurring data but, although SFL itself was developed through similar use of small datasets, it appears that data in this and other SFL studies is often adapted to fit the theory rather than the other way around (Thompson and Hunston 2006: 2). This is shown, for example, by Eggins and Slade's treatment of 'incomplete' clauses. Their application of the model involves the division of the transcript into clauses, the identification of clause components, and the assignment of a mood label.[2] Abandoned or incomplete clauses, however, although numbered are often (where it proved impossible to identify the mood) simply left out of the categorisation (Eggins and Slade 1997: 106-7). As Thompson and Hunston (2006: 3) put it, theory-driven descriptions may constrain observations, in that features that do not fit are often 'not accounted for at all'.

[1] Their focus on mood choices in constructing meaning can be criticised however, as Eggins and Slade point out throughout their analysis, as neglecting discoursal, semantic and generic patterns. Martin (2004: 69), for example, suggests that full exploration of the construction of meaning through grammar requires simultaneous analysis of the three strands of meaning, especially as this often involves focusing on ideational meaning with other strands relegated to semantics and pragmatics; and this can therefore be seen as limiting Eggins and Slade's analysis. It could, however, be argued that Martin's own (2000) investigation of appraisal similarly prioritises the interpersonal over other strands.

[2] The main mood features of each clause was noted on coding sheets, alongside turn number and speaker; clause number with * indicating incomplete clauses; a distinction between independent and dependent clauses with # marking dependent clauses; the Subject of the clause (in parentheses if elliptical); negation, if any; and presence of Adjuncts.

Halliday's functional analysis is in many ways useful for describing speech not least in relating grammatical features to wider social functions pursued by participants. However, as Halliday (1994: xxvii) admits, although he explores prosodic features and the clause complex, his intention is not to produce a model that adequately covers spoken features and, as Eggins and Slade's investigation shows, it cannot necessarily accommodate utterances that do not fit existing descriptions of the clause. Meanwhile, the systemic functionalist adoption of a constituent model, although richly layered and integrated, does not address the suggestion that spoken grammar be analysed linearly rather than by reference to hierarchical sentence constituents relevant to written grammar.

8.3.2 A Grammar of Speech

The principle on which Brazil (1995) constructs his model is that investigation of spoken grammar must recognise, through attested language data, the real-life, pragmatic and purposeful aspects of spoken discourse, as well as its linear nature. Observations underlying Brazil's model are not unlike those made by Halliday, in that both relate grammatical analysis to descriptions of what speakers do with language.[1] In comparison to SFL, however, although applicable in theory to certain written texts, Brazil's approach is driven specifically by consideration of speech and how participants involved in spoken interaction experience language use.[2]

Firstly, picking up (explicitly) on Halliday's distinction between product and process, Brazil realises the suggestion that by approaching grammar from the point at which language is *happening*, different analyses emerge than those produced from the traditional starting point of the completed text. His observation is that spoken communication occurs in time, and that speakers are concerned with how to manipulate linguistic resources at any particular point in 'pursuit of a purpose' (Brazil 1995: 26). If communicative purposes are the outcome or product of speech, rather than a completed text, then the linguist's aim must be to determine how these are fulfilled (Brazil 1995: 39).

The role of pragmatic and purposeful aspects of speech in shaping linearly-organised grammar is evident throughout Brazil's analysis. *Telling increments* are defined by

[1] As in speech act theory (e.g. Austin 1962).
[2] A strength of Brazil's work, for example, is the role given to intonation in marking new and given information.

constituent parts[1] forming *simple chains*, characterised as beginning from an Initial State and progressing through a number of Intermediate States to the Target State.[2] The choice between elements is determined by communicative obligation: what the listener 'needs to be told' (Brazil 1995: 52). This reflects views as to how the apparently wide selection of lexis that can be slotted into paradigmatic slots is actually constrained (syntagmatically) by surrounding choices, and the suggestion that items 'prospect' the next (Sinclair 2004: 88-90).

Simple chains can be elaborated in order to encompass complex utterances. Complex verb sequences, for example, involve the addition of non-finite elements ('*-ing* forms', '*to* forms' and 'pp forms') which, represented by V', can be used as either extensions or suspensions. The suspension occurring before the predicted element in the following example constitutes its own subchain.

Figure 8.1 Suspended sub-chains (Brazil 1995)

She	found	sitting	in her car		this little old lady
N	V	(v'	a	)	N

The need for a linear description of speech is supported by other observations of spoken grammar, including Halliday's *clause complex* and Carter and McCarthy's multiple-coordination. Despite his linear intentions, however, Brazil's model reveals the beginnings of a multi-layered model of constituents. The sequence *at the moment*, for example, can be analysed either as the functional label A (adjunct) or the formal categories PdN (preposition, determiner, noun). The temptation here is not only to move away from a purpose- to a form-driven model but to start thinking in terms of hierarchical organisation, the very feature Brazil strives to avoid. Perhaps the issue is that, although Brazil's model represents a new *perspective* on language analysis it essentially involves giving different terms to traditional categories.

Brazil's handling of multi-word units is similarly mixed. On the one hand, he recognises the way in which context resolves problems of ambiguity posed by analysts not working

[1] These include obligatory nominal (N) and verbal (V) elements, plus optional adverbial (A) or adjectival (E) elements and an additional N, as well as appropriate intonational features, namely a proclaiming tone.
[2] The target state is the modified set of circumstances that comes about as a result of the listener being told what needs to be told.

with used data. Indeterminacy is not relevant to a purpose-driven account of used, contextualised language; and nor is it likely to lead to ambiguity.[1] This reflects Sinclair's (1991: 116-121) observations regarding the role of collocation and pattern in disambiguating between senses, and his suggestion that open-choice and idiom principles can account for different interpretations of the same stretch of language (Sinclair 1991: 109-115). However, although Brazil's work is important in highlighting the way in which language is prospected linearly through previous communicative choices, his focus on grammatical structure and word-class neglects discussion of prospections made by particular lexical items.

The fundamentals of Brazil's model can however be extended to take these aspects into account. Sinclair and Mauranen's (2006) linear unit grammar (LUG), for example, avoids traditional categories and starts with the division of data into chunks, determined intuitively by native speakers and labelled as organisational (O), including interactive (OI) and text-oriented (OT) chunks, or message-oriented (M), further subdivided into categories including message fragments and supplements. The drawback to Sinclair and Mauranen's work, as pointed out by Mason (2007), is that it stops at the point at which OI chunks are discarded to leave a 'cleaned-up', analysable version of message content. However, while LUG does not as yet offer a replacement model, it highlight aspects of spoken discourse that are not amenable to analysis through traditional models and begins to suggest future developments for a truly linear account of spoken (and written) language.

Finally, although consideration of real data[2] is essential to Brazil's approach, the retellings he uses of *The Little Old Lady* monologue are not typical of interactive spoken communication between two or more parties. According to Brazil, the narrators generally made up what they said as they retold the story, and so the data is produced incrementally and linearly (Brazil 1995: 26). However, Brazil's main concern with 'the grammar of telling' rather than of asking (and therefore with extended monologues rather than exchanges characterised by shorter turns) means that, although he intends this focus as a basis on which to extend the possibilities of his model, his later discussion of asking

[1] This rests on the distinction between indeterminancy as the potential for items to be interpreted in more than one way, and ambiguity as the actual failure of a stretch of language to unequivocally yield the interpretation intended by the speaker (Brazil 1995: 71).

[2] Brazil (1995: 24) uses the term 'used language' which he defines as 'language which has occurred under circumstances in which the speaker was known to be doing something more than demonstrating the way the system works'.

exchanges must necessarily rely upon invented examples and reported question-response sequences from the narratives (Brazil 1995: 190). The failure of Brazil's model to fully account for features of speech identified elsewhere can be put down to its basis in staged, lengthy monologues. One example is ellipsis and omission, with incomplete sentences merely marked as ….. in his analysis. Situational ellipsis is not addressed, the only omission dealt with being structural. Nor does Brazil satisfactorily deal with many features identified in spoken data by Carter and McCarthy (2006) or Biber et al (1999). Discourse markers are described in a way reminiscent of traditional analyses as elements which are 'detached' from incremental chains (Brazil 1995: 219), while examples of vague language such as *kind of* are treated together with the elements they modify: *kind of deserted*, for example. This is perhaps unsurprising in that Brazil is using monologue and staged story-telling. However, it is therefore not unreasonable to suggest that the actual language Brazil describes using his model of speech is to some extent reflective of standard written language.

Another limitation is not so much the nature of the corpus but the fact that the approach is corpus-based rather than corpus-driven, in that Brazil finds examples from the data to match preconceived ideas, rather than starting from data description (Brazil 1995: 49). Brazil (1995: 51) goes on to admit that data must be carefully selected in order to fit his model and avoid complexities which his model does not account for.

8.3.3 Corpus Approach

The corpus approach adopted by Carter and McCarthy (2006) and Biber et al (1999) overcomes limitations of Halliday's and Brazil's work in so much as they account for naturally-occurring or attested features of speech. Halliday and Brazil highlight the linear nature of speech and the way in which its grammar is shaped by social and communicative functions fulfilled through speech, but their analyses also raise questions as to whether we have as yet sufficient understanding of spoken grammar to construct theoretical models. Both Halliday and Brazil, for example, neglect consideration of incomplete structures and 'omitted' elements typical of speech. In contrast, the corpus approach uses naturally-occurring data to identify or explore features of spoken grammar and, rather than attempting to develop a theoretical model of grammar (Biber et al 1999: 6-7), instead works within and around traditional structures. Rather than identifying new features, the corpus

approach foregrounds features which do not fit into existing categories, and describes them as functional and purposeful within speech rather than degenerate versions of standard written forms.[1] Although, as Leech (1998) suggests, this could represent a move towards a structural model, based on clines of features more or less restricted to speech, the fact that much of their work remains descriptive highlights continuing difficulties in explaining spoken features within an analytical framework and thus the benefits of descriptive work.

Use of corpus data raises questions as to the size and range of corpora and implications for generalising across spoken (or written) varieties. Datasets drawn on in descriptive studies of spoken grammar are, given current restrictions, impressive: the spoken component of the LSWE (Longman Spoken and Written English), comprises 3436 texts and 3,929,500 words (Biber et al 1999: 25); while CANCODE contains five-million words of spoken English (Carter and McCarthy 2006). Both cover various communicative settings categorised as 'conversational registers' or 'everyday spoken discourse'.[2] However, the criticism levelled at Brazil regarding his limited number of monologues can essentially be levelled at any corpus study: however large or diverse, spoken language corpora can never be completely representative of speech (Carter 2004b: 170-171). Of greater relevance, however, is that Carter and McCarthy's text-driven approach precludes very stringent comparison regarding the frequency with which texted and spoken items occur: few frequency figures are given for the spoken data. Given, perhaps, the limited size of CorTxt, it was possible to quantify findings to a greater extent, but comparisons between spoken and texted corpora must remain cautious.

Exploration of spoken features identified by Carter and McCarthy (2006) in a descriptive analysis of texted grammar is nonetheless beneficial for two reasons. Firstly, our understanding of texting and the extent to which texters draw on spoken language can best be explored using a descriptive, data-based model, given the limitations of theoretical models (Halliday 1994; Brazil 1995). Secondly, however, application of Carter and McCarthy's (2006) findings to a new dataset (texting) can only serve to strengthen and

[1] For example, the term *utterance*, as used by Carter and McCarthy (2006: 177), refers to 'complete communicative units, which may consist of single words, phrases, clauses and clause combinations' including those previously categorised as incomplete sentences; their term *multiple coordination* (Carter and McCarthy 2006: 170) replaces previous descriptions of speech as fragmented; while the terms *headers* and *tails* replace earlier ones of left- and right-dislocation.

[2] CANCODE, for example, is explicitly organised in terms of text types and typical situations ranging from the transactional provision of information (eg, a museum guide commentary) to an intimate collaboration (e.g. a couple decorating a room or siblings discussing their childhood) (see e.g. Carter 2004b: 149-150).

extend their dynamic, flexible framework. As an initial step in this direction, the use in texting of features identified by Carter and McCarthy is explored below.

8.4 Texted Grammar

8.4.1 Spoken grammar, texted grammar?

The following investigation of spoken grammar in texting shows that features typical of conversational speech can be identified in texting. For example, the clausal structure of texted utterances at times resembles string of units identified in speech; there is a great amount of situational ellipsis; *headers* and *tails* occur, although infrequently; there is a fair amount of deictic reference; frequent use of response tokens and vague language; and a speech-like use of discourse markers, illustrated below by *oh*. This suggests texters draw on spoken language to create speech-like informality and intimacy. Differences between spoken and texted grammar emerge, however, suggesting texters are also motivated by other purposes or affordances of the medium. These differences include clipped and disjointed stringing together of clauses; types of ellipsis not commonly found in speech; some difference in use of *this* and *that*; and less evidence of discourse markers such as *you know* and *you see*. Greetings and sign-offs also play a larger role in texting than in speech. These features are explored in greater detail below.

8.4.2 Clause combination

8.4.2.1 Overview

One hundred text messages were randomly-selected from CorTxt ordered according to the time at which text messages were sent (the ones used being sent around 14:30). The following analysis of clause combination in these text messages focuses on relationships between clauses: that is, how they combine in, for example, paratactic or hypertactic relations, to use Halliday's terms. Some points of clarification and definition are needed before the findings are outlined. Firstly, the nature of abbreviation within the clause is discussed in the following section on situational ellipsis and is not here a consideration: so that the degree to which, for example, units are deemed to be combined in a speech-like way does not depend on whether elements omitted from constituent clauses are those typically omitted in spoken language. The term *unit* is used throughout to refer to stretches of text containing at least one clause and ending with a full-stop (corresponding to a traditional *sentence*) but, due to the technology, not necessarily beginning with a capital

letter. Use of *unit* rather than, say, Halliday's term *clause complex,* is motivated by the desire to remain pre-theoretical at this stage and because clause complex is not used by Carter and McCarthy. The term *clause* is itself used loosely to include elements (such as noun phrases) functioning within a unit in the way that clauses traditionally do and which, although not traditionally labelled a clause in the absence of predicating elements, could be *expanded* to constitute a clause in the traditional sense.

Investigation of the 100 text messages revealed 19 that comprise single one-clause units in which clause combination does not occur, most of which are either exclamations (9) or questions (4).[1] Three examples of these text messages in their entirety are as follows (Appendix 8.1).

(8.1) Have you seen who's back at Holby?!

(8.2) I have lost 10 kilos as of today!

(8.3) Now thats going to ruin your thesis!

Clause combination in the remaining 81 text messages can be described as more or less *speech-like*, depending on the number of clauses within each unit, and how clauses and units are combined.

8.4.2.2 Speech-like clause combination

Identification of *speech-like* clausal combination in texting is based partly on intuition and partly on descriptions of spoken clauses by Carter and McCarthy (2006), as well as Biber et al (1999) and Halliday (1994). As mentioned previously, Carter and McCarthy use *multiple coordination* to describe how spoken clauses typically combine, and their description matches that of Halliday's (or Brazil's linear approach): clauses in speech are strung together using connectives, chiefly coordinators (*and*, *but*) and subordinating conjunctions such as *because* and *so* functioning to an extent like coordinators. Carter and McCarthy (2006: 556) also point out that multiple coordination can occur between 'contact clauses'

[1] Defined as such through the respective use of exclamation and question marks.

(coordinated clauses linked not by conjunctions but by commas).[1] In other words, *multiple coordination* embraces clauses linked by punctuation and certain subordinators, as well as coordinators.

Speech-like units in texting comprise just over 20 of the 100 text messages analysed, in that clauses are combined using multiple coordination, including the following examples (Appendix 8.1). Clauses are divided by |, and units with ||.[2]

(8.4) if you see NAME10 | then dont mention | that you've spoke to me, | she wants me | to go to the pub | **and** | i cant be bothered | **and** | my mom goes on holiday tomorrow.

(8.5) I'm in solihull, | do you want anything?

Speech-like units in texting also include those involving structures identified by Carter and McCarthy as typical of speech such as mid-position tag-questions and wh-fronting:

(8.6) that would be good … I'll phone you tomo lunchtime, shall I, to organise something?

(8.7) What i mean was i left too early to check, cos i'm working a 9-6.

As these show, speech-like text messages are characterised by relatively lengthy units, with 8,[3] 7 and 6[4] clauses respectively in the following units.

Figure 8.2 Speech-like text messages

No	Clauses
	Hiya –
1	*thinking about weekend*
2	*and realised*

[1] This occurs not only in literary style or for effect in narratives but, according to Danielewicz and Chafe (1985), in speech (if a comma is interpreted as reflecting, in writing, the role carried out by a rising tone in speech).

[2] *and* (as a coordinating conjunction) in these examples is presented as being outside the clause, whilst subordinators are represented as being part of the following clause.

[3] That this is one clause is based on the fact that the bracketed clause is subordinated to the previous clauses and so the brackets are interpreted functionally - that is, as an aside - rather than indicating the end of the clause. A similar interpretation is thus given to the question-tag, another example of a speech-like mid-position tag-question.

[4] That this is one clause assumes that the hypens can be treated, like commas, as separating contact clauses rather than utterances.

3	*I was assuming*	U
4	*you would stay Saturday night too –*	N
5	*you will ,*	I
6	*won't you?*	T
7	*(tho no prob*	
8	*if you can't)*	

No	Clauses	
	Dude,	
1	*if you see NAME10*	
2	*then dont mention*	U
3	*that you've spoke to me,*	N
4	*she wants me*	I
5	*to go to the pub*	T
6	*and i cant be bothered*	
7	*and my mom goes on holiday tomorrow.*	

No	Clauses	
	Hi-	
1	*seems*	
2	*a small group will be going out on fri-*	U
3	*i'll book a table for a meal at 7.30-*	N
4	*so try to get the earlier train*	I
5	*and pack light*	T
6	*if possible*	
	x	

At the same time, despite having several clauses, the last two text messages above (and arguably the first) comprise only one unit. Most speech-like text messages have only one or two units, with just one exception having three.

Figure 8.3 Three-unit text message

1	*the more holiday photos, anecdotes etc the better, naturally!*
2	*anyway, i'm slaving away, as per so better stop.*
3	*see ya soon tagliatelli!*
	SD

To sum up, prototypical speech-like text messages comprises a limited number (one, two or three) of two- to eight-clause units, joined by parataxis, either coordinators or commas, and/or hypotaxis which chiefly involves *that* and *to*. The last highlight the frequency of reporting and mental process verbs in texting, such as *mention that*, *seems (that)*, and *realised (that), wants (me) to* above.

8.4.2.3 Text-specific clause combination

Several text messages, however, depart significantly from this speech-like prototype. Thirty-eight text messages are here described as *text-specific*, in that they appear to be characteristic of texting rather than spoken language (Appendix 8.1). These can comprise several units, as in the following.[1]

(8.8) Good news indeed. || But Prodi doesnt seem that good too. || In England,its pretty bad too. || Blair is a complete idiot. || Im gonna b more and more involved in political issues. || Im already a unionist, | but | i think | next year i'll b a member in a political party. || Its called Ligue Communiste Revolutionnaire. || I'll tell u more about it |when we'll c each other. Xx

(8.9) Hi. || Do u want | to join me with sts later? || Meeting them at five. || Call u after class.

(8.10) No i'm not gonna be able to. || too late notice. || i'll be home in a few weeks anyway. || what are the plans

As the above examples suggest, some text-like text messages comprise only single-clause units:

(8.11) That's a shame! || Maybe cld meet for few hrs tomo?

(8.12) Sorry. || mail? || POSTALADDRESS

(8.13) Not thought bout it... || Drink in tap & spile at seven. || Is that pub on gas st off broad st by canal. || Ok?

Those with two- or three-clause units tend to show hypotaxis rather than parataxis, and chiefly *that*- or *to*-clauses:

(8.14) Hope | [that] you had a good time last night .\ || Bet | [that] NAME172 is ill. || Why didn't NAME262 go? || Just a text | to ask you | who did your bathroom | and | how much did it cost?

[1] Again, utterances are marked by ||, and clauses by |.

Again, the *that*- and *to*-clauses are concerned with mental processes or reporting: *hope that, bet (that), a text to tell you*.

Each unit comprises fewer clauses than in speech-like text messages (a maximum of 4, rather than 8), but there are often more units per text message than in speech-like text messages (up to 6, rather than 3).

Figure 8.4 Non-speech-like text messages

No	Units	
	lhey,	M
1	*did you have a good time in porague?*	E
2	*I'm in ireland at the mo with the in laws!*	S
3	*Going ok .*	S
	yes	A
4	*love being on hol 2!*	G
		E

	God	
1	*it's really getting me down just hanging around.*	M
2	*I hate not knowing.*	E
3	*What did NAME72 say?*	S
4	*How would she know?*	S
5	*They'd be crazy not to take u.*	A
6	*I'm worried*	G
	X	E

1	*i'm not sure how to break this to you.*	
2	*there's no easy way to put it....*	M
3	*i can't make the friday-night fun.*	E
	sorry.	S
4	*however, feel free to text me during the evening if there are any lulls in conversation.*	S
5	*anyway, hope ur exotic trip goes well.*	A
6	*see u next term.*	G
	rw	E

Noticeable here is the lack of clear relationships between units, with the result that text messages appears fragmentary or disjointed: *I hate not knowing. What did NAME72 say? How would she know? They'd be crazy not to take u. I'm worried*.

Text-like clause combination, then, is characterised by text messages consisting of any number of one-clause units or two- to three-clause units, combined by hypotaxis (mainly *that*- and *to*-clauses). There are several units within each text message, but fewer clauses within each unit, little parataxis between clauses and a somewhat disjointed connection between units.

Other text messages (around 21 of 100) combine both speech-like and text-specific clause relationships (Appendix 8.1). These suggest the need for a cline from more speech-like to more text-specific text messages, or that clausal analysis should assume prototypical speech-like and text-specific text messages and the extent to which text messages meet the criteria of either analysed. In the text message below, for example, the first unit is speech-like in that it comprises four clauses joined paratactically, in this case with *and* or no connectors,[1] while the two units which follow are text-like in each comprising one clause.

(8.15) NAME78, NAME416, NAME366 and NAME417 picking them up from various points | going 2 yeovil | and they will do the motor project 4 3 hours | and then u take them home. || 12 2 5.30 max. || Very easy x

Classification of text messages as more or less speech-like or text-like could be explored across more text messages. Nevertheless, the identification of two prototypical types of clause combination suggests that texters not only construe speech-like language characterised by multiple coordination but also what could be described as text-specific language comprising strings of single-clause units. Across these text messages is a high number of *that*- and *to*-clauses, reflecting the use of mental process verbs and reporting verbs in CorTxt. Evidence of embedded clauses typical of formal written language is, however, missing across this cline of clausal combination. These findings regarding clause combination add another layer to the description of lexical phrases outlined in Chapter 7.

8.4.3 Situational ellipsis

8.4.3.1 Overview

Situational ellipsis, defined by Carter and McCarthy (2006: 181) as 'not explicitly referring to people and things which are in the immediate situation',[2] occurs more frequently in spoken contextualised interaction than in written texts where shared backgrounds cannot be assumed. Its occurrence in texting draws attention to the immediacy of texted interaction and texters' shared backgrounds, as well as their attempts to recreate the spoken informality of speech. However, the elision in texting of some elements and not of others suggests again that texters both construe *speech-like* text messages (resembling spoken language)

[1] As well as a header – see section 8.4.5.

[2] While this may not fully explain why some elements are typically omitted in speech and others are not (discussed later), their definition serves to distinguish situational ellipsis from textual and structural ellipsis (Carter and McCarthy 2006: 181).

and, elsewhere, text messages best described as *text-specific* (characteristic of texting rather than speech). That is, although much ellipsis in texting follows speech-like patterns of omission, medial ellipsis and use of copulars or auxiliaries without the subject pronoun does not. One reason for this is the need, given technological constraints, to shorten text messages. However, the occurrence of both speech-like and text-specific ellipsis highlights the choices open to texters and their active role in construing texted grammar. One caveat concerns the retrieval method: given that investigating ellipsis involves studying what is not realised, occurrences were found by searching for the realised elements (*am*, for example) and determining through sorted concordance lines which were not preceded by expected elements (in this case, *I*).

8.4.3.2 Speech-like ellipsis

Examples of elements often omitted in both texting and spoken language include the initial elements of phrases and declarative clauses: subjects pronouns, auxiliary verbs, and determiners. Examples in texting, where the omitted element is represented by *, are presented below.

Figure 8.5 Speech-like omission of elements in texting

** Dunno, * haven't spoke to NAME10 about it yet.*	(subject pronoun *I* omitted twice)
*Hi- * seems a small group will be going out on fri*	(subject pronoun *it* omitted)
** Just checked.*	(subject and auxiliary verb omitted)
*yes... * staff-room can be pretty unbearable sometimes!*	(determiner *the* omitted)[1]

Elision of subject pronouns in texting often occurs with mental process verbs such as *like*, *hope, think, guess* (in part because of the frequency of these verbs in CorTxt) as it does in speech. Carter and McCarthy (2006) give the spoken examples: *Like your new car*; *Wonder how they got up walls*; and in short replies: *Yeah*. *Think so*. Examples in text messaging include the following.

(8.16) * Thought I didn't see you.

(8.17) Easy mate, * guess the quick drink was bit ambitious.

[1] See Chapter 6 (section 6.4.2.3) of this thesis for discussion on the use of reduced noun phrases (where the determiner is dropped) as a cohesive device across messages.

As in spoken language, subject pronouns in statements with question tags are also omitted in texting, as in:

(8.18) Annoying isn't it.

Turning to interrogative forms in texting, we see that, again as in speech, auxiliary verbs and subjects are frequently omitted from questions. Auxiliary *are* is missing for example in the following questions taken from CorTxt.

(8.19) What * u wearing?

(8.20) * You gonna ring this weekend or wot?

Another speech-like example of elision of auxiliaries in texting is the ellipsis of *are* from the phrase *what you up to?*, which occurs 38 times in CorTxt if alternative spelling of *what* (*wot, wat*) *you* (*u*) and *to* (*2*) are included. Examples in CorTxt include the following.

Figure 8.6 Selected concordance lines for *what you up to* and spelling variants

```
1        They're right little buggers! What * you up to today? Mwah! Xxx"  "
2        pictures of her sister's cat. What * you up to tonight?"  "She mentio
3      al National Hotel. Pls call."  "Wot * u up 2?tb"  "HI NAME252 HOPE
4        Haven't seen you for a while. What * you up to. Are you cushiing tom
5      Speak soon xx"  "Hello chicken, wot * u up to? Boy have i had a stressful
6        for a spin! Shall i get u?"  "What * u up to? Fancy popping into to
7      s the best policy with NAME159! Wot * u up to? X"  "I do apologise for tha
8      roo then?"  "Hey there. Howz u? Wot * u up 2? I'm just watchin tv. Went
9       l Thursday. Hope you are well. What * u up to? X"  "Oh poop about sit j
10      eat time but am shattered now. What * you up to?"  "Cool, i'll do that an
11      to local tonight with NAME219. Wot * u up to? Everything ok? X"  "In th
12      ely today... You all right and what * you up to this weekend?"  "Still a
13      y to hospital."  "Watcha dude. What * you up to tomorrow?"  "Predictio
14     in 2 weeks"  "YO YO YO MONKSTER WOT * U UP 2 FUNGUS? THIS IS NAM
15       will wait til tomoro"  "Hello what * you up to today? Mum and i have
```

The implication of speech-like ellipsis in texting is that much compression of Txt can be seen as part of texters' attempt to construe or perform speech-like informality or intimacy. However, as explored below, medial ellipsis and use of copular *be* or auxiliary without its

subject are also characteristic of texting but neither intuitively speech-like nor noted by Carter and McCarthy in their description of spoken grammar.

8.4.3.3 'Dont worry is easy': subject ellipsis

As Carter and McCarthy (2006: 184) point out, pronouns tend not to be ellipted in certain spoken contexts, as pronoun elision is limited mainly to auxiliary and mental process verbs. Pronouns tend not to be ellipted in speech either with copular *be* or *be* as an auxiliary, as shown by Carter and McCarthy's examples.

Figure 8.7 Pronouns not ellipted in speech (Carter and McCarthy 2006: 184)

A: Will your mother be here too?
B: Yeah, she's coming
(Not: Yeah, is coming)

Non-speech-like omission of the subject with the retention of copular or auxiliary *be* occurs across CorTxt. The third-person form *is*, for example, occurs 33 times (from a total 1576 times) without a subject pronoun. In 18, the missing pronoun is *it*, represented by * in the following.

(8.21) tap & spile at seven. * Is that pub on gas st off broad st by canal. Ok?

(8.22) Don't worry, * is easy once have ingredients!

(8.23) Perhaps * is much easy give your account identification, so i will tomorrow at UNI

(8.24) Hey NAME42, is * rite u put »10 evey mnth is that all?

As these examples show, omission occurs both where *it* functions as a general referent as in * *is easy once have ingredients*, or as anticipatory *it*: *Perhaps * is much easy give your account identification*; and where the reference is anaphoric: *tap and spile at seven. * Is that pub on gas st*. Interestingly, however, there are fewer or no occurrences of *is* (without the subject) as part of present progressive verb phrases.

Other subjects are similarly omitted where *is* is retained. *There* is omitted from the following.

(8.25) Does cinema plus drink appeal tomo? * Is a fr thriller by director i like on at mac at 8.30.

The subject *he* or *she* is missing from 13 examples, including the following.

(8.26) NAME232 came to look at the flat, seems ok, in his 50s? * Is away alot wiv work. Got woman coming at 6.30 too.

This kind of elision, however, occurs more frequently with first-person *I* than other personal pronouns, mirroring the fact that *I* is more frequent in CorTxt than all pronouns except *you*. *Am* occurs without *I* in 334 of 629 occurrences, both as copula and auxiliary, as in the following.

(8.27) * Am on a train back from northampton so i'm afraid not![1]

(8.28) * Am on my way

(8.29) No * am working on the ringing u thing but have whole houseful of screaming brats so * am pulling my hair out! Loving u

Omission of *I* occurs also with the past-simple tense of *be* (*was* and, in two cases, *were*)[2] in 121 of 946 total occurrences of *was*, as in the following examples.

(8.30) Sorry * was at the grocers.

(8.31) Pity, * was in mood for that. So...any other suggestions?

[1] Note, in this example, how the text-specific ellipsis *Am on a train* contrasts with the speech-like *I'm afraid not* towards the end of the text message.

[2] It is possible that *I* (or *we*) is omitted in the following example but it is perhaps more likely that the apostrophe from *we're* is instead missing:

*I don't mind, probs not a great idea if * were having a meal plus will have to carry it around after.*

(8.32) * Was thinking about chuckin ur red green n black trainners 2 save carryin them bac on train

There are also 27 examples of *was* where *it* has been omitted but again these are less frequent than those without *I*:

(8.33) * Was really good to see you the other day dudette, been missing you!

(8.34) * Was a nice day and, impressively, i was sensible, went home early and now feel fine. Or am i just boring?! When's yours, i can't remember.

He is not omitted from any example and *she* only twice, which may in part be because third parties are less frequently referred to in CorTxt; and reflecting the tendency, in speech, not to omit third-person pronouns.

Omission of pronouns before other auxiliary and delexicalised verbs seems equally non-speech-like. Examples where *will* is retained but the subject pronoun omitted are frequent (277 out of a total of 389 occurrences of *will*) and include the speech-like, relatively fixed expression *will do*, which occurs 14 times.

Figure 8.8 Concordances for *will do*

1	t planning to go out but	**will do**	if people want to.
2	r of two will do) xx Ok	**will do**	- downstairs now
3	w very hazy … Yes ok	**will do**	in the morning
4	of your arse again! Aye-	**will do**	my lover! We ju
5	chance to eat cake yet.	**will do**	soon though.than
6	nd? U got a spare fag?	**Will do**	! We are doin fin
7	et iin touch tomorrow x	**Will do,**	if u fancy comi
8	Where shall we meet?	**Will do**	- yes- must meet
9	soon. Xx Okey dokey,	**will do.**	Have a good da
10	en had enough of gep?	**Will do.**	Enjoy ur day. E
11	t an hour and a half xx	**Will do.**	Yes but can w
12	nd happy christmas xx	**Will do.**	Hello! Been stu
13	E46's new girlfriend! ok,	**will do.**	rx Hello my sw
14	v far! See u nxt wk. Ok,	**will do.**	Not sure bout w

However, other uses of *will* without a pronoun appear less intuively speech-like. In the following, *I'll* and *it'll* would appear more speech-like alternatives:

(8.35) Sorry completely forgot * will pop em round this week if your still here? X

(8.36) * Will have two more cartons off u and NAME262 is very pleased with shelves

(8.37) * Will be september by then!

Have, both as auxiliary and delexicalised verb, also occurs frequently without subject pronoun (177 out of 1870 occurrences), as in the following.

Figure 8.9 Selected concordance lines for *have* with no subject pronoun

1	xx" "Just saying hello!	**Have**	of course lost track
2	good, its still cold here.	**Have**	finished work, pootl
3	nfused by your text but	**have**	processed for 24 h
4	just concreting garage	**have**	to cover front and
5	Yey! Would join you but	**have**	just this moment g
6	go to woodlane first as	**have**	had phonecall 4 tic
7	.love,NAME621" "Hello-	**have**	received your pack
8	e been out working and	**have**	just got in. Cya."
9	s x" "Yeah that suits -	**have**	to find out bout cla
10	ten to your adventure." "Have		called home a co
11	s mind of its own roads	**have**	been a mare all da
12	ad u had a good one x" "Have		had a gorgeous d
13	lecture on Freud later.	**Have**	texted from NAME
14	you make more sense.	**Have**	sent number to NA
15	then is it xxx" "O dear!	**Have**	just finished wot ti
16	ill at work?" "sorry, no,	**have**	got few things to d
17	y text to say i'm back -	**have**	been for a while, d
18	got hectic on friday and	**have**	only just now settle
19	aking pizzas..x" "Wow	**have**	just woken up.. Ap
20	bringing tickets to u but	**have**	to go to woodlane

8.4.3.4 'Wine good idea': medial ellipsis

Medial ellipsis, which also occurs in a non-speech-like way across CorTxt, describes the omission of copular *be* in declaratives where the subject (in the cases below, realised by nouns rather than pronouns) is retained. Examples in texting include the following.[1]

(8.38) Evening * v good if somewhat event laden. Will fill you in, don't you worry … Head * ok but throat * wrecked. See you at six then!

(8.39) Hi. Hope ur day * good! Back from walk, table booked for half eight. Let me know when ur coming over.

(8.40) R u sure they'll understand that! Wine * good idea just had a slurp!

These examples are not typical of speech. In speech, although the subject and the copula may be omitted, as in *Ready!* (or, in texting, *Not long to wait now!*), it is unusual for the subject to be present in speech without the auxiliary, particularly where the subject is a noun rather than a pronoun. The nearest speech-like form, according to Carter and McCarthy (2006: 182), is the omission of auxiliaries in interrogatives such as in *The dog bothering you? Shall I throw him out?*. If neither medial ellipsis nor use of copular *be* without its subject is typical of speech-like language, this suggests that texters are not consistently engaged in construing speech-like informality; and that elision must be explained with reference to other factors. The tendency to abbreviate in text messages for reasons of time, effort and space is an obvious possibility, and widely documented (Hard af Segersteg 2002). For example, the preference in speech over *Am on my way* is presumably *I'm on my way*, but we can conjecture that it is not a preferred option for texters due to the increased effort in accessing the apostrophe. Physical constraints of the medium cannot however constitute the full picture, as they fail to explain why some elements are omitted and others not. To take the last instance above, in *Wine good idea just had a slurp!* the auxiliary is omitted, as is the article before *good idea* and the subject in the second clause (presumably *I*), but the article in *a slurp* retained.

Another explanatory factor is that elision is guided by consideration of which elements are necessary for communication in any given situation, as Carter and McCarthy (2006: 181)

[1] The omitted auxiliary is marked by *.

suggest with reference to speech. Carter and McCarthy's description of written language as a necessarily elaborated version of a streamlined spoken variety is an attractive one, as it reverses common notions that spoken ellipsis is simply an abbreviated version of a fuller and richer written structure. However, Carter and McCarthy's argument (throughout their section on situational ellipsis) that what is omitted is 'unnecessary' or surplus to what is needed to fulfil 'the purposes of communication' also fails to fully explain why some elements are omitted and others typically not. For example, as Carter and McCarthy (2006:182) point out, auxiliary verbs can be omitted in spoken interrogatives: *You been eating the biscuits again?* but not with *I*: *Am I interfering?* (not *I interfering?*), yet it could hardly be argued that *am* is necessary for communication in a way that *have* in the first example is not.

Similarly, Carter and McCarthy continue, pronouns tend only to be omitted in declaratives when they occur with auxiliary and modal verbs, or verbs such as *hope* or *wonder*, and not in such cases as the one we saw earlier.

Figure 8.10 Pronouns used in speech (Carter and McCarthy 2006:184)

A: Will your mother be here too?
B: Yeah, she is coming.
(not: Yeah, is coming.)

Nevertheless, as *she* in the above example is easily retrievable from the previous turn, the fact that it is omitted in texting but not in speech challenges the idea that elision is solely determined through the necessity of elements for understanding the overall message. In other words, the construal of distinctly *text-specific* language is only in part driven by consideration of what is necessary for communication, as it is by technological limitations. An alternative argument is therefore that texters *perform* brevity, as much as they perform speech-like informality and deviance, informed by the discourses surrounding and thus shaping Txt (see Chapter 9).

To sum up, while much ellipsis characteristic of texting construes an informal speech-like language, elsewhere we see the construal of non-speech-like language through medial ellipsis and use of copular *be* without its subject. Although presumably motivated by the perceived need to abbreviate and difficulties in accessing the apostrophe to form

contractions, this does not fully address the question of why some elements are omitted and others not, and instead the argument can be proposed that texters perform brevity, just as they perform spoken informality and deviance. As this suggests, the application of this grammatical framework allows for a perception of the patterns characterising Txt that differs from those highlighted in Chapter 7. The observation that texters construe speech-like and non-speech-like grammar is illustrated further in the following sections.

8.4.4 Deixis

8.4.4.1 Overview

Deictic reference includes words and phrases used as 'speakers orient themselves and their listeners in terms of person, time and space in relation to the immediate situation': refer, that is, to the person they are talking to, when they are talking, and where (Carter and McCarthy 2006: 178). Deixis is therefore, according to Carter and McCarthy (2006: 178), particularly likely when people are doing things together and as such is more frequent in spoken than in written language. Deictic reference is realised through three main word-groups: adverbs (*here*, *now*, *then*), pronouns (*I*, *we*, *him*) and determiners (*this*, *that*). Investigation of CorTxt reveals similar deictic uses made of these words, as illustrated in the following text message.

(8.41) Man this bus is so so so slow. I think you're gonna get there before me

The following overview of deictic reference in texting focuses, as does Carter and McCarthy's investigation of spoken deixis, on the determiners *this* and *that* which occur, respectively, 427 and 229 times throughout the corpus,[1] and which were investigated through concordances. The investigation shows that determiners are used in texting to refer to a shared time framework as well as to the 'virtual space' created through the medium and, as with spoken deixis, they play an evaluative role in heightening intimacy and expressing disapproval.

[1] Although, as discussed in Chapter 7, the pronouns *you* and *I* are highly frequent in texting and would therefore be worthy of study.

8.4.4.2 Reference to time

Explicit reference to periods or points of time is made by over three-quarters (315) of the 427 occurrences of *this*, including the following.

Figure 8.11 Frequent uses of *this* in CorTxt

this weekend (62)
this week (48)
this morning (46)
this evening (28) and *this eve* (6)
this afternoon (21)
this time (15)
this year (12)
this Saturday (4)
this Christmas (3).

Selected concordances lines showing *this* + time reference, shown below, indicate that *this* can refer to present, past or future time.

Figure 8.12 Selected concordances for *this* + time reference

1	cked it beast good luck	**this week**	x" "Done it but
2	"Hi, how was teaching etc	**this wk**?	Here all ok but
3	have to do things here	**this afternoon**	so can't meet, but will
4	thinking- if you're free	**this weekend**	come up - busy fri but t
5	"Take it ur not coming	**this weekend**?	Shame. I'm bloody bo
6	sure but it must b around	**this time**	roughly. Anyway happy
7	of time, you still have	**this afternoon**.	Has the postman bee
8	e you heard from NAME159	**this week**?"	"Guess i have been dro
9	to you some time over	**this weekend**?	Xx" "NAME67 should
10	pending it on booze. from	**this day**	on i am be sensible with my
11	vin had my bank statement	**this morning**,	I recon, i'll be working
12	remember if ur around	**this wkend**?"	"Yeah, it's gonna be
13	ghty, what's the excuse	**this time**?"	"Yeah, not too bad, i
14	quick drink - sometime	**this week**?"	"Hello - i'm back - glori
15	so bad. Going to try again	**this afternoon**	x" "Hiya. I'm at a
16	ike to come for a meal	**this evening**?	Let me know asap as i'
17	whether to play football	**this eve**.	Feeling lazy though." "Argh
18	ext from me - a favour	**this time**	- if you are in gep could
19	good day and good luck 4	**this week**!	Love Sx" "Evening! Are
20	"Oh gosh what did he do	**this time**?"	"Wah don't say anything

Reference to a present, ongoing period is the least frequent reference of the three categories, accounting for around a fifth of occurrences in a random sample of 100 references to time:

(8.42) Rightio. 11.48 it is then. Well arent we all up bright and early this morning.

(8.43) Bognor it is! Should be splendid at this time of year.

(8.44) I take it we didn't have the phone callon Friday. Can we assume we won't have it this year now?

Past-time reference is to the recent past, or to events that happened at a specific earlier point in an ongoing period (such as earlier this week), as illustrated in the following examples. These recent-past references makes up around a quarter of all time references:

(8.45) Have you heard from NAME159 this week?

(8.46) Will do. Was exhausted on train this morning. Too much wine and pie. You sleep well too x

(8.47) … we r stayin here an extra week, back next wed. How did we do in the rugby this weekend? Hi to NAME281 and NAME159 and NAME79, c u soon xx"

Most time references are to the near future (just over half of all occurrences). Again, these also include specific future times within an ongoing period, often in reference to plans or arrangements.

(8.48) Bbq this sat at mine from 6ish. Ur welcome 2 come

(8.49) This weekend is fine (an excuse not to do too much decorating)

(8.50) I'm really sorry i won't b able 2 do this friday.hope u can find an alternative.hope yr term's going ok:-)

That, in contrast to *this*, is not used predominantly with explicit references to time. However, there are slightly more references to time (17 out of 150)[1] than to places (11 out of 150). These references are not to past events (as you might expect with *that*) but often to agreed future plans or arrangements: *that weekend*, *that Sat*, and *that day*.

(8.51) Yeah confirmed for you staying at NAME150's that weekend

(8.52) Unfortunately i've just found out that we have to pick my sister up from the airport that evening so don't think i'll be going out at all. We should try to go out one of th

(8.53) NAME87's gonna let me know cos comes bak from holiday that day. NAME125 is coming. Don't4get2text me NAME498's number. Xx

that time can also refer to present or general points in time. It appears as part of appeals to a shared understanding of its significance:

(8.54) Cheers for the card ... Is it that time of year already?

(8.55) Our ride equally uneventful - not too many of those pesky cyclists around at that time of night ;).

(8.56) Need a coffee run tomo?Can't believe it's that time of week already

These uses of *this* and *that* in referring to shared, present or future, time can be related both to the near-synchronous nature of texting and to its interpersonal functions of arranging and coordinating future plans.

8.4.4.3 Reference to place

In contrast to frequent reference to periods or points of time, only 21 (of 427) occurrences of *this* refer to places. These include 4 references to transport:

[1] Given the varied references of that, and the time and space constraints of this thesis, only 150 were categorised. The method followed was that adopted by Sinclair (1991: 84), in that sets of 30 concordance lines were categorised one by one.

(8.57) twenty past five he said will this train have been to durham already or not coz i am in a reserved seat

(8.58) Lovely smell on this bus and it ain't tobacco... Xxx

Watch this space occurs twice and is included here as referring to the 'virtual space' created by the communication, highlighting the fact that texters are unlikely to share physical space. In the text messages above, for example, the transport referred to is that which the texter, and not the textee, is on. Other references included in the place category, such as *world* or *country*, do not necessarily indicate close proximity:

(8.59) It's reassuring, in this crazy world.

(8.60) Can not use foreign stamps in this country.

The fact that texters are not in the same place, and/or not always sure where their interlocutors are, is further illustrated below.

(8.61) Dear relieved of westonzoyland, all going to plan this end too!

(8.62) R u in this continent?

(8.63) What is this 'hex' place you talk of? Explain!

References to place with *that* (11) mainly comprise directions to or suggestions about places of entertainment with the implication that texters have been there and/or will go there at some future point.

Figure 8.13 Selected references to place with *that*

```
1       st bfore 9. No idea wher that pub is."  "We're me
2         tap&spile at seven. Is that pub on gas st off br
3          around there. There's that place we walked pa
4       to come ur way first. Or that pizza place in selly
5              "Can u meet me at that corner shop at half
```

The lack of reference to shared space in texting presumably indicates a potential difference between texted interaction in CorTxt and typical face-to-face spoken interaction represented by examples such as *I'd like to pop in to that little shop over there before we leave* (Carter and McCarthy 2006: 179).[1]

8.4.4.4 Reference to the medium and the interaction

Another 27 uses of *this* refer to the medium itself, either to *this text* (8) or *this message* (3), *this phone* (4) or *mobile* (1) or *this number* (5). The reference is to texts sent or received by the texter or to the number they contact interlocutors on.

(8.64) Hi hope u get this txt~journey hasnt been gd,now about 50 mins late I think.

(8.65) I wonder if you'll get this text?

(8.66) Only just got this message, not ignoring you. Yes, i was. Shopping that is

Most references to *this message* or *this text* refer to the one being written, with the exception of example 8.66, in which *Only just got this message* refers to the one received by the texter. Similarly *this number* and *this phone* can refer either to the texter's or their interlocutor's phone:

(8.67) Realy sorry-i don't recognise this number and am now confused :) who r u please?! NAME12 x

(8.68) Hi its NAME325 in durban are you still on this number

(8.69) Wrong phone! This phone! I answer this one but assume the other is people i don't well

This one (in bold below) refers to number or texts, only where it refers back to previous more explicit references (in bold below).

[1] Although Carter and McCarthy do not make explicit the frequency of deictic reference to *place* in the spoken corpus.

(8.70) I av **a new number,PHONENUMBER**. Wil u only use this one,ta.

(8.71) Fighting fit now but, no, tend not to leave house once have slippers on and especially not in the rain. Have friend down for weekend and busy tomo but may be in moseley sunday eve and if so will give you **a text**. Gosh this one is more a novel than a text.

Sixteen of 150 occurrences of *that* refer to the medium, either *that message*, *that text* or, less frequently, *that call*. The difference between *that* and *this* in referring to the medium is that *that* refers not to the current text message but to previous text messages, either sent or received by the texter:

(8.72) Mm not entirely sure i understood that text but hey. Ho. Which weekend?

(8.73) did u get that message

(8.74) I didn't get the second half of that message

References to *that call* refer to interaction between the texters beyond the texted exchange:

(8.75) 1.20 that call cost. Which i guess isnt bad. Miss ya, need ya, want ya, love ya

(8.76) Hello- thanx for taking that call. I got a job! Starts on monday!

Included in this category is one ironic reference to (idiosyncracies or failings of) predictive texting devices:

(8.77) Nimbomsons. Yep phone knows that one. Obviously, cos thats a real word

The above examples all show texters referring to the medium as they organise their interaction (texted and otherwise) through texting.

Five occurrences of *this* are used metalinguistically to organise the text, as part of the expressions *put it this way*, *at this rate* and *stick my neck out on this one*.

Figure 8.14 Metalinguistic uses of *this*

1	ou soon." "Lets put it **this** way, you will like it.
2	e there nearer 12.30 at **this** rate: -(" "Great! Thi
3	down! B home by 11 at **this** rate still drunk love
4	ick me up from bw" "At **this** rate it could be late
5	a stick my neck out on **this** one and say yes! I'l

One final occurrence is a texter's self-reference:

(8.78) Hi, been v busy with work, glad wk nearly over, v v tired! Ur wk ok? Re tomo, i finish work at twelve so i'll come back here n join u later, if that's ok. Was wondering whether to train in to uni station n train n taxi back as solves car problem - plus **this oldy** might flake early as been sleeping badly this wk. Let me know time if plan changes. See u tomo x

Another 24 occurrences of *that* play an organising role in the discourse by referring back to the interaction.

Figure 8.15 Selected concordances showing anaphoric uses of determiner *that*

1	hat happenend? You kept **that one** quiet! Did he turn
2	b arranged" "Not planned **that way** but it may turn ou
3	i aint holding my breth 4 **that 1**, lol" "I havent spoke
4	? X" "I do apologise for **that previous outburst**! X" "
5	g before she goes back. **That way** transport is less
6	ave fun! Try and think of **that word**, it's driving me nu
7	t get the funding! Forgot **that bit**. best to keep it in t
8	arty, how could I resist. **That kinda thing**. Texting yo
9	i think so! Did you like **that play on words** then! I k
10	of year?" "Oops. 4 got **that bit**. POSTCODE1" "P

8.4.4.5 Evaluative and referential role

Another frequent use of *this* (51 occurrences) and *that* (85 occurrences) is in reference to texters' shared experience, knowledge or background. In many cases, the determiner takes on an evaluative as well as referential role in indicating familiarity or even positive or negative attitudes. Similarly, in spoken language, Carter and McCarthy (2006: 180) identify uses of *this* and *that* referring not to physical but to psychological proximity or distance. *This*, they suggest, in *Now tell me what you think of this new girlfriend he's got?* indicates the importance, immediacy and familiarity of referents, as well as the speaker's positive

attitude. In contrast, they argue that *that girlfriend* would imply a more detached, critical and even disapproving tone.

In texting, concordances for *this* in referring to shared experience or knowledge include the following.

Figure 8.16 Selected concordances showing references to shared experience/knowledge with *this*

1	bored tho.so tell me about	**this** comedy night, i dont know anyth
2	was invited to lunch and had	**this** fish soup, a traditional korean ha
3	to type up something for	**this** outline. Getting a good list of
4	scraped the barrel with	**this** shower of social misfits. Love da
5	the bank" "So shall i sell	**this** currency then? X" "good luck...
6	think i have decided to keep	**this** mp3 that doesnt work. May
7	"If you want to finish	**this** relationship please dont piss me
8	ment..." "NAME187 knows	**this** household all too well... Nasty c
9	;o) h.x" "But what is	**this** mysterious something which is
10	ing. Told that i'm going to	**this** dinner this evening, so not lookin
11	go?I really have to set	**this** essay finished off - only wanted
12	I'm such an idiot for digging	**this** hole for myself. My plan is to get
13	n" "You've got to talk to	**this** rep person for the room, anyway
14	see u again getting bored of	**this** bridgwater banter don't worry see
15	ou soon s xx" "How about	**this** one?" "Thank you i shall and
16	X x x x" "Fucking hell.	**This** man is talking complete shit. I n
17	or msn anymore?" "A lot of	**this** sickness thing going round.
18	gonna make us wait about	**this** bloody job? Tho if the conversati
19	All good mate. you going to	**this** thing in london on the 15 me

The referents, as suggested above, tend to be those in the texter's immediate vicinity (either spatially or temporally): *this mp3*, *this dinner this evening*, *this sickness thing going round*. What is apparent from these examples is the extent to which, while some uses of *this* seem chiefly to be referential, many if not most also play evaluative roles in their indication of familiarity or immediacy similar to that noted by Carter and McCarthy. However, these references are critical rather than positive (in contrast to Carter and McCarthy's observations of speech), with the following making an overtly negative reference.

(8.79) Hello. Damn this christmas thing. I think i have decided to keep this mp3 that doesnt work.

(8.80) Think i might have to give it a miss. Am teaching til twelve, then have lecture at two. Damn this working thing.

(8.81) Not getting anywhere with this damn job hunting over here!

(8.82) Ok but knackered. Just came home and went to sleep! Not good at this full time work lark.

(8.83) what I meant to say is cant wait to see u again getting bored of this bridgwater banter

One example of *this* in a texted narrative serves both to heighten immediacy and express disapproval:

(8.84) I've been barred from all B and Q stores for life!?This twat in orange dungerees came up to me and asked if I wanted decking? So I got the first punch in!!

Remaining references with *this* are neither particularly critical nor positive, but appeal to the familiarity of the reference to the textee:

(8.85) That's ok. I popped in to ask bout something and she said you'd been in. Are you around tonght wen this girl comes?

(8.86) Hey chief, can you give me a bell when you get this. Need to talk to you about this royal visit on the 1st june. NAME182

Concordances for *that* in indicating shared experience include the following.

Figure 8.17 Selected concordances for *that* in indicating shared experience

```
1      can"  "think of anything to add to that list..?"  "Tee hee just read
2       u. Oh just remembered, ur meeting that guy from last wk. Hope ur
3   going to text others too. Keep out of that sun today x"  "Hello! I've book
4    k you for out presents. We will have that party soon-unless I've another
5    ool! Every pen helps! Xx"  "I've set that picture of the big spider i
6       ow dad my photos. Now get on with that marking young lady!"  "NAME28's
7     st did? Quick have a cuppa!"  "Turn that noise down a bit will ya"
```

8 "Ur wrong, u r!" "What the f was **that** recording about?" "What's goi
9 lies already! Sad I know! xxx" "Was **that** girl who came nice?" "What
10 s 2mora night! see you" "Ah i see, **that** quiet little drink is so going
11 at does she do? She aint as old as **that** other bloke is she, its a
12 love u more than ever xxx" "got it! **that** nasal hair has just been pluck
13 u heard about that job? I'm going to **that** wildlife talk again tonight
14 u give me the name and number of **that** landscape chap please"
15 comprehend the time that went into **that** puzzle. it was a true mission.
16 maybe getting their top up. Just like **that** vodafone case!!!" "Pole dancin
17 off so they re cancelled me AGAIN! **that** NAME11 needs to b sacked"
18 me to me, sharpish" "finally! got **that** piece of bacon out from between
19 y hi to NAME224 from me." "Love **that** holiday Monday feeling even
20 i'm meant to b blonde under all **that** ginger! What a tool! Xxx"

These concordance lines suggest that *that* is used in texting to indicate things, people and places familiar to both participants but, rather than referring to things in the immediate vicinity of the texter (as with *this*), it instead refers back to previous shared experience or to events and objects outside the texted interaction. Similarly, Carter and McCarthy suggest that *that* can be used in spoken narratives for things listeners know well, as in: *Then we drove round **that** big roundabout and over **that** bridge, you know the one just as you come into Norwich*. Examples can be picked out of the above concordances which show how *that* indicates events or objects familiar to the textee: *What was the name of that catalogue* is not a reference to a spatially-distant object but is used to recall shared knowledge, as is: *I've got an interview for that job in Exeter!*

In CorTxt, *that* (like *this*) also expresses negative attitudes. Those that express disapproval or criticism (ironically or otherwise) include the following.

(8.87) Gosh that NAME62, what a pain. Spose I better come then.

(8.88) would u fuckin believe it they didnt know i had thurs pre booked off so they re cancelled me AGAIN! that NAME11 needs to b sacked

(8.89) Sorry im getting up now, feel really bad- totally rejected that kinda me thing.

Association of *that* with disapproval makes us interpret other superficially neutral (or decontextualised) examples such as the following in a similarly disapproving way.

(8.90) Dont flatter yourself... Tell that man of mine two pints of carlin in ten minutes please.... X

(8.91) Hello. Sort of out in town already. That NAME219. So dont rush home, I am eating nachos. Will let you know eta. X

That is not, however, solely negative. Other occurrences, such as the first two below, seem to be positively evaluated, while others are not obviously evaluative at all.

(8.92) hiya hows it going in sunny africa? hope u r avin a good time. give that big old silver back a big kiss from me. x

(8.93) Love that holiday Monday feeling even if I have to go to the dentists in an hour

(8.94) Hey mr NAME242 whats the name of that bill brison book the one about language and words x

A final note should be made of the pattern *this*/*that* + noun + *thing* which is used (as in spoken English) in an apparently derogatory or dismissive way. The pattern extends to use of *malarky*, *lark*, *shit* and *stuff*, as well as *thing*.

Figure 8.18 Selected concordances with *this*/*that* + noun + *thing*

```
1     too where you at"  "What with all this working malarky and your comm
2    NAME127 can't drive on saturday to this training thing cos her car has bro
3          and who's NAME49? need to do this website thing, but should be finis
4    catch up sometime. Take care. X"  "This pen thing is beyond a joke. Wo
5        I think i have decided to keep this mp3 thing that doesnt work. May
6     on rs or msn anymore?"  "A lot of this sickness thing going round. Tak
7    Christ im famished."  "Hello. Damn this christmas thing. I think i have de
8        then have lecture at two. Damn this working thing."  "I probably could
9    ome and went to sleep! Not good at this full time work lark. You? Have yo

10   rubbish. And dont try using any of that whacky science shit to prove
11         ok 4 tom?"  "U want me 2 day that census stuff thru? G"  "U'll b
12         material options. We climbed that hill thing first thing - before br
13       bacardi in the house? What was that cap thing she wanted called?
14         really bad- totally rejected that kinda me thing. Think I might
```

The relative length of this structure is interesting. *Damn this christmas thing* could feasibly be paraphrased as *damn christmas* and choice of the former indicates concern for emphasising sentiment over that of brevity.

8.4.4.6 Conclusion

In conclusion, *this* and *that* in CorTxt occur frequently in references to shared and often future time, which can be related to production constraints or affordances and the interpersonal functions fulfilled. Given the expectation and general observance that delays between text messages will be fairly minimal, texters can assume that they and their interlocutor are working within the same time framework (and, if not, the date and time at which text messages are sent is generally recorded by the phone); while frequent references to future time indicate the role of texting in making arrangements or securing future interaction. However, texting does not take place face-to-face and thus the extent to which texters refer to what is around them through deictic reference is limited. The context that texters do share is 'virtual space' and *this* and *that* are used to refer to the medium and to the current interaction. In other words, the investigation of deixis highlights the use of language in referring to, or construing, not only the confines of the immediate interaction but also the shared often future world that lies beyond the texted interaction: as concluded from the phrasal investigation in Chapter 7. Furthermore, the fact that texters tend to be intimates with relationships extending beyond texting means that reference can be made to shared knowledge of current events and situations through *this* and *that*. What is interesting about this is the role of the determiners in evaluating these events and situations: serving perhaps, as with the use of creativity and repetition, as a necessary stance marker in a seemingly constrained medium.

8.4.5 Headers and tails

The terms *headers* and *tails* are used by Carter and McCarthy (2006: 192-196) to describe certain non-clausal items placed either before or after the clause. These items are typically noun phrases and are 'repeated' within the clause with a pronoun. Carter and McCarthy's examples of headers and tails in speech include:

Your sister, is she coming too? (header)
He's amazingly clever, **that dog of theirs** (tail)

(Carter and McCarthy 2006: 194)

Headers and tails are varied, with potentially complex structures,[1] and are typical and functional aspects of spoken language (Carter and McCarthy 2006). Carter and McCarthy reject the terms *left* and *right dislocation* as suggesting a perversion of the 'correct' rules with which written language is described. Instead, headers and tails are 'listener sensitive' (Carter and McCarthy, 2006: 196), helping, in the case of headers, to orient listerners and prepare them for new information and, in the case of tails, clarifying or reinforcing what has been said. One criticism of their description of headers is that it presupposes that the speaker is able to predict the need to orient the listener before they actually begin speaking. Eggins and Slade's (1997) description of the same feature suggests that speakers first provides a lengthy subject, and then use a pronoun with the same referent in order to link the subject to the rest of the clause. Regardless of their exact function, however, it appears that headers and tails occur only infrequently in CorTxt.

Drawbacks to this investigation of headers and tails in texting, as presumably with spoken data, include the lack of regular surface features for use in retrieval, as well as the possible length and complexity of the structures. The process cannot easily be conducted automatically, even with tagged corpora. The procedure followed in this study was to search for pronouns and to look line-by-line through concordances for noun phrases occurring within a wide span to the left (headers) and the right (tails): a time-consuming task only possible given the size of CorTxt.

Headers in texting were found to be particularly infrequent and it proved difficult to retrieve clear examples. The only possible examples retrieved from CorTxt included the following (where headers are in bold and pronoun within the clauses underlined).

(8.95) Well done, blimey, **exercise**, yeah, i kinda remember wot <u>that</u> is, hmm. Xx

(8.96) **NAME78, NAME416, NAME366** and **NAME417** picking <u>them</u> up from various points

[1] Headers, according to Carter and McCarthy (2006: 192-194), may comprise a sequence of noun phrases, or non-finite clauses, and refer to various elements within the clause: subject, object, object complement, prepositional complement. Tails may also constitute complex noun-phrases, as well as prepositional clauses or phrases, and they can occur with statement- and question-tags. They can refer to objects, complements or adjuncts in the preceding clause as well as subjects (Carter and McCarthy 2006: 192-194).

(8.97) **My friend**, she's studying at warwick, we've planned to go shopping and to concert tmw, but it may be canceled, havn't seen NAME54 for ages, yeah we should get together sometime!

While the first is more clearly a header, albeit one obscured in speech-like chains of non-clausal elements, all three could alternatively be described as constituting short responses to previous questions (*exercise; NAME78, NAME416, NAME366 and NAME417*; *My friend*), followed by elaboration (*i kinda remember wot that is; picking them up from various points*; *she's studying at Warwick*). This interpretation is challenged only by punctuation which (as these and other examples show) is often unconventional in texting.

Tails occur in CorTxt, but do so infrequently. Searching the corpus produced 11 examples, including:

(8.98) Hasn't that been the pattern recently **crap weekends**?

(8.99) it's really getting me down **just hanging around**.

(8.100) Oh yes, why is it like torture **watching england**?

(8.101) Yep then is fine **7.30 or 8.30 for ice age**.

(8.102) Hello! Just got here, **st andrews**-boy its a long way! Its cold. I will keep you posted

However, there appear to be very few examples in a corpus of nearly 200,000 words, even if we accept that some examples will have been missed, given retrieval difficulties. Putting aside methodological issues for the moment and assuming that tails occur only infrequently in texting, and headers less often, the general lack of both can perhaps be put down to features of the medium and of texted interaction. The lack of headers reflects the lack of lengthy noun phrases, most subjects being pronouns; as well as the fact that text messages can be edited and scanned (rather than processed online), thus reducing the need (in speech) to orient listeners. The lack of tails is perhaps indicative of the intimacy of the relationship between texters and their familiarity with the texted content.

8.4.6 Vague language

8.4.6.1 Vague expressions in speech

Vague terms such as *or something* and *kind of* are, according to Carter and McCarthy (2006: 202), purposeful and prevalent in spoken interaction and they go on to give the following example.

Between then and **like** nineteen eighty four I just spent the whole time, I mean for that whole **sort of** twelve year period **or whatever**, erm I was just working with just lots and lots and lots of different people.

Vagueness softens otherwise direct or authoritative statements while suggesting shared knowledge or marking group membership. Examples such as the following from Carter and McCarthy (2006: 202) show that vague language can sometimes be described as 'necessary' when interacting, to avoid being too lengthy or explicit (in this case, listing all available drinks). It is therefore not lazy or sloppy but 'motivated and purposeful'.

A: *She doesn't like coffee*.
B: *Well, she can have an orange juice,* ***or something***.

Other examples of vague language include the following.

Figure 8.19 Vague language in speech

like, sort of, kind of, and that, and so on, or something, or anything, or so, whatever, thing and *stuff.*

Of these, write Carter and McCarthy (2006: 203), the three expressions *and things, and stuff* and *and that* prove particularly frequent and flexible in CANCODE.

8.4.6.2 Vague language in texting

Investigation of the above vague expressions in texting reveals firstly that they do occur and that their uses, despite differences, are similar to that in spoken language. Frequencies of vague words and phrases in CorTxt are shown below.

Table 8.1 Vague terms in CorTxt

Vague term	**Frequency**
thing (117) / things (74)	201
stuff (and stuff)	104 (10)
or something	27
sort of	16
kind of	10
or so	10
like (adv)	7
or whatever	6
and that	4
or anything	2
and so on	0

Immediate observations arising from this table is that *and stuff* and *and that* occur very infrequently, while *and things* does not occur at all.[1] *Thing(s)* and *stuff* are, as we may expect through their versatility and potential to occur within other phrases, very frequent; so, perhaps more surprisingly, is *or something*. The vague items listed above are explored below, in ascending order of frequency.

8.4.6.3 or anything, and that, or whatever, like, and or so[2]

The occurrence of these quite lengthy phrases suggests some prioritising of interpersonal and speech-like informality over technological constraints. The least frequent vague item in CorTxt, *or anything*, occurs just twice, although the two occurrences are interesting in their differences.

(8.103) Hello! How's you and how did saturday go? I was just texting to see if you'd decided to do anything tomo. Not that i'm trying to invite myself or anything!

[1] Apart from this, the only other item not to occur in CorTxt is *and so on*.
[2] This group of vague terms are grouped together not so much because of shared functions (although there are similarities) but because of their similarly low frequencies.

(8.104) Did he say how fantastic I am by any chance, or anything need a bigger life lift as losing the will 2 live, do you think I would be the first person 2 die from N V Q? XXX

The first occurrence above is a clearer example of *or anything* as a vague item, used by the texter to further soften their indirect request to join in with any planned activities. In the second, in contrast, the texter is more specifically asking whether 'he' has said anything nice about her at all. This distinction is reinforced by the surrounding lexical choices: *or anything* in the first example is accompanied by other attempts to hedge the request, and to distance the texter from it: *I was just texting to see if you'd decided to do anything*; while the first is surrounded by emotional, figurative items which heightens the emotional intensity and the intimacy between the texters: *by any chance, need a bigger life lift as losing the will to live, the first person 2 die from N V Q*. In other words, while the second is not strictly speaking an example of vague language, it does seem to be working on a highly interpersonal level. Bearing in mind that the fact that there are only two occurrences of this particular phrase mitigates against its significance in CorTxt, it is interesting in the way that it draws attention to the varied interpersonal and evaluative functions fulfilled in texting.

The next most frequent item, *and that*, occurs four times in CorTxt. It functions similarly to the more frequent *or something*, explored below, in the sense that it indicates longer lists of people, activities or things than is actually mentioned or 'needs' to be mentioned, given the shared understanding and background knowledge. The effect of this is, as in speech, that of familiarity and intimacy between texters:

(8.105) Pick you up bout 7.30ish? What time are NAME159 and that going?

(8.106) Erm … ill pick you up at about 6.45pm. That'll give enough time to get there, park and that. X

(8.107) Sorry to be a pain. Is it ok if we meet another night? I spent late afternoon in casualty and that means i haven't done any of y stuff42moro and that includes all my time sheets and that. Sorry. Xx

Two further examples of *and all that* appear to have a more evaluative function than *and that*. In both cases, the phrase also indicates the texter's awareness of their figurative language, and an assumption as to their interlocutor's awareness.

(8.108) here is my new address POSTALADDRESS-apples&pairs&all that malarky

(8.109) Dont give a monkeys wot they think and i certainly don't mind. Any friend of mine&all that**!** Just don't sleep wiv NAME227, that wud be annoyin! Xxx

There are 5 occurrences of *or whatever*, again similar in use to *or something*.

Figure 8.20 Concordances for *or whatever*

```
1            "... Oh and bring wine or whatever you want to drink if
2         We have a lift. Bring wine or whatever you want to drink if
3       as happy as a pig in clover or whatever the saying is! X"  "O
4            I cud stay at NAME159s, or whatever, so dont worry about
5          that2worzels and a wizzle or whatever it is?! Xx"  "That's rid
```

The meaning in some of these examples is somewhat similar to that of *or something*, particularly in *I cud stay at NAME159s, or whatever*, the difference being a heightened sense that the outcome is not important to the texter and it does not matter which option is picked. As well as around 30 uses of *whatever* as determiner or pronoun (as above), there are also 6 non-clausal exclamations, which signify that the texter is uninterested in, or does not care about, what has been said. This is a highly informal, modern usage.

(8.110) whatever, im pretty pissed off.

(8.111) She's borderline but yeah whatever.

(8.112) Yeah whatever lol

Like as an informal adverb expressing vagueness occurs 7 times in CorTxt, 3 in relation to approximate time. The adverb occurs before the noun phrase it modifies (*an hour, hills and shit*; and, in one case, the 'verb phrase' *proper tongued*), and after the adjective *creepy*. The

use of this adverb is interesting, as it tends to be stigmatised in speech, and it does seem that the examples in CorTxt accompany socially taboo words or topics.

(8.113) Hmm ok, i'll stay for like an hour cos my eye is really sore!

(8.114) Serious? What like proper tongued her

(8.115) Hey cutie. How goes it? Here in WALES its kinda ok. There is like hills and shit but i still avent killed myself. NAME500.

(8.116) I'll text NAME227 now! All creepy like so he won't think that we forgot

Like also occurs 13 times in other vague expressions. *Something like* occurs 8 times, of which 5 are followed by *that* and one by *this*. As discussed below in relation to *or something* (that is, *or something like that*), this phrase serves to indicate uncertainty, hedge authority, or make something less specific.

Figure 8.21 Concordances for *something like*

```
1      king of you- ahh!"  "Something like communications and campai
2     africa."  "Ermines or something like that. Its driving me mad!"
3           auto centre (or something like that) on Newport road. I
4     early evening... 6, 7 something like that..."  "Not at home so will
5      es un buen tiempo or something like that. My love to NAME219 a
6       a while to get over something like dengue fever."  "I'll check wh
7     into perspective when something like this happens and it is nice
8         it's like2be told something like that! Xx"  "We can share the
```

stuff like that occurs once and *things like that* twice:

Figure 8.22 Concordances for *stuff/things like that*

```
1     didn't really talk bout stuff like that, but he's really lookin4w20
2   not seeing me, then says things like that! NAME281 said he'd co
3           young to look at things like that!"  "Ooh sorry. Table at i
```

Finally, *and the like* occurs twice.

Figure 8.23 Concordances for *and the like*

```
1       ed time, sipping sherries and the like.... anyway, have big fun. c u
2      all the details of courses and the like. should be good"  "You kno
```

These uses of *like* again have similarities in terms of function with *or something*, and *and that*.

The 9 occurrences of *or so* in CorTxt are in relation to time. This use of the phrase would presumably also be frequent in speech, but its use here solely in relation to time perhaps highlights the role which texting plays in making arrangements.

Figure 8.24 Concordances for *or so*

```
1      ou make it a bit later. 8.30 or so."  "Yes no problem. s
2        ut can you give me an hour or so please as have just g
3      le!"  "Ok, prob bout an hour or so. Bring baccy."  "If.its.o
4       "Can you leave it half hour or so as i am cooking som
5      ng for town in nxt half hour or so. Let me know when u
6       i meet you in the next hour or so?"  "See you there an
7       be cool to meet up for a hr or so"  "Who is this?"  "S.
8         the room. ouch"  "10 mins or so"  "Yes. be there in a
9        you sms reports in a month or so - it'll be worth having
```

8.4.6.4 Kind of *and* sort of

Kind of and *sort of* function similarly in hedging propositions. *Kind of* occurs 4 times as a vague item, used to hedge outcomes, assertions and demands. The use of *kind of* as a response token in these two examples serves to hedge texters' answers as to the state of a car and success of a journey.

(8.117)

A:	Dare i ask... Any luck with sorting out the car?
B:	Kind of. Took it to garage. Centre part of exhaust needs replacing. Part ordered n taking it to be fixed tomo morning.

(8.118)

Kind of. Just missed train cos of asthma attack, nxt one in half hr so driving in. not sure where to park.

In the following, *i kind of am* hedges the extent of the texter's feelings and serves to sound almost apologetic. In the second, *thats the kind of thing* similarly softens the texter's assertion.

(8.119) Hi NAME242, where are you? We're at NAME343's and they're not keen to go out i kind of am but feel i shouldn't so can we go out tomo, don't mind do you?

(8.120) I thought i'd get him a watch, just cos thats the kind of thing u get4an18th. And he loves NAME51's so much!

The use of *some kind of* in the example below similarly softens the texter's criticism or themselves.

(8.121) I must be some kind of obsessive party animal... Or perhaps i'm actually less full of busy. Didn't see anyone on tues, so still have cards, sorry! Apparently its a french thing to give cards in the new year, you could try that... Have fab hols, see you soon for another friday night bash xx

What kind of time do you want to meet? in the following hedges a demand, by making it clear that the texter is not demanding to know a specific time. *A meet and greet kind of affair* in the second example similarly softens a texter's request.

(8.122) Hiya. How was last night? I've been naughty and bought myself clothes and NAME189 very little ... Ready for more shopping tho! What kind of time do you wanna meet?

(8.123) NAME242, im NAME318.. On the snowboarding trip. I was wondering if your planning to get everyone together befor we go..a meet and greet kind of affair? Cheers, NAME318

Two further occurrences are also evaluative rather than referential, although it is not clear that the purpose is vagueness. *Any kind of fumbling* in the first example adds to its assertiveness, while *Kind of person who* similarly emphasises the category into which NAME79 is placed.

(8.124) Aah! A cuddle would be lush! I'd need lots of tea and soup before any kind of fumbling!

(8.125) NAME79 said kiss, kiss, i can't do the sound effects! He is a gorgeous man isn't he! Kind of person who needs a smile to brighten his day! Xx

Sort of occurs 16 times in a vague sense. The following 5 refer to an approximate time.

Figure 8.25 *Sort of* in reference to approximate times

```
1     Spose I better come then. What sort of time are we kicking off? A
2      e of booze or coffee!"  "What sort of time if ur there 4 ish u ca
3       "  "You dirty stop out! What sort of time today. I'm in college
4     When ur ready Let us know what sort of time would b good 2 com
5        from now on!"  "Ok, so what sort of time would be convenient
```

The following 8 downplay an event, action or object.

Figure 8.26 *Sort of* in reference to events, actions and objects

```
1       x"  "On me way x"  "Hello. Sort of out in town already. That
2    the wrong thing but i already sort of invited NAME79-tho he m
3      t somethin out thankyou its sort of an emergancy"  "On way
4      hes also in denial!"  "They sort of offered it to me already-all
5        nt! What do you mean they sort of offered it to you? Are you
6    how are you - we are having a sort of first birthday party for NM
7          i have necklace, i have sort of cardi, no shoes tho... Do
8     Will be good to see u."  "Oh sort of pleasure. Going to c siste
```

The following 3 serve to make statements less specific.

Figure 8.27 *Sort of* used to downplay statements

```
1    out not bein good enough, that sort of thing. I can't keep dealing
2    ber! I think I should get some sort of medal!"  "Tottenham lost!"
3      eat though"  "D'you get some sort of payslip for claims u make
```

8.4.6.5 or something

Or something occurs 34 times in CorTxt, and these can be grouped into four main usages: to indicate a range of options, to hedge otherwise authoritative sounding propositions, to hedge otherwise harsh criticism or comment, or to indicate uncertainty. The first usage,

accounting for 18 occurrences, seems to suggest that the texter is not being prescriptive but leaving a degree of choice open to the textee.

(8.126) Let me know if you need anything else. Salad or desert or something... How many beers shall i get?

(8.127) Hello! Good week? Fancy a drink or something later?

(8.128) Maybe i could get book out tomo then return it immediately ..? Or something.

The use of the phrase to mitigate otherwise authoritative assertions accounts for another 6 occurrences including the following.

(8.129) I got another job! The one at the hospital doing data analysis or something, starts on monday! Not sure when my thesis will got finished

(8.130) Nationwide auto centre (or something like that) on Newport road. I liked them there

(8.131) Que pases un buen tiempo or something like that

Another 5 occurrences express uncertainty: that is, that the texter is not sure exactly what has happened, or what to do:

(8.132) Should I have picked up a receipt or something earlier

(8.133) Just hopeing that NAME281 wasn't too pissed up to remember and has gone off to his sisters or something!

In the final 3 occurrences, *or something* softens otherwise harsh statements:

(8.134) Have you emigrated or something? Ok maybe 5.30 was a bit hopeful...

(8.135) Hi, wlcome back, did wonder if you got eaten by a lion or something, nothing much

happened here while you were adventuring

(8.136)

Have you not finished work yet or something?

8.4.6.6 stuff *and* thing

The first point to reiterate is that *and things*, apparently 'extremely frequent' in spoken language (Carter and McCarthy 2006: 203), does not occur in CorTxt, while *and stuff* occurs just 10 times.

Figure 8.28 Concordances for *n stuff* / *and stuff*

1	ool cool! Just been doing work	**n stuff**. Busy at work? Have a nice
2	ool in Newquay with the flumes	**n stuff** tomorrow? say yes, you know u
3	How are you getting to work	**and stuff**. I miss you craziness on
4	Glad your having a good time	**and stuff**. X" "Recieved on rout" "And
5	Speak to you later bout times	**and stuff**. Xx" "Am here now and
6	n! Hope you had a good day	**and stuff** wiv your bike went well! Mwah!
7	and where abouts is the tent	**and stuff** cos we will need it for venice
8	f to the bank to sort out rent	**and stuff** then ill give you a call,
9	stressed out by the crowds	**and stuff** but still had some great times.
10	m not bad. Bumming around. Pub	**and stuff**. Glad you're having a good

The terms *stuff* (104), *things* (75) and *thing* (117) do however occur fairly frequently in CorTxt and are used, as could be expected, in a variety of similar ways, although it would appear that *stuff* has a somewhat narrower range of usages than the other two. Many uses of *stuff*, *thing* and *things* can be grouped into two main categories: *stuff* or *things* to do, and actual objects, generally belongings. There are 8 occurrences of *stuff* in the sense of stuff to do.

Figure 8.29 Concordances for *stuff* to do

1	oney! I cant go to the talk,	**stuff** has piled up that i must get do
2	I can't stay over as have	**stuff** to do early on sat but maybe
3	u stay til fri but will have	**stuff** to do before f comes. Does that
4	on the go! Am trying to get	**stuff** done instead, before leaving tom
5	uiet here,trying to finish	**stuff** - article, s and m- before NAME24
6	be home by half six as doing	**stuff** for mother." "I am! Yum yum
7	tomo to have whole day to do	**stuff** on fri. See you at lunchtime
8	l be there at 3. point me at	**stuff** that needs doing." "I can help

There are 24 occurrences of *things* to do, of which the following are examples.

Figure 8.30 Selected concordances for *things* to do

1	On my way to campus now - **things** to do for conference but let m
2	t u! That'll learn me2do **things** ina rush!" "Yes that's fine,
3	erday. Took on too many **things** in one day! Congrats for finally
4	tter than all the great **things** that we have done! In a strange
5	very behind with getting **things** ready. x" "Is NAME2 coming to
6	"sorry, no, have got few **things** to do. may be in pub later."
7	ya! Sorry but got a few **things** to get done tomorrow, thanks for
8	up tremough and get a few **things** done- where are you at?" "Hey
9	ry just Think before ya do **things**! Hugs x" "Hi - dad just texted m
10	old fella. gs x" "Why do **things** always take longer than planned

There are just 4 occurrences of *thing* to do and these appear more evaluative than *things* or *stuff*:

(8.137) Ah, well that confuses things, doesnt it? I thought NAME90 was friends with NAME159 now. Maybe i did the wrong thing but i already sort of invited NAME79- tho he may not come cos of money.

(8.138) Ha. You don't know either. I did a a clever but simple thing with pears the other day, perfect for christmas.

In the second category, where the reference is to actual objects, there are 51 occurrences of *stuff*, including the following.

Figure 8.31 Selected concordances for *stuff* in reference to actual objects

1	to me! I'm guarding NAME262's **stuff** wiv my life at the mo,
2	home this weekend to get more **stuff** as came with very little.
3	d u can stay over! XXX" "How much **stuff** will u have with u as we
4	easier for u to sort, esp if u hav **stuff**. I hav v little time to get
5	"Ok. See ya soon." "Hi, have **stuff** to buy, car is on meter
6	ll keep u posted. I can maybe get **stuff** if quick and easy. Choc
7	hem 4 me plse, w collect w other **stuff**. Vodka is 4 anyone 2 have.
8	be back NAME392s goin 2 collect her **stuff**. Ive told him he has 2 clean
9	I spoke to ur mum, i will drop ur **stuff** off at their flat tomo.
10	ng in, in 20 mins to pick up some **stuff**." "The thing is i'm not
11	tar he can meet u and keep your **stuff** in london til we come back fr
12	sure thing mate haunt got all my **stuff** sorted but im going sound anywa

13 d NAME50. Let me know cost of **stuff** you get for them, dont go mad.
14 "I did yeah, just to pick up the **stuff** that i wanted2take home. We
15 rning. Just finding places for my **stuff** to go for 3 weeks maybe at tash
16 night or will it be pain to carry **stuff**? If yes, can drop it off monday.
17 "My dad says if u bring all your **stuff** to london on the eurostar he
18 , great idea. Need slash want some **stuff** from town anyway. See ya later
19 x" "Now i'm confused, i sent you **stuff**. Honest!. Gimmie a kiss x x x x
20 card ready." "I could try and get **stuff** later and make them here... Or

Two occurrences of *stuff* are ambiguous as to whether actual objects are being sorted, or things being done.

(8.139) Okey dokey, i'll be over in a bit just sorting some stuff out.

(8.140) Did u find out what time the bus is at coz i need to sort some stuff out.

There are 9 occurrences of *things* as objects:

Figure 8.32 Concordances for *things* as objects

1 i can remember. You should see the **things** i stole from behind the bar
2 n. I will see if i can think of some **things** i need to buy! What time you
3 ve spoken to mum. There's a few small **things** on the list i have her.
4 to ask a few questions about sending **things** off to print. Let me no
5 357 and NAME215 i hope they post **things** out. Im off on hols tomorrow
6 some xmas! Hope u get lots of nice **things**, eat lots and get smashed! :-)
7 farmers market, bought lots of lovely **things**. How's your head?" "Have fun
8 thirty. Need to know cos will bring **things** for supervision if staying
9 "Sorry computers, brb x" "Bloody **things** takin for ever to load, brb x

There are 10 occurrences of *thing* as an object:

Figure 8.33 Concordances for *thing* as object

1 tely right I can't find a single **thing** to year to the christmas d
2 if late!X" "Der. Just sending **thing** to dad now, so do have ad
3 I'll be forced to send the real **thing** round.....xx" "I know. C
4 i still need 2go.did u c d little **thing** i left in the lounge?it was f
5 ap mirage thing. One more little **thing** maybe? X" "Is the cd so
6 ed on me car this month-bloody **thing**!will try n get down soon h
7 ceiving text messages. Bloody **thing**. How are you. How was bi
8 bed is quite possibly the best **thing**.. ever" "Sorry if i sounded

```
9           where near ready. Not packed a thing yet!"  "are you getting dre
10           sure a phone call is the last thing she wants, but would appr
```

In an evaluative use of *thing*, *things* and *stuff*, the three words post-modify nouns. Examples with *thing* include these.

(8.141) This pen thing is beyond a joke. Wont a Biro do? Don't do a masters as can't do this ever again! X x

(8.142) A lot of this sickness thing going round. Take it easy. Hope u feel better soon. Lol x

(8.143) Neither [in sterm voice] - i'm studying. All fine with me! Not sure the NAME357-NAME215 thing will be resolved, tho. Anyway. Have a fab hols

The structure *this/that + noun + thing* was described above in section 8.4.4 on deixis[1] but, as the examples above show, the structure *noun + thing* can also occur with other determiners: as in *the NAME357-NAME215 thing*. There are 60 occurrences of this structure in CorTxt, and other examples include the following.

Figure 8.34 noun + *thing*

the car thing, *this mp3 thing*, *a Sunday afternoon thing*, *the family thing*, *work thing*, *the whole hugh laurie thing*, *nhs thing*, *that hill thing*, *my birthday thing* and *stranger person thing*.

The structure also occurs 9 times with *things*, including: *need to ask about car things*; *i have family xmas things*; *the bloody Sequin stud things on my top*. In most cases the addition of *thing* or *things* creates a negative or dismissive tone.

Stuff also occurs 19 times in the above structure.

Figure 8.35 Concordances for noun + *stuff*

```
1           not too busy with third year stuff the we must rondevous. If
2             hear about horrible period stuff... Just had my most painless
3           re's more of that evaluation stuff than i thought! X"  "Hi, was
```

[1] Where it was pointed out that other words such as *lark* could also take the place of *thing*.

4 tests. Need to sort library **stuff** out at some point tomo - got l
5 sold. Email me your bank **stuff** and i will zap you wot youre owe
6 at end of term but yes work **stuff** doesn't change. Anyhow, see
7 ant make monday, bday **stuff** is ongoing, despite my best
8 ant me 2 day that census **stuff** thru? G" "U'll b able to b
9 guys doing the car parking **stuff** tomorow? Cheers, NAME505 T"
10 "Got all the centre parks **stuff** through and we can pre book
11 ood luck with daily life **stuff** x" "Gosh that NAME62, what
12 and borrow some kitchen **stuff** if u need. X" "Cheers. Thanks
13 nd romantic weekend away **stuff**. Have a great time if i don't"
14 could give to dad, shower **stuff**, if we want. Mum says which be
15 U can just chuck my car **stuff** in my room if u want!" "Hi, got
16 been and put some mastik **stuff** on the side of the bath, badly!
17 Been following strike **stuff** on fr radio too. Mad country!
18 oops just sent materials **stuff** to your bham address - let me
19 ME349 pulled out, family **stuff**. Still quite a few comin" "HI.

The remaining 10 occurrences of *stuff* refer to things in a more abstract sense, many of which also appear to be evaluative (in positive, negative or dismissive senses) in contrast to its use in referring to actions or objects. Examples include these.

(8.144) Good stuff, will do.

(8.145) Okey dokey swashbuckling stuff what oh.

(8.146) Yeah work is fine, started last week, all the same stuff as before, dull but easy and guys are fun!

Things and *thing* also occur in a more abstract sense, 19 and 5 times respectively, of which the following are examples. Again, examples of *things* appear to be somewhat evaluative: there are, for example, *mundane things*, *good things*, and *bad things*:

Figure 8.36 Selected concordances for *things* used in an abstract sense

1 but it's good 2 learn new **things** about yourself! Glad we r coo
2 shocked at the most mundane **things**. I am writing essay in thrid
3 hat comes of having too many **things** on the go! Am trying to get s
4 nd NAME159 may be many **things**, but they've never asked Me
5 king like a few hundred. its **things** like my next rent that are wo
6 chilled out and fun and good **things** will happen!, bit me a NAME3

7 Hope that explains a few **things** but doesnt change things!X"
8 things. He really has had bad **things** happen 2 him in the past. He
9 d some very young views about **things** that i thought might annoy u!
10 so she can contact u about **things** once in the uk! Cant believe

The five occurrences of *thing* in an abstract sense are also evaluative:

(8.147) It is a good thing I'm now getting the connection to bw

(8.148) Just got part Nottingham - 3 hrs 63miles. Good thing i love my man so much, but only doing 40mph. Hey ho

(8.149) I don't know, same thing that's wrong everyso often, he panicks starts goin on bout not bein good enough …

There are other uses of *thing* and *things*. 5 occurrences of *thing* and 3 of *things* are used in reference to events:

Figure 8.37 Concordances for *thing* and *things* in reference to events

1 ll excited i was. Enjoy your **thing**!" "Are you all moved? All
2 need to stay for the whole **thing**, there's only so much mul
3 you feel about the whole **thing** but obviously i want to fin
4 ood mate. you going to this **thing** in london on the 15 me ne
5 then if we miss NAME108s **thing**" "If there are no other opti
6 serbia. Def the most insane **thing** i've ever been 2. If theres

7 know when and where. Enjoy your **things**." "Hello - did you ring earlier?
8 be attending any more of the **things** at the stannary. Also i heard tha
9 ou sorry we had so many other **things** going on! Enjoy canada!" "Yes

Things is also used in reference to the status quo or to life in general. Within this category are included enquiries about interlocutors' well-being or responses to such enquiries. With just two occurrences of *how are things?*, the more typical question is *how's things?*, occurring 19 times.

(8.150) Boo. How's things? I'm back at home and a little bored already :-(

(8.151) NAME242, how's things? Just a quick question.

The statement *hope things are ok* or *went well* also occurs:

(8.152) hope things went well at 'doctors' ;) reminds me i still need 2go.did u c d little thing i left in the lounge?

The word *things* also occurs when texters describe their own well-being, as in the following.

(8.153) Whats that coming over the hill..... Is it a monster! Hope you have a great day. Things r going fine here, busy though! NAME148 x

(8.154) The whole car appreciated the last two! Dad and NAME307 are having a map reading semi argument but apart from that things are going ok. P.x

Other examples where *things* refers to the state of events include the following.

(8.155) It certainly puts things into perspective when something like this happens

(8.156) Er, hello, things didn't quite go to plan – NAME62 is limping slowly home followed by aa and with exhaust hanging off

(8.157) Ah, well that confuses things, doesn't it?

Things is also used in 9 occurrences to refer to what people have to say:

Figure 8.38 Concordances for *things* in reference to things said

```
1           i called her various things to avoid having two SBs. And N
2    tho cos he said some of the things that his real dad said to his
3            "U say the sweetest things boss!"  "U seen NAME268 this
4     t not seeing me, then says things like that! NAME281 said he'd
5      mo! u two gals run out of things to say yet?"  "u tryin to catch
6           Emmerdale? Said nice things about house mates? Didnt touch
7     pay me later 4 saying nice things x x"  "U will b fine when is the
8    want you to tell me naughty things because i want your hard cock i
9           all the embarrassing things I may of texted over the passed
```

Moving on to *thing*, six occurrences co-occur with *kind of* or *sort of*:

Figure 8.39 Concordances for *thing* with *kind of/sort of*

1	ew ppl hav had the same kinda	**thing**.she's better now stil a bit d
2	ein good enough, that sort of	**thing**. I can't keep dealing wiv it,
3	ally rejected that kinda me	**thing**. Think I might work till we i
4	ow could I resist. That kinda	**thing**. Texting you is always inter
5	t some numbers. what sort of	**thing** were you looking for" "Try
6	h, just cos thats the kind of	**thing** u get4an18th. And he loves

Thing is also used in various phrases: *first thing* (7), *sure thing* (6), *the thing* (4):

(8.158) Done it but internet connection v slow and can't send it. Will try again later or first thing tomo.

(8.159) Sure thing big man. i have hockey elections at 6, shouldn't go on longer than an hour though

(8.160) That's the thing with apes, u can fight to the death to keep something, but the minute they have it when u let go, thats it!

Thing is used to describe people (6), generally an interlocutor, as is one occurrence of *things*:

Figure 8.40 Concordances for *things/thing* used to refer to people

1	hout me and NAME270, you poor	**things**..." "Hello. How's summer
2	cope i'm sure x" "Hi Sweet	**Thing**. Glad you got home safe. I
3	u made me say that! You wicked	**thing**. Ok going to work after corr
4	him into the post box, poor	**thing**. Have a good party! I am ha
5	ovely. See ya on tues. You lucky	**thing** with your big hols to go!" "
6	n you can. X x x" "Honey sweet	**thing** we still on for tonight?" "N
7	and only has days left-poor	**thing** :) very glad u will both b at

To sum up, all three are used in vague reference to actions (things or stuff to do) and objects, as well as in a more abstract and evaluative sense (such as *the good thing is* or *this mp3 thing*). *Things* and *thing* are also used to describe events such as *this thing in London*; and *things* to describe the state of events, as in *How's things?* and to refer to what people

have to say (*run out of things to say yet?*). *Thing*, meanwhile, also collocates with *sure* and *first* and is used in the phrase *the thing (with something) is* …. These occurrences indicate, as well as the versatility and evaluative nature of the words, the intimacy and informality of texted relationships as construed through recourse to features of spoken interaction.

The above investigation of vague language in texting shows firstly that vagueness is a feature of texted as well as spoken grammar. As in speech, the vague language used in texting indicates shared background as well as interpersonal awareness as texters hedge assertions and requests. It can also play an evaluative role, mainly by marking dismissive stances. Despite the frequency of terms such as *stuff, thing(s)* and *or something*, other markers of vagueness are less frequent: *or anything, and that, like, or so*. However, the fact that vagueness occurs in texting highlights again the importance of the interpersonal over physical constraints and how texters draw on features of spoken language to perform speech-like informality. The investigation also adds to that carried out in Chapter 7 by focusing on potentially significant categories of phrases with shared functions which do not necessarily each occur with sufficient frequency to be identified in quantitative wordform counts: as do the following investigations.

8.4.7 Fixed tags

Fixed tags are question tags which do not vary in form, used in speech to check understanding or confirm that something has been agreed: *So we're meeting at 7 outside the pizza place, okay?*; *Don't tell anyone about this, yeah?* Other items used include: (*all*) *right, eh, don't you think*? (Carter and McCarthy 2006: 198).

Investigation of the above show that *yeah* and *right* are both used as fixed tags in CorTxt, although infrequently, with 3 and 4 occurrences respectively.

(8.161) Hello, yeah i've just got out of the bath and need to do my hair so i'll come up when i'm done, yeah?

(8.162) PS U no ur a grown up now right?

Tags in texting often constitute a phrase rather than a word:

(8.163) Just sent you an email – to an address with incomm in it, is that right?

These no doubt occur also in speech but, as outlined below in discussion of *ok/okay*, their use in texting suggests attempts to be less blunt given the physical constraints.

Ok/Okay is more frequently used. It is tagged, in five examples, onto the end of a unit to confirm that something has been, or is to be, agreed.

(8.164) I'll leave around four, ok?

(8.165) Spoons it is then okay?

(8.166) dont make ne plans for nxt wknd coz she wants us to come down then ok

However, fixed tags in texting are more frequently separated by full-stops (14 occurrences), rather than being tagged on, including:

(8.167) Hello NAME12. They are going to the village pub at 8 so either come here or there accordingly. Ok?

Elsewhere, other phases are used: *would that be ok?* (1), *hope (that's/it is) ok* (7), *if that's ok* (22), *sound ok* (1), *seem ok?* (4) and *is that ok?* (33). This is perhaps because these phrases are less abrupt than *okay/ok* alone, especially when there is no recourse to light-hearted tone or friendly facial expression.

(8.168) Oops sorry. Just to check that you don't mind picking me up tomo at half eight from station. Would that be ok?

(8.169) Got it. Seventeen pounds for seven hundred ml – hope ok**.**

(8.170) Please give it 2 NAME79 or i will pick it up on Tuesday evening about 8 if that is ok.

(8.171) Thought we could go out for dinner. I'll treat you! Seem ok?

(8.172) I though we shd go out n have some fun so bar in town or something – sound ok?

(8.173) It's £6 to get in, is that ok? Xx

These are not so much fixed tags in the sense that Carter and McCarthy intend, in that the questions are full clauses. However, they do fulfil the same function as fixed tags, whilst highlighting potential differences between spoken and texted grammar. That is, their use highlights on the one hand the interpersonal nature of texting, and on the other, the fact that distance between texters and the potential impersonality of the medium mitigates against the use of the more blunt fixed tag, *ok?*. Again, interpersonal demands are prioritised over the need for brevity.

8.4.8 Response tokens

8.4.8.1 Response tokens in speech

Response tokens are defined by Carter and McCarthy (2006: 922) as words or phrases 'used to acknowledge what a speaker says, and to indicate on the part of the listener interest or engagement in what is being said' and as such refer to a whole preceding utterance (Carter and McCarthy 2006: 189). In spoken language, according to Carter and McCarthy (2006), response tokens can be minimal: words or sounds such as *yeah* and *mm*; or non-minimal: adjectives and adverbs which include: *absolutely*, *certainly*, *definitely*, *fine*, *good*, *great*, *indeed*, and *really*; as well as phrases or clauses such as *Is that so?*; *By all means*; *Fair enough*; *Not at all*; *True enough*; *Of course*; *What a pity!*. Because of their role as response tokens, these words and phrases are more frequent in spoken than written language. As can be seen from Carter and McCarthy's examples, one other feature of response tokens is that, where they do not comprise full turns, they occur at the beginning of the turn. Exceptions to this are when they are pre-modified (as in *jolly good* or *most definitely*, to take Carter and McCarthy's examples) or where they occur in pairs or clusters.

A: If you need some more just order some more. All right?

B: Right. Fine.

(Carter and McCarthy 2006: 191)

Clustering of response tokens, Carter and McCarthy note, is particularly frequent in certain contexts such as phone conversations or at certain points in a conversation such as topic changes or closings.

8.4.8.2 Overview of texted response tokens

The following exchanges illustrate the use of response tokens in texting. Evident in these text messages is their informality and the fact that (unlike the spoken example above), response tokens in texting are likely to precede further comment or statements; in other words, they form part of a longer turn.

(8.174)	A:	Was the farm open?
	B:	**No**! But we found a diff farm shop to buy some cheese. On way back now, can i call in?
	A:	Yeah do! Don't stand to close tho- you'll catch something!

(8.175)	A:	I'll pick you up at about 5.15pm to go to taunton if you still want to come. X
	B:	Yeah go on then, bored and depressed sittin waitin for phone to ring... Hope the wind drops though, scary
	A:	Pansy! You've been living in a jungle for two years! Its my driving you should be more worried about!
	B:	Thanks. Fills me with complete calm and reassurance! X

As in spoken language, response tokens can cluster in pairs or threes, as with *ok not a problem* and *oh right ok* in the following.

(8.176)	A:	What do u reckon as need 2 arrange transport if u can't do it, thanks x
	B:	I'm gonna say no. Sorry. I would but as normal am starting to panic about time. Sorry again! Are you seeing NAME79 on Tuesday? Xx
	A:	Ok not a problem will get them a taxi. C ing NAME79 tomorrow and tuesday. On tuesday think we r all going to the cinema. X
		Oh right, ok. I'll make sure that i do loads of work during the day! NAME281's got a really nasty cough today and is dry n shot so that should really help it!

Other exchanges show more than one response token (or more than one cluster of response tokens) can occur in one text message, as texters respond to different utterances within the

previous text message. Text messages in the following exchange [1] show how fairly inexplicit tokens relate back to utterances in a previous text message.

(8.177)	A:	Have you got Xmas radio times. If not i will get it now
	B:	No. Yes please. Been swimming?
	A:	Okay. No no, just shining on. That was meant to be signing, but that sounds better.

However, response tokens are often more explicit, especially where there is apparent room for confusion, as in *Glad you made it, I am fine* (rather than *Fine*), and *Good that he apologised* in the two exchanges below.

(8.178)	A:	We made it! Eta at taunton is 12:30 as planned, hope that's still okday?! Good to see you! :-xx
	B:	Yes i will be there. Glad you made it.

(8.179)	A:	Yeah he got in at 2 and was v apologetic. NAME159 n NAME156 had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You? Xx
	B:	Yeah I am fine, just getting up as was totally tired from 2 day at work! The nurse said that I would b sleepy 2 day after my jabs. I believe NAME281 as I no how much he likes u and the reason is so believable. Good that he apologised though. Make the most of 2 day together xxx

8.4.8.3 Survey of response tokens in CorTxt

Following this initial overview, response tokens in CorTxt were defined and retrieved quantitatively according to three criteria of position, form and function. The first criterion was that lexical items to be considered as possible response tokens occurred in text message-initial position. The item *wow*, for example, can be assumed to be a response token in this text message

(8.180)	Wow v v impressed. Have funs shopping!

[1] These examples are cited also in Chapter 6 on creativity, where the role of repetition in connecting messages is described.

but not in this, where it occurs at the end of the text message, and refers instead to the texter's own comment.

(8.181)	You'll never believe this but i have actually got off at taunton. Wow

As the initial overview above suggests, however, not all response tokens occur in text message-initial position. This is illustrated by *sorry* and *oh dear* in the following.

(8.182)		
	A:	We are hoping to get away by 7, from Langport. You still up for town tonight? X
	B:	What are you doing in langport? Sorry, but I'll probably be in bed by 9pm. It sucks being ill at xmas! When do you and NAME37 go2sri lanka? Xx
	A:	We are at grandmas. Oh dear, u still ill? I felt Shit this morning but i think i am just hungover! Another night then. We leave on sat. Xx

The fact that these non-initial response tokens are neglected in the following survey is acknowledged.

The focus on initial response tokens meant that response tokens could be retrieved by arranging all text messages in alphabetical order and manually picking out relevant items that occurred more than three times. One drawback to this method is, of course, that text messages are not explored within the wider exchange and the limitations this has to a study of response tokens are acknowledged. Response tokens were also defined by form: identified as being minimal or non-minimal tokens comparable to Carter and McCarthy's adjectives, adverbs and phrases and, in most cases, outside the clausal structure. This however raises questions as to when response tokens cease to be tokens and become full responses. For example, two- or three-word phrases such as *can do*, *fair enough* or *of course not* are clearly response tokens while responses to a previous text message such as *You'll end up a workaholic but rich* is not; but of less certainty are phrases such as *Glad you made it, Sounds god damn near perfect* or *That would be great*. The decision to include phrases such as these allows us to explore the observation that texted response tokens are often more explicit (due to delay between turns) than in speech.

The third criterion in identifying items as response tokens is that of function. Responses are identified as acknowledging a previous comment or indicating interest, rather than either providing information or functioning as discourse markers prefacing following comments. This means that every instance of each text message-initial response token was judged to be functioning as a response, bearing in mind the aforementioned limitations of doing so without reference to the preceding text messages.[1] The fact that the token is a response, rather than prefacing what the texter is about to say, is also considered. *Well*, for example, is deemed in most cases to preface following comments, and is not included as a response token.

(8.183) Well I'm going to be an aunty!

(8.184) Well I might not come then...

(8.185) Well good morning mr NAME242. Hows london treatin' ya treacle?

The response tokens retrieved through the above procedure include items which can be described as non-speech-like or text-specific: variants of *ha ha*, *tee hee*, *lol*, *xxx* and various emoticons.

Table 8.2 Text-specific response tokens

Text-specific response token	**Frequency**
ha/ha ha/haha/hahaha	32
lol	28
:-) / :-(/ ;-) / ;-(/ ;)	10
xxx	8
tee hee	4
Total no	82

[1] It is assumed in this process that use of items such as *yes* or *nope* in response to specific yes-no questions can also be described as response tokens on the following basis. Although it could be argued that responses to yes-no questions do more than 'acknowledge' the speaker or 'indicate interest', at the same time, as Carter and McCarthy (2006: 201) point out, yes-no questions in fact often function as prefaces to further questions. *Yes* and *nope*, for example, can therefore be included as response tokens, on the basis that they acknowledge the yes-no question prior to further questioning. Longer responses to yes-no questions, however, would not be included as 'tokens'.

Ha ha, *haha*, *hahaha*, *tee hee* and *lol* indicate laughter and presumably acknowledge amusing comments made by interlocutors.

(8.186) Ha ha - had popped down to the loo when you hello-ed me. Hello!

(8.187) Lol! U drunkard! Just doing my hair at d moment. Yeah still up 4 tonight. Wats the plan? Xxx

(8.188) Tee hee. Off to lecture, cheery bye bye.

Ha, however, seems to indicate triumph along the lines of 'I win!' rather than laughter:

(8.189) Ha! I wouldn't say that I just didn't read anything into way u seemed. I don't like 2 be judgemental....i save that for fridays in the pub!

(8.190) Ha. You don't know either. I did a clever yet simple thing with pears the other day, perfect for christmas.

Many of these response tokens cluster with other tokens:

(8.191) Lol no ouch but wish i'd stayed out a bit longer

(8.192) Lol! Nah wasn't too bad thanks. Its good to b home but its been quite a reality check. Hows ur day been? Did u do anything with website? X

(8.193) Lol! Oops sorry! Have fun. Xxx

Only on five occurrences do these tokens occur by themselves: *Haha*, *Haha ha*; *Ha ha!*; *Ha ha!*; *Ha.* and *Lol!*.

Emoticons, which indicate pleased or dissatisfied responses, occur on their own in 7 text messages, including ;), ;-) and :-):-).

It could be surmised that use of emoticons by themselves, as with the other tokens above, serves to close the exchange: they acknowledge the preceding text message but do not open up further interaction. In other text messages, emoticons cluster with other response tokens.

(8.194) :-) yeah! Lol. Luckily i didn't have a starring role like you!

(8.195) :-(oh well, c u later xxx

(8.196) ;-) ok. I feel like john lennon.

In two text messages, the emotion expressed by the emoticon is then elaborated on in other ways:

(8.197) :-(that's not v romantic!

(8.198) :-(sad puppy noise

Xxx, representing kisses, tends to be used to sign off and so occurs at the end of text messages, but in the 8 occurrences included in this category, they occur as full turns and so presumably fulfil the dual function of responding to the previous text message and ending the interaction.

The remaining responses (that is, those felt to occur also in speech) are divided below into minimal and non-minimal responses. The minimal response tokens are listed below in Table 8.3, in order of frequency in CorTxt. Where items are grouped together (*yup*, *yea*, *yeah*, *yeh*, *yep* and *yes*, for example, and *mm*, *mmm*, *mmmm*) it was felt that these were often spelling variants and could most usefully be studied together, although the impact that this has on frequency is acknowledged.

Table 8.3 Minimal response tokens

Minimal response token	**Frequency**	**Example(s)**
yup (10) / yea (12) / yeah (244) / yeh (15) / yep (57)/ yes (270)	608	***Yeah*** *no probs - last night is obviously catching up with you... Speak soon xx*
ok / okay (37)	369	***Okay****, good, no problem, and thanx!*
nah / no / nope	241	***Nope*** *thats fine. I might have a nap tho! S*
o (8) / oh (211)/ ohhh (1)	220	***Oh dear****, poor u. Getting away for bit, sounds like good idea. Can do thurs but fri eve already booked, sorry. Call if u need to talk betw now + thurs. Sxx*
ah	48	***Ah*** *but it could be a lot worse...*
ooh (32) / ooo (7) / oo (1)	40	***Ooh*** *goody jars! Ta. Maybe six? See you then at half past five-ish on weds?*
okey doke (17) / okey dokey (6) / okie doke (1) / okie dokie (6) / oky doky (1) / okie pokie (1)	32	***Okey doke****. I'm at home, but not dressed cos laying around ill! Speak to you later bout times and stuff. Xx*
mm / mmm / mmmm	22	***Mmm*** *- delightful image! Am home tonight so 8 sounds good - talk later*
er/erm/err	21	***Er*** *yep sure. Props?*
hmm/hmmm	18	***Hmm*** *ok, i'll stay for like an hour cos my eye is really sore!*
oops / ooops	14	***Oops*** *- am at my mum's in somerset... Bit far! Back tomo, see you soon x*
wow	8	***Wow*** *didn't think it was that common. I take it all back ur not a freak! Unless u chop it off:-) xxx*
aah	10	***Aah*** *i see! You should try watching lost wiv em!*
aw/aww	5	***Aw*** *thats excellent, well i hope you have a wicked time and see you sometime next week. x x*
yay	4	***yay****! finally lol. i missed our cinema trip last week :-(*
phew	3	***Phew****. Hope you didn't let on that she shouldn't try to do anything with it.*
Total no of minimal response tokens	**1663**	

Minimal responses in CorTxt resemble spoken utterances. As in speech, texted minimal responses occur as full ‘turns’, as in speech. There are around 60 occurrences, including these examples.

Figure 8.41 Texted minimal responses occurring as full turns

Ah!	Ok mate	Yeah that works too
Mmmm.	Ok Thankyou	Yeah well done! X
Nah	Ok that is cool x	Yeah whatever lol!
No choice	Ok x	Yeah, no problem.
No idea	Ok, will do	Yeah. Yeah
No its not that bad!	okay	Yep
No longer!	okay thanks	Yes
No thanks	Okey doke	Yes fine
No worries!	Okey doke, again	Yes indeed ta
Nope.	Phew!	Yes please
Ok	Yay!	Yes that's right
Ok :(	Yea	YES!
Ok bud	Yeah	Yes, true. Ok.
Ok dont worry	Yeah cheers	Yup
Ok great	yeah I know! xxx	Yup cool mate.
Ok good idea	Yeah mate.	wow wow wow!
Ok hon	Yeah not a prob	Yeah that works too

However, as the examples above suggest, response tokens are more often followed by further comments or much fuller 'turns':

(8.199) Okey doke. I'm at home, but not dressed cos laying around ill! Speak to you later bout times and stuff. Xx

(8.200) Oops - am at my mum's in somerset... Bit far! Back tomo, see you soon x

(8.201) Wow didn't think it was that common. I take it all back ur not a freak! Unless u chop it off:-) xxx

As in speech, and as in the examples below, minimal response tokens in texting frequently occur in clusters.

(8.202) Okay, good, no problem, and thanx!

(8.203) Nope thats fine. I might have a nap tho! S

(8.204) Yeah no probs - last night is obviously catching up with you... Speak soon xx

(8.205) Er yep sure. Props?

(8.206) yay! finally lol. i missed our cinema trip last week :-(

Other possible response tokens include *er* and *hm*. The inclusion of *er* and *hm* might seem inappropriate, but investigation of the examples suggests that these can often be considered responses to previous turns rather than discourse markers prefacing responses. In some of the occurrences below, *er* clusters with other response tokens (*Er yeah*); in others, it is given in response to interlocutors' questions, before an answer is given.

(8.207) Er yeah, i will b there at 15:26, sorry! Just tell me which pub/cafe to sit in and come wen u can

(8.208) Erm. I thought the contract ran out the4th of october.

(8.209) Erm... Woodland avenue somewhere. Do you get the parish magazine, his telephone number will be in there.

(8.210) Err... Cud do. I'm going to NAME281's at 8pm. I haven't got a way to contact him until then.

Sixteen examples of *hmm* were retrieved.

(8.211) Hmm well, night night x

(8.212) Hmm, too many of them unfortunately... Pics obviously arent hot cakes. Its kinda fun tho

(8.213) Hmm. Shall i bring a bottle of wine to keep us amused? Just joking! I'll still bring a bottle. Red or white? See you tomorrow. X

The remaining non-minimal response tokens are listed below, in order of frequency.

Table 8.4 Non-minimal response tokens

Non-minimal response token	**Freq.**	**Examples**
thank u / thank you; thanyou/ thanks / thanx	164	***Thank you*** *v much my far away friend! It is a beautiful sunny day, i am just about to treat myself to something new and i am looking forward to the rest of the month! See you soon! Xxx*
sorry / sos (4) / soz (1)	141	***Sorry*** *babes x im drunk i did not mean to bother u x x*
good	126	***Good****. Should be on your way by now. Have a great time. Speak soon. Xxx*
cool	105	***Cool cool****. Make sure u do! Is remedis free entry? Xxx*
cheers	40	***Cheers*** *babe. Have a good one yourself. Ps have you got our coat and gloves*
well done	28	***True*** *2ish is fine if it's ok with you.*
sure	19	***Sure****! Thank you x*
ta	19	***Ta*** *and see ya!*
gosh	17	***Gosh*** *that was an exciting message, i've just finished ASDA now heading home for some cheese and crackers, having an early night tonight me thinks, enjoy work.*
excellent	16	***Excellent****! I'm having my hair done in the afternoon so cud come round after that. Xx*
of course (not)	16	***Of course*** *and he was really scared!! Xx*
shit	16	***Shit*** *that is really shocking and scary, cant imagine for a second. Def up for night out. Do u think there is somewhere i could crash for night, save on taxi?*
will do	16	***Will do****. Was exhausted on train this morning. Too much wine and pie. You sleep well too x*
sweet	15	***Sweet****. Where to.*
sound	14	***Sound****! C u tom.*
god	13	***God*** *you don't even know your own children's names, and we're expected to remember Mothers Day.*
don’t worry	12	***Don't worry****, is easy once have ingredients! Hope you had good eve + see you sat. Sxx*
fine	12	***Fine****! Perhaps we could combine it with doing lunch or something. Anyway, have a fab christmas and a great new year xx*
right	12	***Right*** *fine, well at least i know where i stand. i wont bother anymore! i really like you but clearly this is not what you want.*
fab	11	***Fab****! I'll try and control the tears %-), may take week off - have wedding w/e b4.*

don’t know/no/dunno	10	***Don't know*** *i'm not there. Like i said i was in birmingham but now getting nearer to manchester. We're doing the ucas fair thing. Home very late wednesday night. And you*
great	10	***Great****! Lovely 2 c u, have a good trip back x*
what?	9	***What****? All of it or not?*
and you/u	8	***And you****. See you mon. X s*
I don’t know/no	8	***I don't know*** *then. NAME79 hasn't long left NAME281's and they're both dirty!*
indeed / indeedy	8	***Indeed*** *and by the way it was either or - not both!*
bless sb	7	***Bless ye****. Thank ye.*
definitely/defintately/defo	7	***Definitely****. Have fun studying (well try anyway!) X*
done	7	***Done*** *mate. did it last week*
true	7	***True*** *2ish is fine if it's ok with you.*
absolutely	6	***Absolutely****. See you there.*
blimey	6	***Blimey*** *u must hav spent some money! I cant go to the talk, stuff has piled up that i must get done tonight,eg apply to that job. Sounds good tho, our mums r going! x*
bloody hell	6	***Bloody hell*** *well done. I will ring you later to wake you up*
can do	6	***Can do****. What time?*
tit	6	***Tit****! Have a nice day xxx*
cheeky	5	***Cheeky****. U have to activate it for it to work abroad, i didn't realise. Okie, c u soon x*
I’m fine	5	***I’m fine*** *now, took quite a while tho. Sorry again, see you at 10 then. NAME14 x*
fair enough	4	***Fair enough****. Plus we'll prob only go to pubs, not clubs.*
go on then	4	***Go on then*** *wot on earth was it?*
I’m good	4	***Im good*** *really busy at work. Just chilling this wkend. Say hi to everyone else from me. We have to meet up soon*
kind of / kinda	4	***Kinda****. First one gets in at twelve! Aah. Speak tomo xx*
never mind	4	***Never mind****, it was short notice anyway! I have to get the last bus home... Happy new year to you anyway! Are you well? X*
easy	3	***Easy*** *mate, guess the quick drink was bit ambitious... Now on train and well up 4 weekend in london. Have a good one, if u manage to show yr face*
groovy	3	***Groovy****!*
hell yeah	3	***Hell yeah****. Pen it in girl*
hope so	3	***Hope so****, they did notice at least! Its nice2hav em back, at the mo anyway, i'l keep u updated! I hav looked thru*

		sum of the vast amounts of porn they bought!xx
nice	3	***Nice**!*
pardon	3	***Pardon**?*
same to you/u	3	***same to you** sweet we hope to see you in 2006!*
whatever	3	***Whatever**! Its ok were safe... I'm not driving. R u driving home christmas eve*
wicked	3	***Wicked**! So exciting! Was it 590? Cant wait 2 c u already! Xxx*
Total no of non-minimal response tokens	980	

As in speech, non-minimal responses can also occur alone in a text message: in fact, this appears to occur more frequently than with the minimal response tokens. *Thank you* or *thanks*, for example, occurs 17 times by itself in a text message, if we include *Thank you sweetie x* and *thank u by the way x, thanks mate*, *Thanks NAME112*, and *Thanks nellie*. Frequently, however, the non-minimal response tokens occur as part of longer text messages, as seen in the above examples, and also cluster with other tokens.

A final type of response tokens are here termed 'frames', in that they do not constitute fixed phrases but show lexical or grammatical variation within a basic structure. Examples are shown below.

Table 8.5 Response token frames

Response token frame	**Elements inserted into frame**	**Frequency**
sounds …	alright cool (3) delightful excellent fab (3) fair enough fine to me (2) god damn near perfect good (to me [3]) (21) gr8 (1) / great (3) horrendous like (9) a cool party a good idea a nice chilled outbreak a plan (4) cutting edge xx	53

	you've had a fun day ominous wise	
not …	2 bad (2) a bad idea (2) a lot (2) a prob a problem at the moment (2) at those rates doin 2 bad easily good (2) guilty hugely much (4) really (7) sure (13) to worry (3) today / tomo / tomorrow / tonight (5) very good with me you won't yet (8)	59
that would be	a no great (3) lovely(,) thank you (2) much better your area	8
that sounds …	about right fine good great like a plan (2) lovely really good terrible	9

that's	a bit weird a fab idea a no then? a shame alright amazingly lucky better (2) cool (5) d thing [= the thing] disgusting easier fantastic news fine (5) funny good (3) great gud idea it mad me (2) naughty not funny not he question ok (7) probably not a true synopsis of events quite a revelation really good of you ridiculous so mean so unfair the best bit of translating i've ever seen the trouble true (3) what I thought but …	54
that's	good news quite sad so depressing so true the decision that I was helping you to make!	6
what a …	bad mummy I am a bummer a good match a honey you are a joke a nuisance a pair (2) a shit a surprise	12

	a twat (2)	
glad	flat sorted funeral went well it went well (2) to hear to see (2) u got them ok u had fun (2) u like it wedding was good you could come you had a good time you had a lovely time you had fun you liked it you liked them	18
silly …	billy girl (3) question you	7
Total number		226

As these phrases suggest, the distinction between response and response token is a gradual cline rather than a clear dichotomy, and decisions made between what constitutes a clausal reply and what is rather acknowledgement of interlocutors' comments are to some extent arbitrary and subjective.

Premodification of the response tokens occurs, as in speech, but the only premodifier identified was *very*.[1] Ten examples of *v* or *very* were found (including *velly good*, presumably an attempt to mimic an accent):

(8.214) V nice! Off 2 sheffield tom 2 air my opinions on categories 2 b used 2 measure ethnicity in next census. Busy transcribing. :-)

(8.215) Velly good, yes please!

(8.216) Very strange. NAME33 and NAME209 are watching the 2nd one now but i'm in bed. Sweet dreams, miss u xx

[1] Other intensifying adverbs suggested by Carter and McCarthy include *jolly* and *most*.

8.4.9 Discourse markers

8.4.9.1 Definition

Discourse markers can be defined both by form and by function. With respect to form, they comprise various words including *anyway*, *cos*, *fine*, *good*, *great*, *like*, *now*, *oh*, *okay*, *right*, *so*, *well*; as well as phrasal or clausal items such as *you know*, *I mean*, *as I say*, *for a start*, *mind you* (Carter and McCarthy 2006: 209-211). Schiffrin (1987), who explores the extent to which the semantic meaning of such words and phrases shapes their use as discourse markers, looks at *oh* and *well*, *and*, *or* and *but*, *so*, *because*, *now* and *then*, *y'know* and *I mean*. In their role as discourse markers, the above can be distinguished from their non-discursive functions in that they occur outside the clause (Schiffrin 1987: 37-40; Carter and McCarthy 2006: 208). As non-clausal elements, they cannot easily be classified in terms of word-class but 'are best considered as a class in their own right' (Carter and McCarthy 2006: 209; see also Schiffrin 1987). This class is ultimately one defined primarily by function: as Schiffrin (1987: 41) suggests, discourse markers form a 'functional class of verbal (and non-verbal) devices which provide contextual coordinates for ongoing talk'. According to Carter and McCarthy, discourse markers 'link segments of the discourse to one another in ways which reflect choices of monitoring, organisation and management'. This includes such functions as opening or closing interactions, sequencing interaction and marking topic shifts, and monitoring shared knowledge. In so doing, however, as Carter and McCarthy make clear, discourse markers also play a more interpersonal role in indicating formality and power relations between speakers and highlighting their evaluations of the interaction.

8.4.9.2 Discourse markers in CorTxt

Investigation of speech-like discourse markers in CorTxt reveals, firstly, that they do occur in texting; and secondly, that they are involved in organising discourse and marking topic boundaries as well as focusing attention and diverting, but not so much in sequencing text or in monitoring discourse. The relative lack of sequencers (not explored below) can be explained by the length and complexity of text messages, but the infrequency with which monitoring discourse markers such as *you know* and *you see* occur is interesting, given the highly interpersonal nature of texting, and is outlined below. This is followed by investigation of *oh*, a discourse marker which organises discourse, marks boundaries and focuses or diverts attention, and occurs frequently in CorTxt.

8.4.9.3 you know *and* you see*: monitoring discourse in CorTxt*

Markers identified in speech as monitoring listeners' understanding and involvement are found only infrequently in texting. According to Carter and McCarthy, shared knowledge is commonly monitored in speech by (*you*) *see*, which 'projects the assumption that the listener may not have the same state of knowledge as the speaker', and by y*ou know*, which 'projects the assumption that knowledge is shared or that assertions are uncontroversial, and reinforces common points of reference, or checks that the listener is following what is being said' (Carter and McCarthy 2006: 221). Spoken examples include these.

Figure 8.42 Occurrences of *see/you see/you know* in speech

You see, since I've damaged my back in that fall, I find it difficult to climb the stairs without help.

You do it like this. Cut the branches right back, see, then cut them into smaller pieces.

If you got the earliest train in the morning and then, you know, like, got the last train back at night, it might be cheaper that way.

In CorTxt, however, the phrase *you see* occurs as a discourse marker just 3 times.[1]

(8.217) Ah you see. You have to be in the lingo. I will let you know wot on earth it is when NAME50 has finished making it!

(8.218) Its normally hot mail. Com you see!

(8.219) Hiya NAME196, have u been paying money into my account? If so, thanks. Got a pleasant surprise when i checked my balance -u c, i don't get statements 4 that acc

Nor are there many examples of *you know* as a discourse marker.[2] Three occurrences of *you know* (or *u no*) occur as monitoring discourse markers, one as part of the phrase *You know what i mean*:

[1] This can be contrasted with the very frequent occurrence of *see you* (521) in texting (explored further in Chapter 7 and below in section 7.4.10 on Greetings and Sign Offs). In illustration of the difference between speech and texting, *you see* occurs 28,925 times in the Bank of English (accessed October 2004), compared to only 7985 *see you*'s.

[2] Again, this can be constrasted with occurrences of another phrase, this time *let you know*, which occurs 116 times in CorTxt (Chapter 7). As with *see you*, this phrase is involved in management of texters' wider interaction which stretches beyond texting and which perhaps indicates concern with future arrangements.

(8.220) Oh yes I can speak txt 2 u no! Hmm. Did u get NAME12s email?

(8.221) Well. You know what i mean. Texting

(8.222) You know, wot people wear. T shirts, jumpers, hat, belt, is all we know. We r at Cribbs

In summary, the infrequency of both *you see* and *you know* is perhaps unsurprising given the constraints on time and space in texting. However, given that interpersonal considerations often override any need to be brief, and contrasting the infrequency of these discourse markers with the frequent use of *see you* and *let you know* (see Chapter 7), it is interesting that these phrases do not occur more frequently.

8.4.9.4 oh *in texting*

Oh, according to Biber et al (1999: 1083) the most frequently occurring interjection in speech, occurs 262 times in CorTxt.[1] Its role as a discourse marker is associated with its use as an interjection in expressing surprise, disappointment and other emotions in that it is used to respond to new or surprising information (Schiffrin 1987; Biber et al 1999; Carter and McCarthy 2006: 115) or, in Schiffrin's (1987: 74) words, when 'speakers shift their orientation to information'. According to Schiffrin (1987: 74), *oh* facilitates 'information state transitions' by creating a joint focus of attention and marking the saliency of information and interlocutors' shared knowledge in relation to it.

Oh in CorTxt can be categorised into categories suggested by Schiffrin (1987): in repair initiation and completion; to preface requests for elaboration, suddenly remembered questions and unanticipated answers; accompanying receipt and recall of information; and marking shifts in subjective orientation: all of which show language users responding or readjusting to information, old or new. As with other features of spoken grammar, their occurrence in texting highlights the importance of interpersonal considerations over physical constraints, and how speakers draw on spoken language to achieve this. It is also evident that, as in speech, *oh* clusters with other discourse markers to form phrases such as *oh yeah*, *oh right* and *oh my god*.

[1] Including 211 occurrences identified above as response tokens.

Figure 8.43 *oh* in texting

Initiating self-repair (Schiffrin 1987: 74)

I did get your emails... Were they about the party, and the rep? No more responses to invite - **oh**, except NAME335. We can hustle on friday. Sure NAME293 will come. Was that what you meant?

Oh what crap luck. You need to spend some time in a warmer clim- oh how lucky! Ta for translation, perfect x

Oh no i take it all back! They've just given me my presents! Chocs and a cool magnet of a big lizard that's there. Feel bad now4doubtin them! Take care! Xxx

Ok. Not much to do here though. H&M Friday, cant wait. Dunno wot the hell im gonna do for another 3 weeks! Become a slob- oh wait, already done that! X

No! Oh alright then.

Initiating repair of others

A: No still no sign of the card. and yeah there is someone running against me. hopefully in for a good competition now. bit sacred though

B: [text messages not in corpus]

A: Oh no. its POSTALADDRESS[1]

Completing repair made by others (*oh sorry*) (5 occurrences):

Oh sorry i forgot that you're not at school! Ok cool see you later x

Oh sorry didnt realise you had sent that to me too. Think dad replied for us all. Just watched the weirdest film. How was your nigh? X

Oh sorry i thought you were just being silly. You didn't tell of you were going. Its ok here still. Did you see NAME33 in liverpool?

Oh sorry! I didn't really read the message properly. My previous answer kinda applies to the first two question. As for the third: yes it is excellent. It doesn't even require the receiver thing. My computer found it easily. Thank you very much. How's 2006?

Oh sorry, hadn't thought of that.

Questions and answers[2]

Requests of elaboration (*oh gosh, oh why*) (2 occurrences)

Oh gosh what did he do this time?

[1] In this text message, A asks after his birthday card and learns that it has been sent to the wrong address.
[2] Spoken questions prefaced by *oh* are those that request information whilst also acknowledging or responding to old information. Requests for elaboration, for example, indicate that information has been received while soliciting new or further information.

Oh why? she asks innocently.

Suddenly remembered questions (*oh and* and *oh yeah and*)

Hiya can we make it 15 min later gotta in buy a tax disc! Oh and do you have a video camera?

Oh and are you and NAME324 coming to NAME117's housewarming on Saturday?

Morning NAME242, do you think we could have that money tomo, cheers ducks. Oh and do you fancy doing an hour of editing tomo evening if i can book a suite x x x

Hey bird. i'm gonna email you my very almost completed diss in a min. do what you can and give me a shout when you're returning it. thanks. oh and how are you anyway

Hiya, ru getting a train down on sat? Do u need me to pick u up? Any idea of time yet? Ru staying at mine? Just checking! Oh, + how is new job - any nice men?

I'm coming back on Thursday. Yay. Is it gonna be ok to get the money. Cheers. Oh yeah and how are you. Everything alright. Hows school. Or do you call it work now

Acknowledgement of unanticipated answers

Oh right ok then good job i dont need credit aint it lol

Oh are they? Sorry. Don't have your number. I phoned a friend instead!

Oh ok. I had visions of your coming here! But ok, see you then.

Oh okay. You can go jogging everyday

Oh right! Lol. Yeah send em to mum's. Cheers. Xx

Oh right, ok. I'll make sure that i do loads of work during the day! NAME281's got a really nasty cough today and is dry n shot so that should really help it!

Oh right. Things r ok just in college at the moment and working lata!! Fun! How abt u

Oh didn't think of that. For how long?

Oh, i thought it was weds

Recall of information (*oh* or *oh yeah*/*yes*)

Oh yeah i remember-happy shoppin! I am keepin NAME219 company in taunton while NAME281 buys her Xmas present. Hi to NAME62. X

```
 off v soon- c u pronto! Xx  HELLO! OH BUM I COMPLETELY FORGO
 u x  Yeah kinda went off on one... Oh forgot about the eggs on the w
 ed soon. What you up to  Oh fine.  Oh Fuck completely forgot about t
 forgotten quite how silly we were  Oh yeah i missed dad's birthday.
t will soon. cheers english accent  Oh yeah i did. Very funny. I'd forg
   shopping requests- be good! Xxx  Oh yes your party trick! Journey fi
```

Self-prompted recall (oh and)

Oh and by the way you do have more food in your fridge! Want to go out for a meal tonight? X

```
    m Corporation St. Are we still on? Oh and bring photos for extra ente
cently.   ... Are you in the pub?  ... Oh and bring wine or whatever you
 so will return to tropical paradise!  Oh and by the way you do have m
 i pop round to print my cv if u r in? Oh and happy valentines day!x  H
      ok- i will send a copy on monday oh and happy new year to you all!
  ant to go out for a meal tonight? X  Oh and i didn't get the funding!
 leep a bit before i go to work later. Oh and i did have a really good ni
 So much trivia, can't control glee... Oh and i own a pole dancing pole.
    progressively unwell as i watch .. Oh and i just cooked a rather nice
    the end of the year. thanks again. oh and i booked the thai place.  H
   ily anyway! Good luck for tomo ...  Oh and I'd like that conversation b
    tible with the mp3. Cheers bud. X  Oh and i'm not normally this slow
  girlfriend word for word please! Xx  Oh and if you get a camera, get o
  ot in. Cya.  Oh- and it gets worse!  Oh and thanks for the focus, with
      onto campus. Have a good lunch. Oh and well done england xx  I'll b
    edit for suggesting going dancing? Oh and you have a letter.  Not goo
     do coffee tomorrow as per normal? Oh just remembered you're probabl
  don't mind but cldn't get hold of u. Oh just remembered, ur meeting t
  ting... I want more.. Lots more ..x  Oh my god I just realised I'm goin
      elcome back to my address book (oh and a very happy new year!) lo
   will there be any pussy out tonite? Oh shit im not single- bugger.  Ye
```

Acknowledgments of the receipt of new information (*oh I see*).

Oh i see, well i won't come out at all then, unless i find out where NAME114 is! have a good night whatever you do!

```
   h I was out with you lot. Be good !  Oh- a new phone- how nice of san
    working and have just got in. Cya.  Oh- and it gets worse!  Oh and th
     K.  Ok.  Ok.  Ok. Are you driving? Oh ... Can you just tell me when a
ing base! Feeling ok, will ring later. Oh, NAME76 has just come back fr
     to check up on you. Sleep well x  Oh. Getting all excited i was. Enjo
n any full, but i had no notification, oh, hotmail's down 2. A web thing
  ll excited i was. Enjoy your thing!  Oh. I almost bought her somethin
  aybe i will ring NAME7...  Oh i am!  Oh i see i need to tell her by tomo
    need to tell her by tomorrow do y  Oh i see! I have just made a red c
    e! I have just made a red cabbage  Oh i see! I have just made red cab
    t cool at all. Keep me posted xxx  Oh blindz, yes. Have a good trip a
  l!  oh i think im having an orgasm!  Oh i want to play with the cat!i am
      She has web cam its quite funny  Oh. Is he there now? I spoke to hi
     too much will you? Take care x    Oh is that so, well if i do get the
       s seems daunting. Do you mind?  Oh my get you in the independent
   d time, hi to NAME43 and NAME2! Xx  Oh. My supervisor is on that. NAME
   ding out the back to surprise you.  Oh. the new students are here. ca
     er Xmas present. Hi to NAME75. X  Oh yeah they have lap belts. Just
       phone beeped mid class!  NAME4. Oh You're going to think i'm always
  h honey u will feel better soon x x  Oh honey, I so no that feeling - le
      will have 2 spend Mr STREET XXX  Oh hunney, i know the feelin! A dri
iit really b a hard winter this yr! x  Oh nice one, your lucky! um was t
  uld drive down with a coat for you!  Oh unintentionally not bad timing.
cool,id love 2 watch it!!silly NAME87!oh ur going 2 wear them in rems
   ou get it all done. see you soon x  Oh they've just decided that he's t
     u guys. Txt me wen u get there x  Oh god i'm sorry to hear that mate
```

Shift in subjective orientation[1]

One example is the intensification of position, expressed in CorTxt with such clusters as *oh dear, oh crap, oh well, oh my god, oh yes, oh no*. These can either respond to a previous text message or introduce new information.

Prefacing names and vocatives, to express, or appeal for, sympathy or apology:

```
number. I phoned a friend instead!  Oh babe I am sure he is missing
k, boobs out and u will be fine xx  Oh babe, he has probably had a f
s missing u 2 - I miss u loads xxx  Oh babe! You sound like uoy're ha
 E561? Does she like your present?  OH BABY !.. I'm working tonight a
planned!   I can bring alcohol! Xx  Oh babe, having a shit time at the

  i dont know. where do you fancy?  Oh honey u will feel better soon x
   eing you soon xx  Happy birthday oh gorgeous one. Sorry but your c

 ne who had a girlfriend and ended  Oh NAME43, i'm sorry but i don't th
 and working lata!! Fun! How abt u  Oh NAME14 just got message as be
anteed tear jerker! is spoons busy  Oh NAME654. The party was fab, the
't abandon his potential winner! Xx  oh NAME321! I'm in bed with a cold
```

Negative subjective orientations, including the expression of sympathy in phrases such as *oh dear* and *oh sorry*, of resignation in *oh well* and shock or surprise in *oh crumbs* or *oh my god*

```
 thats wicked! See you tonight! Xx  Oh crap - i'm out! Any day time sl
out! Any day time slots available?  Oh crap!  I've already got 1200 wo
tor comes on. I've written 50words. Oh crap! X  London is fine but cold
t it to be finished.  I could cry. Xx  Oh crap. Be on line five sixish and

 they said. Last class coming up.  Oh crumbs don't apologise - it's a
iscuss it with me! See you sat xx  Oh crumbs. Colour i suppose unle

 favourites, whatever they were xx  oh help! i dont know. where do yo

 pain!  Hello, how are you feeling? Oh dear i do indeed seem to be c
 ne you later - when is good time?  Oh dear - i fear for your head too.
           ousin your new outfit.  Oh dear i've forgotten who i am.  o
.  Oh dear i've forgotten who i am.  oh dear today is gonna get merry-
 ed to talk betw now + thurs. F xx  Oh dear, sounds weird. Why cant
    about you?  We are at grandmas. Oh dear, u still ill? I felt Shit t
cant you text it? Bit worried now.  Oh dear, what has he broken now
That was nice of dear old NAME124.  Oh dear-i'm sorry! It means somet
  h dear, what has he broken now?!  Oh dear. Hope they're better for M
   ee you Monday. Stay germ-free x  oh dear. It was a gud game thoug
  h dear. It was a gud game though  Oh dear. Well nice to confirm how
      oney. Very happy!  Who? What? Oh dear. Would like to go to a gig
  o confirm how necessary you are!  Oh dear. You ok? Thought I didn't
uite like Saturday night to myself, oh dear! Was thinking, do you fan
a get merry- we r drinking already!  Oh dear, poor u. Getting away for
   you like my newly found capitals? Oh dear bit giddy with all the trau
  ay, pick your cocktails wisely xx  Oh dear I think a vodka red bullmi

es, i'm not feeling too great either! Oh well at least you earnt twenty f
 g about him forgotten how this felt, oh well if nothing happens at least
    ur crazy text message collecting. Oh well its working. Thanks for th
```

[1] As well as managing information transitions and reorientations, *oh* can also preface the evaluation of information or 'shifts in subjective orientation' (Schiffrin 1987: 95).

```
      long! Will give fifteen min warning.  Oh well never mind  oh well never
          min warning.  Oh well never mind  oh well nevermind. This weather is
           ing Sunday and no work tomorrow  oh well that's good news yours sh
         . Do you know which school it is?  Oh well! I doubt he'll go to any.Th
       this Schnapps tastes like shit!  ;-(  oh well, c u later xxx  ;)  ;-), ju
      of it packing. I'm fed up of packing  Oh well, give us somat diff 2 pick
        years experience- i cant beat that! Oh well, has its advs that im not o
    ive us somat diff 2 pick apart fri :-)  Oh well. Figgy pudding all round t
        well. Figgy pudding all round then  Oh well. There's always the new r
          go on the rides but no-one dares. oh well.  Yep, at six by the bull!

     out that. I'll let you know later. Xx  Oh god already regretting coming
       omo then. I'll even buy you a drink  Oh god sorry we are eating curry i

          trictly come dancing on Saturday  Oh my god i'm being taken to cou
          my god i'm being taken to court.  Oh my god it's a new mike and thi
            d babes:)hannah x x  Oh my god  Oh my god i am so utterly falling f
     god i am so utterly falling for him:)  Oh my god i have a timer too! Ok i
      , a bit of gentle persuasion and all  Oh my god you must be kiddin yo
         t!!!lookin good babes:)NAME69 x x  Oh my god  Oh my god i am so ut
        t called she thinks she's pregnant. Oh my god!  see you then, we're a
    dal. I thought I'd had a difficult day  Oh my god! NAME92s gonna win then
           my god! NAME92s gonna win then…  OH MY GOD! WHERE ARE YOU
          ight for two weeks,how exciting!x  Oh my god, there are no taxis goi
                    ARE YOU? WHY? WITH WHO?  Oh my god, that pom documentar
         are no taxis goin to have 2 walk.  Oh my god, this is awful! i hate it
         y the bouncers! hope work is ok x  Oh my god. You were right. It is v
         ry pink. It looks awesome though.  Oh my goodness, how much long
          t the job anyway! How you doing?  Oh my, b'ham is looking u white g

        n the way that I have chosen to. X  Oh i'm so sorry ... It must be a te
        fore xmas ... With NAME145 too ..?  Oh i'm sorry sweetheart. It must b
         5 and 8.30? ice age looks wicked!  Oh shit sorry dude. See ya later
       !  Oh shit sorry dude. See ya later  Oh sorry didnt realise you had se
          Did you see NAME33 in liverpool?  Oh sorry to hear about horrible per
          hose! No problem. See you later.  Oh, sorry, didn't realise from your
            ank you very much. How's 2006?  Oh sorry. No I havent.Is it any goo

        set. Thanks. Good luck for result.  Oh that's crap of her. She said sh
          or is on that. NAME43 SURNAME11.  Oh. That's strange. Perhaps he's
         corner. Sorry that sounds lame. X  Oh thats good then, at least you k
        ll. There's always the new ranges.  Oh what a pain 4 u!  Oh what crap
          new ranges.  Oh what a pain 4 u!  Oh what crap luck. You need to s

          you get back. Hi to emma. Bye! X  Oh blow!  Oh bugger! That all sou
          ck. Hi to emma. Bye! X  Oh blow!  Oh bugger! That all soundsa bit cr
     ell, I suppose it is her birthday! Xx  Oh bugger.  Oh cheers thats wick
           though. see you tomorrow kid. x  Oh poop about sit jobs. It sounds
          kay. You can go jogging everyday  Oh poo. Never mind i'll have the re
       . It's so crap. Good luck though. X  Oh poor baby x never drink with a
       ount of the girls who tried hee hee  oh poor you!think of the money.!
         t lunchtime then! See you then xx  Oh you meanie. Pop in for a drink

          far to much work to do! hows you  Oh no don't cry - drinking wine 4 u
        ine 4 u again in support of u! Xxx  Oh no i am pulling a funny face!
       e of wine and 5 bottles of breezer!  Oh no I've just found the bridgwate
           ow4doubtin them! Take care! Xxx  Oh no im on second bottle of wine
        t found the bridgwater one. Sorry.  Oh no never in blue jeans might b
         enders - what are we going to do?  Oh no! Damn it...what we gunna d
         ow. Tea and weetabix mini s rule!  Oh no! Have found flaw in essay a
        too long n doesn't flow as well...  Oh no! Well you'll get it back in 2
        ing up x  Oh no. its POSTALADDRESS  Oh no. what are you gonna do.  O
```

Wishes or desires

Excellent still drunk on Tuesday from Sunday, oh I miss those days with him!! Tell him I got a new phone so my best friend can be kept in the loop!! NAME192 is away working and NAME289 is here as he has lost his door key and he has no where else to stay, haven't seen him 4 ages and he's quite cute but beer goggles mite b helping that!? X x

Oh I so hope so bit bizarre really but think he is really great. Quite surprised I think that but it's good 2 learn new things about yourself! Glad we r cool was a bit worried I had been replaced by NAME156! Xxx

Well, i envy you greatly as i am on my tod with a glass of wine watching boules waiting for everyone to turn up! Oh for hot choc and duvet.. How sad i look! See you soon and sleep well! Xx

Positive orientations

```
   ing crossed and waving in the air!  Oh yeah i did. thank you. haven't
    7px. Except i can't b 29 anymore.  Oh yeah baby! I have everything cr
     Cos my phone is better than yours oh yer  Could do balti. Twice in on
     s and texted, gawd bless him xxx  Oh yes I can speak txt 2 u no! H
      hen home. Have a good evening x  Oh yes! Wheres chuckles?  Oh y
ll as the funny ones? It is funny ... Oh yeah know what you mean. S
      ing x  Oh yes! Wheres chuckles?  Oh yes, came with a vengeance.
     enlist mums help later though… X  Oh yes, i'm carrying on the greenl
     issing our drinking sessions Xxx  Oh yes, why is it like torture watc
. i'll try them later, see ya soon x. oh yeah and yes there was snow
       bout thomas moore! Speak soon.  Oh yes and texted, gawd bless hi
     at so i'm happy. See you soon.xx  Oh yes all fine here- just been to

u guys get suits ill be your devil...  Oh fab. It's just become my favour
    as NAME188 should be back by then. Xx  Oh fantastic. That is the best
     a go to bed soon. What you up to  Oh fine.  Oh Fuck completely forg
       e between genius and madness is oh so very fine.  NAME32, i emailed u
    finished new look. ASDA at 4 til 8 oh the joys!hows the college work

   asleep to another jim carrey film.  Oh good. er..remind me what post
    bout lessons and getting home ...  Oh good film. im watching law and
     h gosh what did he do this time?  Oh great- so you can come up on
      t way … Who are you going with?  Oh, great NAME242, you're a star!

   it is her birthday! Xx  Oh bugger.  Oh cheers thats wicked! See you

    one. Followed by a lot of marking. Oh how I love my job !  Well actua
  k hes goin to go far;hes a natural!  oh i think im having an orgasm!  O
     t  Oh no. what are you gonna do.  Oh now, quick!  Oh ok. I had visio
  temptin? Maybe i will ring NAME5...  Oh i am!  Oh i see i need to tell h
     me up though and i want to sleep  Oh i will. Can't get enough of sitt

 ish i cud say i went 2 the lecture,...oh! glad u txt,last nite was fun!im
```

Discrepancies between texters' perspectives (Schiffrin 1987: 97).

Oh details details. Well, i shall give up on the match marking then. Watch out for hte exdrummer mind - by definition they don't drum any more.

Oh don't worry its so not worth it x

Oh hard to say. Neither as i wish but hey that's life. I'm on the allotment enjoying sun. Any plans tonight?

Oh he knew you were there... Very jealous. Hope he did all our favourites, whatever they were xx

Oh sort of pleasure. Going to c sister outlaw in somerset. Thanks. Good luck for result.

oh I think hes goin to go far;hes a natural!

Oh you have to be up very early to catch me hungover ... Mm but you know that's not true ... Anyway great - we can start drinking at lunchtime then! See you then xx

Oh erm revisit and rehash last years h&h costume maybe rauch it up a bit bring out the brown shorts again x and if you guys get suits ill be your devil...

Oh I'm farting about looking at phones. Lol. Sunday night I'll be stressed. Think I know what I'm doing and have reason to do it in the way that I have chosen to. X

Oh scary movie 4 and silent hills stary at 8.15 and 8.30? ice age looks wicked![1]

As the cognitive task fulfilled by *oh* in online, spontaneous speech is not one relevant to texting,[2] the suggestion here is that *oh* is used in texting for pragmatic effect: that is, texters attempt to recreate the pragmatic effects of *oh* on interlocutors in spoken interaction. In speech, *oh* marks 'the division of conversational labor between speaker and hearer' and 'speaker/hearer alignment toward each other' by making explicit their orientation towards information (Schiffrin 1987: 99-100). Its occurrence in texting, where texters have little cognitive need to mark information transition points, suggests that texters draw on the interpretation of *oh* in speech to recreate the conversational, informal sharing of information. This picture of *oh* as an indicator of stance within intimate relationships is reinforced by its frequent occurrence as a preface to negative subjective orientations or reorientations.

8.4.10 Greetings and sign-offs

Reading through CorTxt suggests that one other category of often non-clausal elements plays a potentially larger role in texting than in speech: greetings and/or vocatives and closings. Over one tenth of text messages in CorTxt start with greetings, including the below.[3]

Table 8.6 Frequent greetings in CorTxt

Hi	508 occurrences
Hey	477
Hello	366

[1] In the final three, it appears as though the texter suspected that their reply would not meet their interlocutor's expectations.
[2] Using *oh* to manage their constant redistribution of knowledge and the readjustment of the relative saliency of items of information may, suggests Schiffrin, provide speakers with time to focus on informational (rather than cognitive) tasks.
[3] Greetings were identified by sorting the text messages in alphabetical order, as with response tokens, and focusing on the text message-initial elements.

Hiya	103
Heya	33
Howdy	6
Hero	3
Total	1496

Of these, *hey* is particularly likely to be followed by a vocative. These include:

Table 8.7 *Hey* + vocative

Hey dude	32
Hey babe/babes/baby	30
Hey buddy	21
Hey honey	14
Hey hun	11
Hey gorgeous	8
Hey chief	8

Both *hey* and *hi* are frequently followed by an interlocutor's name: 70 and 115 occurrences, respectively.

Very few text messages (28) begin with enquiries into a interlocutor's well-being, including:

(8.223) How's it going? Got any exciting karaoke type activities planned? I'm debating whether to play football this eve. Feeling lazy though.

(8.224) hows my favourite person today? r u workin hard? couldn't sleep again last nite nearly rang u at 4.30

(8.225) Hows the champ just leaving glasgow!

The possible significance of closing sequences was suggested by the frequency of phrases such as see you soon described in Chapter 7. Closing sequences in texting include *see you + time expression* and the symbol *xx*. *See you* is spelt both conventionally (521 times) and

as the homophones *c u* (182), with 99 occurrences of *see u* (although no occurrences of *c you*). *See ya* and *c ya* occur 134 and 6 times, respectively. Taking all these spellings into account, there are 942 occurrences.

Figure 8.44 *See you/c u* + time expression[1]

See you / c u etc	soon	164
	later	105
	then	59
	there	60
	tomorrow	48
	tomo	44
	next …	26
	Monday	14
	Friday	14
	Sat	13
	again	8
	when …	7

Interestingly, *bye* and *goodbye* occur infrequently, with 9 and 3 occurrences respectively. Despite the need for brevity, texters prefer *see you* or *speak to you*. This may be because *see you* serves to soften the bluntness of exchanges which must presumably be concluded more abruptly than in face-to-face or telephone encounters. It also reflects the role which texting plays in co-ordinating activities and confirming future arrangements. These findings unsurprisingly reflect those of Chapter 7, and allow them to fit into a wider description of texted grammar.

8.4.11 Summary

Using Carter and McCarthy's (2006) description of spoken grammar, the above investigation of texted grammar explored clause combination, situational ellipsis, deixis, headers and tails, vague language, response markers, fixed tags, and discourse markers. Two observations emerge: firstly, that non-clausal elements of texted grammar can be

[1] Any discrepancies between these figures and those cited for 3-wordform phrases in Chapter 7 are due to the fact that these encompass spelling groups, and those in the previous chapter did not.

accounted for by reference to, and in contrast with, a model of spoken grammar; secondly, that there are differences between spoken and texted grammar.

Clauses, for example, are frequently combined in a way reminiscent of speech in that text messages contain a limited number of two- to eight-clause units joined in paratactic relationships. Other text messages, however, comprise several units with fewer clauses combined by hypotaxis (chiefly *that* and *to* clauses) along with somewhat disjointed connections between units that are not reflective of spoken language. Clause combination in texting, then, is sometimes but not always speech-like, but what it tends not to include are embedded clauses typical of formal written language. Another example, within the clause, is the substantial amount of situational ellipsis. Ellipted elements not only include those also ellipted in speech, but also those more typical of written language: medial elements as in *Wine good*, and the use of *is* rather than contracted spoken forms such as *he's*, in: *Is away a lot.*

Other differences emerge in relation to deictic reference. Determiners *this* and *that* are used predominantly when texting in making future arrangements as well as in reference to the places where these plans will take place, the 'virtual space' created through the medium, and to aspects of texters' shared background. As in speech, the determiners also play an evaluative role although, in contrast to Carter and McCarthy's (2006) observations, *this* in texted utterances such as *Damn this christmas thing* appears to be as (or perhaps more) critical or dismissive as those with *that*. Alongside response tokens which Carter and McCarthy identify in spoken interaction, particularly variants of *yeah*; *ok*/*okay*; *nah* and *no*, and *oh*; *thanks* or *thank you* and *sorry*, *good* and *cool*, we find tokens that are not typical of speech: representations of emotions, such as *lol* and *ha ha*, and emoticons. Also non-characteristic of speech is the fact that response tokens tend to occur not as full turns but as prefaces to lengthier text messages.[1] The frequent use of frames such as *sounds alright*/*cool*/*fab* and *that's a bit weird*/*cool*/*good* highlight an apparent need in texting to be more explicit than in speech, as does the use of lengthy fixed tags such as *would that be ok?*.[2]

[1] Or in the middle rather than the beginning of messages, although this was not the main focus of the current investigation.
[2] Although raising questions as to the distinction between tokens and full responses,

Vague language of a speech-like nature is also found in texting, perhaps surprisingly given their often lengthy expression, although some items frequent in speech, chiefly *and things* and *and stuff*, do not occur and others, such as *and that* and *like*, are fairly infrequent. Discourse markers typical of speech also occur and, although there is little indication that texters explicitly monitor interlocutors' involvement, speech-like markers illustrated by *oh* are used to show texters' orientation to information. Finally, greetings, vocatives, and sign-offs also play organisational roles outside the traditional clausal structure, as could be expected from the turn-taking mechanism of texting.

As in previous chapters, these findings highlight the importance of the interpersonal over the constraints of the medium, and suggest that texters often draw on the resources of spoken language in creating speech-like informality.

8.5 Chapter Summary

The realisation that spoken grammar is distinct from that of written texts has led to a questioning of traditional grammatical categories and development of analytic models attempting (in differing ways) to incorporate spoken as well as written data. These range from Halliday's functional-systemic analysis to Brazil's (1995) linear grammar of speech, as well as corpus descriptions of spoken grammar (Biber et al 1999, Carter and McCarthy 2006), all of which recognise that grammar is shaped in different communicative situations by differing constraints on its production and the social functions it fulfils; and that grammatical features be analysed with reference to real data. Despite insights offered by analytic models, limitations when applying them to naturally-occurring, interactive speech suggest a need for a descriptive model grounded in corpus investigation of naturally-occurring speech.

CorText was investigated for features found to be typical of spoken grammar by Carter and McCarthy (2006). While based on comparison with spoken grammar, the findings nonetheless suggest texted grammar can begin to be described as a grammar in its own right. Its features can in part be explained in reference to the situational circumstances of texting, in which participants involved in informal and intimate communication face physical constraints of the texted medium. For example, text messages are visual without recourse to prosodic features, gestures or expressions, and thus dependent on the written

language to convey meaning and fulfil otherwise speech-like communicative purposes: hence the evaluative use of *this* and *that*, emoticons and representation of laughter seen in *lol* and *ha ha*, and the sign-off, *x*; while the fact that texters do not share physical space shapes their use of deixis. Other physical constraints, chiefly the limited character allowance and tiny keyboard and screen size, may similarly discourage lengthy speech-like features.

However, features identified in texting must also be explained in terms of the way the medium is exploited by texters in fulfilling functions to which they choose to put the technology. These functions are in many cases most typically associated with spoken interaction: text messages are, for example, private and sent between intimates (Ling and Yttri 2002) and use of speech-like grammatical features construes an appropriately informal arena. At other times, texts fulfil transactional functions such as requests or coordinating future arrangements.

In conclusion, texters have the time and opportunity more typically associated with writing to craft text messages designed to fulfil speech-like purposes: suggesting features identified in speech are likely to occur, but that these arise more from functions of texting than its constraints. Seen in this light, we can describe texters as actively and selectively drawing on spoken resources to create or perform speech-like informality and thereby heightening the sense of intimacy; while their responses to other aspects of the medium result in grammatical features distinct to texting. Chapter 9 argues that these can be explained by the concept of performativity, and in particular the performance of brevity, spoken informality and deviance. The implications of an account which credits users with active roles in shaping language and describes them as drawing on existing resources is that texted grammar should be described, not in comparison to speech nor as existing on a spectrum of written and spoken varieties, but as a language variety in its own right. Further research could therefore move away from comparison with speech, and explore texted grammar in its own right as shaped by the communicative situation in which it is produced

Although primarily driven not by frequency counts in the data itself but by an existing model, the investigation adds to the previous chapter by retrieving and interpreting features according to grammatical or discourse functions neglected in a phrasal approach. In so much as similar features arise in both studies, as well as in the investigation of everyday

creativity, the thesis highlights how the same features are interpreted in different ways through distinct approaches. However, given the similarities in conclusions reached, the studies reported in this thesis can also be seen as complementary.

.

CHAPTER NINE
CONCLUSION

9.1 Contributions made by this thesis

9.1.1 Introduction

This investigation into the language of texting contributes to the existing field of knowledge in three ways: firstly, it develops the methodology used in corpus investigation of language variation; secondly, it adds greatly to a description of texting as a language variety in its own right; and, thirdly, it proposes that one way of understanding the language of texting lies in the concept of *performativity*. This concept focuses attention away from technological constraints and highlights instead the agency and linguistic awareness of users who exploit the medium for their own purposes. These three areas are explored in greater depth below.

9.1.2 Methodological insights

The contributions made to methodology by my thesis are threefold, and include its discussion of issues relating to corpus compilation; its role as an example of corpus-driven investigation; and its application of the descriptive spoken grammar model outlined by Carter and McCarthy (2006). Taken together, these aspects pave the way for further research into texting and other forms of electronic communication and, more generally, for researchers engaging in corpus-driven investigations of other varieties.

Designing and compiling a text message corpus for linguistic analysis, as described in Chapter 4, raises considerations also addressed when compiling other corpora: issues surrounding corpus size and representativeness, and ethical considerations. However, the distinct nature of texting raises new challenges (not previously addressed in such depth by linguistic researchers), and so paves the way for future linguistic research into texting and other electronic forms of communication. What makes texting such a challenging data set is that it is highly personal and its language so unconventional. The fact that text messages are sent between intimates complicates data collection although, as this thesis shows, individuals are willing to submit text messages either online and anonymously or in person.

The benefits of the latter approach to data collection (adopted in this study), whereby participants are known to the researcher, is firstly the researchers' familiarity with participants' backgrounds and their texting practices; and secondly, that it gives the researcher control over issues of text message *selection* (whereby participants select interesting or unusual text messages for submission) and *transcription error* (whereby mistakes are made when transcribing text messages), issues which must nonetheless be acknowledged. This data-collection method also has beneficial implications for corpus representativeness, assuming that the best way for corpora to be deemed representative is to carefully define their parameters and to limit generalisations to the sample. As for corpus size, this is inevitably an issue with texts averaging 17.2 words, but the use of a corpus of less than 200,000 words can be justified pragmatically as well as by reference to the number of texts included (11,067) rather than the number of words; and by prioritising corpus stability over size. Finally, texting raises new challenges in regard to ethical issues, complicated already by the lack of ethical frameworks for linguistic researchers and the fact that guidelines must always be interpreted according to the particular dataset. Anonymisation, given the personal yet unconventional nature of texted language, will always be particularly difficult when compiling text message corpora, yet this thesis laid out a framework for reference by future researchers. The thesis also sought to justify use of a largely un-normalised corpus, according to Sinclair's 'clean-text' policy and the meaning inherent in respellings (Sebba 2007). Throughout discussion of these issues in chapter 4, what emerges is a complex and justifiable balance between *principles* (such as representativeness and protection of participants) on the one hand and *pragmatics* (what can be done given the nature of the data and the goals of the research) on the other, the need for which is also acknowledged in the general literature on corpus design (Sinclair 1991; Hunston 2002; McEnery et al 2006). My work therefore serves as guidelines which subsequent researchers of texting (and other forms of electronic data) can interpret and adapt to their data sets.

As well as illustrating how to compile a corpus of text messages, my research provides an example of corpus-driven investigation of a language variety. The strengths of a corpus-driven approach, as outlined in chapter 3, are that initial decisions as to what features to explore (that is, what is interesting in the particular set of data) are made according to the data itself, which in effect means that the features forming the focus of the ensuing investigation are those which are significantly frequent in the corpus. This *inductive*

approach, as many corpus linguists would argue (Sinclair 1991; Tognini-Bonelli 2001), ensures that language descriptions remain true to the data being explored and avoids the imposition of theoretical assumptions (beyond recognition of the *wordform* and reliance on frequency as an indicator of significance). The limitations of the corpus-driven approach are that it is difficult to produce either a framework or a point of comparison for analysis of other varieties, and analysis must instead proceed, as it were, on a case-by-case basis. The study outlined in Chapter 7 is therefore important in two ways. Firstly, it shows one way in which the course of the investigation of a variety can be determined by features arising from consideration of wordform frequency. In the present case of the text message corpus, CorTxt, the investigation proceeded from the observation that many of the frequencies of the most frequent wordforms are reversed from the order found in other corpora, and that *a* in particular is relatively significant in comparison to *the*. In other studies, avenues of research may alternatively emerge from consideration of keyword lists. Secondly, while the study in Chapter 7 illustrates the need for researchers to proceed flexibly and in response to what their data tells them, what it cannot do is set out a model which can be applied rigorously to other text types. In other words, the argument which this investigation supports is that there is a need for in-depth exploration of single language varieties alongside large, quantitative studies such as those employing Biber's multi-dimensional framework which, although enabling comparisons between varieties, makes presuppositions about the nature of a variety based on analysis of other varieties. As well as adopting an inductive approach, however, the wordform-frequency study is also corpus-driven in the other sense with which this term is often used, in that it assumes a *lexical grammar* model of language. As discussed further below, this study shows that unusual wordform frequencies (such as those of *a* and *the*) can most convincingly be explained with reference to the phrases in which the wordforms occur.

The third and final contribution which this study makes to methodology lies in its application of Carter and McCarthy's description of spoken grammar to another dataset, motivated by perceived inadequacies of traditional models in describing the structures used in CorTxt. To an extent, Carter and McCarthy's framework poses a challenge for corpus-driven investigation. Features such as heads, tails and situational ellipsis could not have been identified through frequency counts and concordances but only through close analysis of smaller texts: for this methodological reason, it is argued here that their approach is best described as *text*-driven rather than corpus-driven (Bednarek 2006). Their model also

challenges the principles of corpus-driven work because it posits new observations within a largely traditional grammatical structure and so highlights possible limitations of lexicogrammatical frameworks which, as pointed out by McEnery et al (2006), work solely on the lexical level and neglect other grammatical relations included in Carter and McCarthy's descriptive model. However, unlike other more analytic and radical models of speech which reject traditional models (Brazil 1995), the strength of Carter and McCarthy's model lies in the fact that it is driven by data, albeit in what is described here as a *text*-driven way that differs from the corpus approach outlined above. It can therefore be used (as in this study) alongside corpus-driven approaches to enrich or supplement findings.

A further contribution made by my study is that the productive application of the model to a set of non-spoken data (CorTxt) serves to highlight the utility of the model by revealing its relevance to the linguistic description of electronic communication. This means that the application of Carter and McCarthy's (2006) descriptive model can begin to highlight specific similarities and differences between spoken and, in this case, texted communication. This is significant, given the amount and significance of research comparing written and spoken language varieties (Biber 1988). However, the conceptualisation in many previous studies that these varieties lie along a cline from writing to speech (Baron 2000) should be challenged, not only because it focuses overly on the mode of production and neglects user-related aspects such as the functions to which the medium is put (hence Biber's multi-dimensional framework), a point I shall return to below, but because it presupposes a simplistic and comparative relationship between the situational factors and linguistic features of each language variety. Instead of claiming that speech-like features in texting render it 'a little like speech' or 'somewhere between speech and writing', a more effective conceptualisation is that users are drawing on their awareness of other discourses (in this case, the informality of spoken interaction) in order to fulfil certain functions through texting, and that this leads to the occurrence of 'speech-like' features in texting. This approach is suggested by McCarthy (1993) in his study of how advertisers draw on features of spoken language to produce informal adverts which appeal directly to readers, and its relevance to texting is returned to below. To return for the moment to methodological issues, Carter and McCarthy's (2006) model serves as a data-driven resource by which to explore and support assertions regarding the differences between the realisations of spoken and written grammar.

Through these three aspects (the new challenges raised by compiling a corpus; the role of the investigation in illustrating a corpus-driven approach; and the application of Carter and McCarthy's text-driven spoken grammar model), this thesis paves the way for future research into texting and other electronic varieties. It outlines and addresses various challenges raised by compilation of a small corpus of electronic data; and argues for two approaches into the data: firstly, a corpus-driven approach which does not impose pre-corpus categories; and, secondly, the careful application of findings of text-driven descriptions of spoken grammar which not only indicates the value of such an approach but also, as we shall see below, highlights the extent to which, and the ways in which, users draw on knowledge of spoken discourses to fulfil certain functions through the texting medium.

9.1.3 Description of texting

The contribution which my thesis makes towards the description of texting as a language variety is, again, threefold. Its achievements lie not only in its confirmation, through a large dataset, of assertions regarding variation in spelling and in texting styles; but also in the fact that my study goes beyond the abbreviations which form the basis of many investigations of Txt to explore other language features such as everyday spoken creativity and features of spoken grammar in texting. Finally, it also highlights the fact that texting cannot be defined solely in relation to speech; or to writing. Instead, the third aspect of my research is that texting must be approached as a variety in its own right.

9.1.3.1 Written and spoken creativity in texting

The argument that spelling variation in texting, as in other text types, carries social meaning rests on the assumption that the choice of variants is constrained by orthographic principles which ensure that respellings are understood and interpreted in the way intended by writers (Sebba 2007; Shortis 2007a,b). The implication of this argument, outlined in Chapter 5, is that respellings in texting follow principles and therefore reflect and extend existing practices seen in pre-standardisation variation and early printing practices as well as modern advertising, graffiti and music fanzines. This has received a limited amount of textual support: Shortis (2007a,b) explores one or two examples, Hard af Segersteg (2002) notes the use of 'conventional' respellings in her corpus of 1152 text messages, and Kesseler and Bergs (2004) trace back the use of variants such as *bcoz* and *missd* in their

corpus of valentine text messages to love letters written by 'fallen' Victorian girls. My systematic investigation of a corpus of 11,067 text messages showed exactly how one large network of texters drew on the orthographic principles of English and how their respellings reflected and extended existing practices, and so serves to further support assertions made by other researchers. Respellings in my corpus which reflect spoken pronunciation include the colloquial contractions *goin*, *ad* and *ya*, the colloquial respellings (a category not used in other studies reviewed here) *nope* and *nah*, and the regiolectal spellings *dat* and *sumfing*. These co-occur with phonetic spellings (*u*, *r*, *tonite* and *cum*); abbreviations such as *stil*, *bac* and *tomo*; and the frequent omission of apostrophes. However, my study also serves to challenge or refine previous work based on smaller amounts of data: for example, not evident in my corpus are initial-letter sequences such as AAMOF (*as a matter of fact*) documented by Crystal (2008). Respellings that appear to extend existing trends and which are not covered in other accounts were found to include the omission of <h> in *wat*, *wen* and *wich*; clippings such as *tomo* and *tho*; double letter reductions such as in *stil* or *gona*; and letter homophones *ne* (*any*) and *ur* (*you're* and *your*). Although some examples differ from previous accounts, this study provides evidence as to the choices which texters have, sometimes in how they spell one word: *you*, for example, can be spelt phonetically as *u* or as the colloquial contractions *ye* or *ya*; and *what* occurs both as *wot* and as *wat*; while sequences such as *cumin* combine both phonetic spelling (as the <o> becomes <u>) and colloquial contraction (the omission of the final <g>). As such, this study supports the argument that texters have the potential to creatively and appropriately construe meaning through the spellings they select.

The other achievement of this thesis lies, however, in its identifying that creativity in texting is not limited to variation in spelling, but can be seen also in creative patterns reflecting spoken everyday creativity. The assertion made above that spelling in texting is creative is repeated across the literature (Crystal 2008; Shortis 2007a,b; Hard af Segersteg 2002; Kasesniemi and Rautianen 2002), but the source of creativity is each time limited to spelling, with the argument made that creative respelling reflects metalinguistic awareness, a sense of appropriacy, and playfulness. While studies into spoken discourse highlight the prevalence and significance of playful language use such as idiom manipulation, metaphor extension, new word coinages and structural and phonological repetition, even Carter's (2004) own discussion of webchat focuses largely on exploitation of graphic and

orthographic resources. My study, however, shows that everyday creativity of the kind identified in spoken language also occurs in text messages such as the following.

(9.1) Am watching house – very entertaining – am getting the whole hugh laurie thing – *even* with the stick – indeed *especially* with the stick.

(9.2) I've got some salt, you can rub it in my open wounds if you like!

Furthermore, everyday creativity is seen to play a significant role in allowing users to express themselves despite the restrictions of the medium. In the absence of other resources, such as prosody and tone, and without space for the lengthy expression possible in extended written prose, texters exploit and transfer spoken creativity across to texting in order to fulfil expressive, speech-like functions. In other words, everyday creativity is very appropriate for the texting medium.

The picture so far is, as suggested in the previous section, that texting draws on other discourses: on the one hand, it reflects and extends the respellings and abbreviations seen across contemporary and historical written texts and, on the other, it calls upon the creative resources used in spoken conversation. As stated previously, my argument is not that this reflects observations made by other researchers that texting like other forms of electronic communication sits on a cline between the typical realisations of speech and writing and as such mixes elements of the two but, instead, that it is a distinct variety which (as stated above) draws on texters' awareness of established discourses. This argument rests on the observation arising from my research that texting differs in significant ways from both written and spoken modes, suggesting that it may not be helpful to describe it simply as somewhere between speech and writing. The first indication of these differences can be found in the word frequency list described in Chapter 6. As mentioned in section 9.1.2, unlike frequency lists generated for general spoken and written corpora (Leech et al 2001), *the* is not the most frequent wordform in texting but is replaced by *you* and instead occurs with a similar frequency to *a*, which usually has a third of the occurrences of *the*. Further analysis suggests that the similar frequencies of *a* and *the* cannot be explained solely by the omission of *the* or by reference to the 'speech-like' nature of texting, but must instead be addressed with reference to the particular phrases in which *a* and *the* occur. As O'Keeffe et al (2007) suggest, the most frequent phrases in a corpus provide a 'fingerprint', in much the

same way as a wordform frequency list, and in texting we see the frequent occurrence of phrases such as *have a good time*, *give me a bell* and *a bit of* which differ from those found most frequently in spoken and written corpora and so not only explain the frequencies of *a* and *the* but also distinguish texting from both written and spoken texts.

A similar picture emerges through application of written and spoken grammars to texting. Although the application of Carter and McCarthy's descriptive model of spoken grammar reveals typically spoken features in CorTxt, it soon becomes evident that differences between the two exist. For example, while some ellipsis reflects that occurring in spoken language, texters also produce utterances such as: *Wine good* and *Is film by French director*; while many text messages are speech-like in their clausal combination, others combine clauses into short, disjointed sentences; and while there is deixis in texting, it largely relates to time rather than place.

My argument is that, as discussed in Chapter 2, too great a reliance on mode can reduce the issue to the question of whether a variety is speech-like or writing-like, whilst ignoring clines of other situational features found to shape language varieties. In other words, the focus on mode prioritises the impact of constraints of the medium and overlooks the influence of users, their relationships with interlocutors and the purposes to which the medium is put. Consideration of functions and users of the technology, on the other hand, allows us to perceive of the constraints as *affordances* exploited by users according to their capabilities. This highlights the agency of users, as well as their awareness of language use in drawing on other discourses. In the case of texting, what is physically a constrained medium is exploited by its users to achieve informal, intimate and expressive purposes through choices made in spelling and the use of everyday creative patterns, and which reveals itself in distinctive phrasings recurring across the corpus. Although comparison with 'speech' and 'writing' forms a convenient starting-point for analysis of a 'new' variety, the implication drawn from my thesis is that a variety such as texting should at least be considered along a multi-dimensional framework, rather than a cline between spoken and written modes, and ideally as a variety in its own right, whose users draw on their awareness of other discourses in construing it.

9.1.3.2 Further comparative study: texting and other varieties

Comparisons drawn between CorTxt and other varieties can add further credence to the argument outlined above that texting be considered a distinct variety. The analytical focus in this thesis has been primarily on comparisons of texting with general language, with reference to the BNC and, to a lesser extent, to the Bank of English. Other comparisons were made throughout between texted and spoken language, most notably in observations regarding differences in creative language use and between the grammatical features of CANCODE and CorTxt, but also in comparisons made between collocate lists of CorTxt and the spoken subcorpora of the BNC. More thorough and delicate comparisons with spoken corpora could, however, strengthen the argument. For example, in Tagg (2007), I report on differences between CorTxt and two corpora of written and spoken data: FLOB (comprising one million words of written English compiled in the 1990s); and COLT, a similarly-sized corpus of spoken English from the same decade.[1] Keyword comparison with FLOB reveals differences between CorTxt and the written corpus which would be expected from differences between CorTxt and the BNC and from the speech-like nature of texting (Table 9.1). Words significantly more frequent in text messaging than in FLOB include first- and second-person pronouns (*you*, *I*), as well as references to time (*soon*, *just*); and markers of phatic conversation: *hope*, *hi*, *hello*, *sorry* and *thanks*. Words significantly frequent in FLOB in comparison to CorTxt include a number of closed-class, rather than open-class, words: *was*, *which*, *he*, *and*, *his*, *of* and *the* (Tagg 2007: 274-275), again not surprising given the comparison with the BNC but meriting further investigation.

Table 9.1 Keywords in CorTxt, in comparison with FLOB

Positive keywords in CorTxt	Negative keywords in CorTxt
u	as
x	their
you	by
I	was
xx	which
Im	he
Hi	and
I'm	his
soon	of
hope	the
hello	
sorry	
pam	
hu	

[1] FLOB represents an attempt to create a modern version of the 1960s LOB corpus; COLT is a Corpus of London Teenagers' Language.

thanks just	

Differences reported in Tagg (2007) between keywords in CorTxt and COLT suggest that text messaging also differs from the spoken corpus, although a number of the differences are simply due to variant spellings in CorTxt, including *u* (*you*), *wot*, *ok*, *x*, *im*, *r* (*are*), *ur* (*you are*), and *luv* (*love*). However, some of the words more frequent in COLT are contractions such as *it's*, *that's*, *what's*, perhaps reflecting the suggestion that other types of ellipsis are often favoured in CorTxt (*Is* instead of *it's*, for example), as are contractions without the apostrophe: *its*, *thats*, *whats*. Other words more frequent in COLT than in CorTxt appear to be characteristic of spontaneous speech produced on-line: hesitation markers such as *erm*, *oh*, *yeah*, *er* and *mm*, as well as several function words *that*, *know*, *what*, *of*, *he* and *it*, the significance of which could be explored in greater depth. Further differences include the significantly higher frequency in CorTxt of *will*, *soon*, *hope*, and *text* and *thanks*, which can be accounted for not only by the need for brevity, which may explain the choice of *will* over the more lengthy *going to*, but by the role of CorTxt in managing interaction and co-ordinating future activities (Tagg 2007: 275). Such findings bolster the argument that, due to features of the 'unique' situation in which text messages are sent, texting differs in significant ways from spoken language. Further more finely-grained comparisons could be made between texting and the conversational and task-oriented subcorpora of the spoken BNC (Leech et al 2001: 223-246).

Table 9.2 Keywords in CorTxt, in comparison of COLT

Positive keywords in CorTxt		Negative keywords in CorTxt	
u	wot	that	no
x	hu	know	don't
ok	luv	he	right
xx	text	erm	she
xxx	thanks	oh	it
will	am	what	what's
soon		it's	one
im		laugh	mm
hope		yeah	he's
r		that's	of
its		like	because
h		goes	
pam		er	
ur			

Detailed comparisons between texting and computer-mediated forms of communication could also confirm texting as a distinct variety, whilst at the same time strengthening the more general argument, outlined in Chapter 2, that care must be taken in assuming similarities between varieties simply because of their shared electronic mode (Herring 2001; Thurlow and Brown 2003). Empirical studies exploring such distinctions include Gains' (1999) study of personal and business emails; Hard af Segersteg's (2002) investigations of SMS, email, instant messaging and webchat; and Baron and Ling's (2002) comparison of IM and SMS. North's (2007) study of MSN also begins to suggest differences between that and texting. However, in order to reinforce claims made in this thesis regarding the status of texting as a distinct variety, more specific comparisons could be made between texting and other electronic domains with reference to word frequency, creativity, phraselogy and grammar. With this in mind, it is hoped that further studies of emerging electronically-mediated language varieties such as MSN or Twitter can use this study as a point of comparison.

Comparisons are of course also needed between texting and varieties which are not electronically mediated. Although such comparisons were made throughout this thesis, most were narrowly concerned with respellings. Txt has been seen to resemble older forms of informal communication such as nineteenth-century working class letters (Kesseler and Bergs 2003) and telegrams (Crystal 2003: 425) and to reflect respellings occurring in other domains where prescription is relaxed, such as graffiti (e.g. Blume 1985) and trade names (e.g. Pound 1912). Although the general observation made in Chapter 5 was that Txt drew on respellings found in earlier and other texts (a point made also by Shortis 2007a,b), some distinctions emerged in terms of 'new' forms of respelling in CorTxt. My initial attempt to run CorTxt through VARD2, a spellchecker trained on historical texts (Baron and Rayson 2008), also revealed huge differences between Txt and the spellings typical of sixteenth century writing. Distinctions however should not be looked for solely in the spelling. For example, although noting a similar syntax, Kesseler and Bergs (2003) make initial observations regarding the differences between their corpus of letters and Txt, noting a simpler, more explicit style in the latter and more use of symbols (see Chapter 5). Where corpora are available, more systematic and delicate data-based comparisons could be made regarding the similarities and, more importantly, the differences between Txt and other varieties that focus not only on respellings but a range of linguistic features, thus

highlighting more precisely the aspects of texting that distinguish it from other types of language use.

The description of texting which emerges from my research is for the moment one which draws on written discourse (through, for example, the reflection and extension of spelling practices documented in CorTxt) as well as on spoken discourse (in that creativity in texting extends from spelling to the occurrence of everyday creativity, and also in the identification of spoken grammar features). This does not necessarily mean that texting should be described as combining spoken and written language. Instead, CorTxt reveals features apparently specific to texting which can be linked not just to constraints of the medium but to its exploitation by users in actively fulfilling expressive, informal purposes in creative and appropriate ways. Pedagogical implications of the view, which build on arguments and proposals made elsewhere (Kress 2000; Carter and McCarthy 2004; Carter 2007; Cook 1997), are outlined in Chapter 6 in relation to everyday creativity.[1] Here, the argument is made that the agency and awareness of texters, as well as the choices they make and the prevalence of creativity and variation across text messages, can be captured with the notion that texters are putting on or *performing* texting identities. Informed by the discourses surrounding texting, they create illusions of brevity through abbreviation and ellipsis, they perform spoken informality through colloquial contractions and certain features of spoken grammar, and they construe deviance and group membership through use of what sometimes appears as esoteric texting 'code'.

9.1.4 Performativity

The concept of *performativity* which I use in describing texting behaviour is taken from work into gender and language by the feminist theorist Butler (1990).[2] The concept can be found in the work of speech act theorist Austin (1962), who suggested that illocutionary acts like 'I promise' do not describe pre-existing states of affairs but bring one into being: in Butler's (1990) words, they constitute an 'aspect of discourse that has the capacity to produce what it names'. Butler (1990) extended this concept to address the construction of gender identity, arguing that men and women construct and reaffirm gender roles through the act of performing them. In other words, gender is neither biologically determined nor

[1] Similar arguments made about spelling, drawing on Kress (2000), are discussed alongside those regarding everyday creativity in Tagg (forthcoming).
[2] And see Cameron (1998) for an application of Butler's (1990) ideas to language analysis.

assigned at birth, but is forged through people's own reaction to cultural expectations. As Cameron (1998: 271) puts it: 'Gender has constantly to be reaffirmed and publicly displayed by repeatedly performing particular acts in accordance with the cultural norms'. The parallel with texting lies in the observation, made by Shortis (2007a,b), that the established use of certain respellings in text (and, I argue, other features) emerges from interaction rather than prescription. In Shortis' words:

> Respellings in Txt are 'natural', functional and uncodified and are interpreted and replicated by immersion rather than by formal instruction … it is an orthography remade by users in their practices rather than one which depends on being received, learned and directly replicated in the manner of the acquisition of standard language spelling accuracy.
>
> (Shortis 2007a: 2)

In other words, texters do not learn the features used in texting, but acquire them from interlocutors and affirm them through repeated use. In gender theory, this notion of performativity extends well-established beliefs that gender roles are socially rather than biologically determined to suggest that people are active producers of feminine and masculine roles, rather than having them passively assigned early in life. It also explains variability in how gender is constructed, as people can and do perform gender differently depending on whom they are with and other circumstances (Cameron 1998: 272). The relevance of this concept to texting behaviour is thus threefold. As well as being useful in explaining texter's linguistic choices, *performativity* enables us to assign an active role to participants, and explains variation in their behaviour.

The observation that texters are performing brevity, as well as performing spoken informality and deviance, helps to explain the linguistic choices which texters make in drawing on other discourses in order to fulfil the functions to which they put the medium. The suggestion that texters are performing brevity rests on the observation that the choices made cannot wholly be explained either in terms of a need to be brief, or by comparison with spoken ellipsis, or by reference to whether the omitted elements are necessary for communication. Texters are not necessarily or always 'being brief', as suggested by the fact that while certain elements are omitted across text messages others of equal length and of similar necessity in enabling communication are retained; the choices do not always match those made by speakers; and apparently 'abbreviated' text messages are frequently both lengthy and expressive. Alongside this abbreviation and ellipsis, texters also choose to draw on respellings taken from written discourses in which they mark deviance from expected

norms and membership of non-mainstream groups; and on spoken discourse in order to construe an informal and intimate language through texting. What emerges is a *performance* of brevity, spoken informality and deviance, informed by texters' awareness of and access to features from other varieties of language, as well as by discourses that circulate with respect to the abbreviated language of Txt and by texters' own experience of their own and interlocutors' texting practice, which their continued performance serves to affirm. The idea that people draw on dominant cultural discourses in the construction of linguistic identity is one used also in gender theory (see Coates 1998).

The active role that this suggests for texters in construing Txt also reflects that which the concept of performativity assigns to men and women in the construction of gender identity. Butler (1990) rejects arguments prevalent in feminist theory that gender roles were inevitably imposed by culture upon male and female bodies and argued that, although constrained by dominant perceptions of gender identity, people were able to choose what form their performance of gender took, and thus shape their own gender identity. According to Cameron (1998: 281), men and women do not simply learn and enact different ways of speaking. Instead, they inherit 'a much broader set of gendered meanings' which shapes their language and behaviour in complex and varied ways. Similarly, Txt is not an abbreviated language encouraged solely by physical constraints of the medium, but emerges from texters' chosen behaviour within a framework of related discourses, past texting experience, technological factors, and interpersonal relations. The agency of users and their exploitation of possibilities afforded by technology is similarly noted by other linguists (North 2006).

Variation in texting behaviour, once user agency is recognised, would appear inevitable. Again, this reflects gender theories which recognise that gender cannot be studied in isolation from specific social contexts or relationships (Butler 1990); that is, a person's performance of gender identity varies over time and according to contextual circumstances. Gender is not what you *are*, it is what you are doing at any particular time. Even for those adhering to rigid and traditional gender roles, argues Cameron (1998: 281), being masculine or feminine cannot involve performing the same regardless of the circumstances.

Texters are similarly unlikely to 'give the same performance' but will vary their texting style according to their interlocutor (peer, for example, or parent), the reason for texting

(quick requests versus chatty text messages), and the circumstances in which they are texting (public settings such as on a bus, say, or private settings on the sofa) (Shortis 2007a,b), all of which determine the apparent need for brevity, as well as the degree of speech-like informality and the extent to which deviance and group membership are marked through particular spellings. The notion of gender identity as theatrical performance can also be transferred to texting. As Osborne and Segal (1993) put it in their interview with Butler, gender is 'a kind of improvisational theatre, a space where different identities can be more or less freely adopted and explored at will' which is interesting to consider in relation to the creative language play seen in CorTxt.

My argument here is not to compare the construal of Txt to the construction of gender identity per se, but to highlight the usefulness of the term *perform* to describe how texters actively respond to varied circumstances, expectations and discourses and it is the effects of this performance that creates a text-specific language and which itself affirms the nature of Txt.

9.2 Limitations and future research

The limited size of CorTxt and its representativeness must, as in any corpus study, be acknowledged. Conclusions based on less than 200,000 texted words cannot be generalised to make statements about texting in general, but remain statements about the corpus with implications for wider descriptions and for more extensive comparisons with different text varieties. This is true of any corpus research which attempts to generalise beyond the data included within the corpus itself. CorTxt is particularly limited, however, in terms of its range. Firstly, contributors to the data form what I call a 'texting network' and are either known to me personally, or are friends of people I know. This in many ways is advantageous: not only in proving a ready and reliable source of data in that I could ensure text messages were not invented, but also in enabling me to interpret text messages in the light of my background knowledge and not dissimilar texting practices. At the same time, however, participants were not selected to be representative of the general texting population and the limitations this imposes on generalisations must be acknowledged. In fact, and this is my second point, contributors to this study are generally in their twenties or older and tend to be well-educated and highly literate people, whose texting practices are likely to differ from other sectors of the texting population. This is not to say that this

'group' of texters are any less interesting than teenagers, for example (despite or perhaps because of the focus, in other studies, on younger people) but, again, it does limit generalisations. Finally, however, it must be pointed out that, regardless of either the amount or range of data, a study of this nature, breadth or depth could not be described as a complete or perfect description of texting. Certain avenues of research were not explored in this study, either through limitations of the corpus or because of time and space constraints: it is not possible, of course, to be comprehensive within the limits of a doctoral thesis. Furthermore, although the thesis explores a number of approaches to data description, the findings are preliminary and only go some way towards increasing our understanding of the medium.

Further research could, however, overcome the above limitations and build on these initial findings. A larger set of data could be collected and a wider range of participants recruited through the use of websites into which people could enter their text messages, or from recruitment drives in local schools or universities. If the corpus was more carefully controlled than the present one in terms of the composition of its participants, it may also prove informative to subdivide the corpus into subcorpora based on gender, age or other social variables and to explore how linguistic features varied between different user groups. This was not done in the present study partly because of the corpus size and composition. With a limited number of participants and no attempt to control for variables such as age and gender, it was felt that analysis along gender or age lines would in fact produce statements skewed by participants rather than by social variables. A second caveat regards the validity of demarcating groups according to preconceived and to an extent arbitrary categories as a starting-point for the study of linguistic differences. An alternative is to start from shared linguistic features from which to determine shared social characteristics of those using the features, allowing social groups to emerge from the data: in Sealey's (2008) words, a 'data-driven approach to the identification of descriptive categories, both linguistic and social'. This approach is currently being developed by Sealey (2008) but could form a useful future avenue of research.

CorTxt, as well as any future corpora, can also be productively mined for other linguistic features which which it was not possible for practical reasons to include in this thesis. These possible avenues of research include the exchange structure of text message

conversations, either with reference to conversation analysis,[1] or to Sinclair and Coulthard's (1975) initiation-response-feedback (IRF) structure. Initial analysis suggests that IRF exchanges occur in texting.

(9.3)	A:	What is the plural of the noun research?	I
	B:	There generally isn't one. It's an uncountable noun - u in the dictionary. pieces of research?	R
	A:	Yes i thought so. Thanks.	F

There is also the suggestion that differences between the structure of spoken and texted exchanges may emerge through this approach. For example, the following exchange could be analysed as three initiations within one text message (I (1), (2), and (3)), with three responses occurring in the second (R (1), (2), and (3)):

(9.4)	A:	I am back. Good journey!	I (1)		
		Let me know if you need any of the receipts.		I (2)	
		Shall i tell NAME150 you like the pendent?			I (3)
	B:	Good!	R (1)		
		No, don't need any receipts—well done! (…)		R (2)	
		Yes, please tell NAME150. What's her number, i could ring her			R (3)

Another avenue for further research is the construction of narrative in texting. Narratives in CorTxt appear unsurprisingly to be highly personal and evaluative, and they contribute to identity construction through texting. Examples include:

(9.5)	We took hooch for a walk toaday and i fell over! Splat! Grazed my knees and everything! Should have stayed at home! See you tomorrow! Xx

(9.6)	Just got up. have to be out of the room very soon. …. i hadn't put the clocks back til at 8 i shouted at everyone to get up and then realised it was 7. wahay. another hour in bed.

[1] Building on initial work by Hutchby and Tanna (2008) who conduct a conversation analysis of 1250 messages.

(9.7) Honestly i've just made a lovely cup of tea and promptly dropped my keys in it and then burnt my fingers getting them out!

Another feature emerging from consideration of narrative is evaluation (Hunston and Thompson 2000) and the depiction of stance in texting, which includes but is not limited to, use of emoticons. These give positive and negative evaluations as well as indicating irony.

(9.8) Glad it went well **:)** come over at 11 then we'll have plenty of time before claire goes to work.

(9.9) Not really dude, have no friends i'm afraid **:(**

(9.10) Ahhhh...just woken up!had a bad dream about u tho,so i dont like u right now **:)** i didnt know anything about comedy night but i guess im up for it.x

(9.11) Hm good morning, headache anyone? **:-)**

As we have already seen, however, evaluation is also signalled by texters drawing on creative resources and on non-core and idiomatic language. Other stance markers associated with spoken discourse (Biber et al 2002; Carter and McCarthy 2006) are likely to occur, given the occurrence of spoken discourse markers documented in the current study.

Finally, as mentioned throughout this thesis, it is hoped that this example of the compilation of a text message corpus and investigation of the language used serves to pave the way for other researchers in their own explorations of this emerging medium.

REFERENCES

Abercrombie, D. (1969) 'Voice qualities' in Markell, N. (ed.).

Aijmer, K. and B. Altenberg (eds) (1991) *English Corpus Linguistics.* New York: Longman.

Aitchison, J. and D. Lewis (eds) (2003) *New Media Discourse.* London: Routledge.

Androutsopoulos, J. K. (2000) 'Non-standard spellings in media texts: the case of German fanzines' *Journal of Sociolinguistics* 4/4: 514-533.

Archer, D. and J. Culpeper (2003) 'Sociopragmatic annotation: new directions and possibilities in historical corpus linguistics' in Wilson et al (eds) pp. 37-58.

Archer, D., J. Culpeper and M. Davies (2008) 'Pragmatic annotation' in Ludeling, A. and M. Kyto (eds), 613-642.

Austin, J. (1962) *How to Do Things with Words.* Oxford: Clarenden.

BAAL (2006) 'Recommendations on Good Practice in Applied Linguistics'. Accessible: http://www.baal.org.uk/goodprac.htm. Accessed: 26th October 2007.

Baker, M., G. Francis and E. Tognini-Bonelli (eds) (1993) *Text and Technology: in Honour of John Sinclair.* Amsterdam: John Benjamins.

Baker, P. (2004) 'Querying keywords: questions of difference, frequency and sense in keywords analysis' *Journal of English Linguistics* 32/4: 346-359.

Bang, M. (2009) 'Representation of foreign countries in the US press: a corpus study'. *PhD thesis.* Birmingham: University of Birmingham.

Barnbrook, G. (1990) 'The computer analysis of spelling variation in Chaucer's Canterbury Tales'. *MPhil thesis.* English Department, University of Birmingham.

Baron, A. and P. Rayson (2008) 'VARD 2: A tool for dealing with spelling variation in historical corpora'. *Postgraduate Conference in Corpus Linguistics*, Aston University, 22nd May.

Baron, N. (1998a) 'Letters by phone or speech by other means: the linguistics of email' *Language and Communication* 18: 133-170.

Baron, N. (1998b) 'Writing in the Age of Email: the impact of ideology versus technology' *Visible Language* 32/1: 35-53.

Baron, N. (2000) *Alphabet to Email: How Written English Evolved and Where it's Heading.* London: Routledge.

Baron, N. (2002) 'Who sets email style? Prescriptivism, coping strategies, and democratising communication access' *The Information Society* 18: 403-413.

Baron, N. and R. Ling (2003) 'IM and SMS: a linguistic comparison'. *AOIR 4.0 International Conference of the Association of Internet Researchers*. Toronto, 16th-19th October.

Beard, A. (1998) *The Language of Sport.* London: Routledge.

Beard, A. (1999) *The Language of Politics.* London: Routledge.

Becher, T. (1987) 'Disciplinary discourse' *Studies in Higher Education* 12: 261-274.

Bednarek, M. (2006) 'Epistemological positioning and evidentiality in English news

discourse: a text-driven approach' *Text & Talk* 26/6: 635-660.

Bennett, J. and G. Smithers (eds) (1968) *Early Middle English Verse and Prose.* Oxford: Clarendon Press.

Bex, T. and R. Watts (eds) (1999) *Standard English: the Widening Debate.* London: Routledge.

Bhatia, V.K. (1993) *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, D. (1988) *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (1995) *Dimensions of Register Variation.* Oxford: Oxford University Press.

Biber, D. (2003) 'Compressed noun phrase structures in newspaper discourse: the competing demands of popularization vs economy' in Aitchison, J. and D. Lewis (eds), pp. 169-181.

Biber, D. and V. Clark (2002) 'Historical shifts in modification patterns with complex noun phrase structures: how long can you go without a verb?' in Fanego, T., M. Lopez-Couso and J. Perez-Guerra (eds), pp. 46-53.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Blume, R. (1985) 'Graffiti' in van Djik, T. (ed.), pp. 137-148.

Bolinger, D. (1946) 'Visual morphemes' *Language* 22: 333-340.

Bolinger, D. (1961) 'Syntactic blends and other matters' *Language* 37/3: 366-381.

Bolinger, D. (1976) 'Meaning and memory' *Forum Linguisticum* 1: 1-14.

Bowdre, P.H. (1982) 'Eye dialect as a literary device' in Williamson and Burke (eds), pp. 145-177.

Bradley, H. (1919) *On the Relations between Spoken and Written Language with Special Reference to English.* Oxford: Clarendon Press.

Brazil, D. (1995) *A Grammar of Speech.* Oxford: Oxford University Press.

Brinker, K. (2001) *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden (5. Aufl.).* Berlin: Erich Schmidt.

Butler, J. (1990) *Gender Trouble.* London: Routledge.

Burnard, L. and T. McEnery (eds) (2000) *Third International Conference on Teaching and Language Corpora*, Peter Lang.

Butt, D., R. Fahey, S. Feez, S. Spinks and C. Yallop (2000) *Using Functional Grammar: an Explorer's Guide.* Sydney: NCELTR.

Cameron, D. (1998) 'Performing gender identity: young men's talk and the construction of heterosexual masculinity' in Coates, J. (ed.), pp. 270-284.

Cameron, D. (2006) 'Metaphor in everyday language' in Maybin, J. and J. Swann (eds), pp. 46-53.

Carey, J. (1989) 'Consumer adoption of new communication technologies' *IEEE Communications Magazine* 27/8: 28-32.

Carney, E. (1994) *A Survey of English Spelling.* London: Routledge.

Carter, R. (1999) ‘Standard grammars, spoken grammars: some educational implications’ in Bex, T. and R. Watts (eds), pp 149-168.

Carter, R. (2004a) ‘Grammar and spoken English’ in Coffin et al (eds), pp. 25-39.

Carter, R. (2004b) *Language and Creativity: the Art of Common Talk.* London: Routledge.

Carter, R. (2007) ‘Discussant: Response to special issue of Applied Linguistics devoted to language creativity in everyday contexts’ *Applied Linguistics* 28/4: 597-608.

Carter, R. and S. Cornbleet (2001) *The Language of Speech and Writing.* London: Routledge.

Carter, R. and M. McCarthy (eds) (1988) *Vocabulary and Language Teaching.* London: Longman.

Carter, R. and M. McCarthy (2004) ‘Talking, creating: interactional language, creativity, and context’ *Applied Linguistics* 25/1: 62-88.

Carter, R. and M. McCarthy (2006) *Cambridge Grammar of English: a Comprehensive Guide.* Cambridge: Cambridge University Press.

Carter, R. and J. McRae (1996) *Literature and the Learner: Creative Classroom Practice.* Harlow: Longman/Pearson Education.

Chafe, W. (1982) ‘Integration and involvement in speaking, writing, and oral literature’ in Tannen, D. (ed.) pp. 35-53.

Chafe, W. and D. Tannen (1987) ‘The relation between written and spoken language’ *Annual Review of Anthropology* 16: 383-407.

Cheng, W. (2006) ‘Describing the extended meanings of lexical cohesion in a corpus of SARS spoken discourse’ *International Journal of Corpus Linguistics* 11/3: 325-344.

Cheng, W., N. Greaves and M. Warren (2006) ‘From n-gram to skipgram to concgram’ *International Journal of Corpus Linguistics* 11/4: 41-433.

Chomsky, N. (1957) *Syntactic Structures.* The Hague: Mouton.

Coates, J. (ed.) (1998) *Language and Gender: a Reader.* Oxford: Blackwell.

Cobb, T. (2003) ‘Review of Alison Wray 2002 Formulaic Language and the Lexicon’ *Canadian Journal of Applied Linguistics* 6/1: 105-110.

Coffin, C., A. Hewings and K. O’Halloran (2004) *Applying English Grammar: Functional and Corpus Approaches*. London: Arnold.

Colby, B. (1987) ‘Coherence in language and culture’ in Steele, R. and T. Threadgold (eds).

Collins English Dictionary (1994) Glasgow: HarperCollins.

Collins, R. (2004) *Interaction Ritual Chains.* Princeton: Princeton University Press.

Collot, M. and N. Belmore (1996) ‘Electronic language: a new variety of English’ in Herring, S. (ed.) pp. 13-28.

Conrad, S. and D. Biber (eds) (2001) *Variation in English: Multi-Dimensional Studies.* Harlow: Longman.

Conrad, S. (2001) ‘Variation among disciplinary texts: a comparison of textbooks and journal articles in biology and history’ in Conrad, S. and D. Biber (eds), pp. 94-107.

Cook, G. (1994) ‘Repetition and learning by heart: an aspect of intimate discourse, and its implications’ *English Language Teaching Journal* 48/2: 133-141.

Cook, G. (1997) 'Language Play, Language Learning' *English Language Teaching Journal* 51/3: 224-231.

Cook, G. (2000) *Language Play, Language Learning.* Oxford: Oxford University Press.

Cortes, V. (2002) 'Lexical bundles in freshman composition' in Reppen, R., S. Fitzmaurice and D. Biber (eds), pp. 131-145.

Coulthard, M. (ed.) (1992) *Advances in Spoken Discourse Analysis.* London: Routledge.

Coulthard, M. (2004) 'Author identification, idiolect, and linguistic uniqueness' *Applied Linguistics* 25/4: 431-447.

Coulthard, M. (2007) 'The linguist as detective'. Talk given at an English Langauge Postgraduate seminar, University of Birmingham, 30th November 2007.

Cramb, A. (2003) 'Girl writes English essay in phone text shorthand' in the *Daily Telegraph*, 3rd March.

Crystal, D. (1998) *Language Play.* Cambridge: Cambridge University Press.

Crystal, D. (2001) *Language and the Internet.* Cambridge: Cambridge University Press.

Crystal, D. (2003) *The Cambridge Encyclopedia of the English Language.* Cambridge: Cambridge University Press.

Crystal, D. (2008a) *Txtng: the Gr8 Db8.* Oxford: Oxford University Press.

Crystal, D. (2008b) *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell.

Crystal, D. and D. Davy (1969) *Investigating English Style.* Bloomington: Indiana University Press.

Dahl, H. (1979) *Word Frequencies of Spoken American English.* Michigan: Verbatim.

Dahlmann, I. (2007) 'How big is big enough? Methodological considerations on the determination of corpus size for the study of frequent multi-word units (MWUs) in spoken language'. *Corpus Linguistics Conference 2007*. University of Birmingham, 27th -30th July.

Danielewicz, J. and W. Chafe (1985) 'How 'normal' speaking leads to 'erroneous' punctuating' in Freedman, S. (ed.), pp. 213-225.

Davies, E. (1987) 'Eyeplay: on some uses of non-standard spelling' *Language and Communication* 7/1: 47-58.

DeVito, J. (1966) 'The encoding of speech and writing' *Speech Teacher* 15: 55-60.

Dewey, M. (2007) 'English as a lingua franca and globalisation: an interconnected perspective' *International Journal of Applied Linguistics* 17/3: 332-354.

Doring, N. (2002) '"1 bread, sausage, 5 bags of apples I.L.Y" - communicative functions of text messages (SMS)' *Zeitschrift für Medienpsychologie* 3.

Doyle, R. (1993) *The Snapper.* London: Penguin.

Dunlop, M. and A. Crossan (2000) 'Predictive text entry methods for mobile phones' *Personal Technologies* 4/2: 134-143.

Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence' *Computational Linguistics* 19/1: 61-74.

Eggins, S. and D. Slade (1997) *Analysing Casual Conversation.* London: Cassell.

Eisenstein, E. (1979) *The Printing Press as an Agent of Change.* Cambridge: Cambridge University Press.

Ess, C. (2002) 'Ethical decision-making and Internet research: recommendations from the AoIR ethics working committee'. Accessible: www.aoir.org/reports/ethics.pdf. Accessed: 22nd October 2004.

Fairon, C. and S. Paumier (2006) 'A translated corpus of 30,000 French SMS'. *LREC.* Geneva.

Fanego, T., M. Lopez-Couso and J. Perez-Guerra (2002) *English Historical Syntax and Morphology.* Amsterdam: John Benjamins.

Faulkner, X. and F. Culwin (2005) 'When fingers do the talking: a study of text messaging' *Interacting with Computers* 17/2: 167-185.

Ferguson, C. A. (1983) 'Sports announcer talk: syntactic aspects of register variation' *Language in Society* 2/2: 153-72.

Firth, J. R. (1935) 'The technique of semantics' *Transactions of the Philological Society* 36-72.

Fishman, J. (ed.) (1968) *Readings in the Sociology of Language.* The Hague/Paris: Mouton.

Flach, J., P. Hancock and K. Vicente (eds) (1995) *Global Perspectives on the Ecology of Human-Machine Systems.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Fortunati, L. (2001) 'The mobile phone between orality and writing'. *Third International Conference on Uses and Services in Telecommunications*. Paris.

Gains, J. (1999) 'Electronic Mail - a New Style of Communication or just a New Medium?: an Investigation into the Text Features of E-mail' *English for Specific Purposes* 18/1: 81-100.

Gairns, R. and S. Redman (2003) *Natural English Upper Intermediate.* Oxford: Oxford University Press.

Ghadessy, M., A. Henry and R. Roseberry (eds) (2001) *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: John Benjamins.

Gibson, J. J. (1986) *The Ecological Approach to Visual Perception.* Hillsdale, N.Y.: Lawrence Erlbaum.

Gibson, I. (2004) 'Text tidal wave hits record high' in *The Guardian*, 27th April.

Giles, H., W. Robinson and P. Smith (eds) (1980) *Language: Social Psychological Perspectives.* Oxford: Pergamon.

Gleason, J. (1980) 'The acquisition of social speech routines and politeness formulas' in Giles, H., W. Robinson and P. Smith (eds), pp. 21-27.

Gledhill, C. (1995) 'Scientific innovation and the phraseology of rhetoric. Posture, reformulation and collocation in cancer research articles'. *PhD Thesis,* Aston University.

Gledhill, C. (2000) *Collocations in Science Writing.* Tübingen: Gunther Narr Verlag.

Goffman, E. (1959) *The Presentation of Self in Everyday Life.* New York: Doubleday.

Gong, J. and P. Tarasewich (2005) 'Alphabetically constrained keypad designs for text entry on mobile devices'. *CHI 2005*.

Greaves, C. (2005) 'Introduction to ConcGram©'. *Tuscan Word Centre International*

Workshop. Certosa di Pontignano, Tuscany, Italy, 25th - 29th June.

Grinter, R. and M. Eldridge (2001) ‘‘y do tngrs luv 2 txt msg?’’ in Prinz et al (eds), pp. 219-238.

Grinter, R. E. and M. Eldridge (2003) ‘Wan2talk? Everyday text messaging’. *ACM Conference on Human Factors in Computing System (CHI)*.

Groom, N. (2003) ‘A corpus study of the patent specification genre’. *MPhil thesis.* English Department, University of Birmingham.

Groom, N. (2006) ‘Phraseology and epistemology in humanities writing: a corpus-driven study.’ *PhD thesis.* English Department, University of Birmingham.

Grover, D., M. King and C. Kuschler (1998) *Patent No US5818437, Reduced keyboard disambiguating computer*: Tegic Communications Inc.

Haestrup, J. (2001) *US Patent No 6,223,059. Communication Terminal Having a Predictive Editor Application.*

Halliday, M. A. K. (1978) *Language as Social Semiotic: the Social Interpretation of Language and Meaning.* London: Hodder Arnold.

Halliday, M. A. K. (1985/1989) *Spoken and Written Language.* Oxford: Oxford University Press.

Halliday, M. A. K. (1994) *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. K. and M. R. Martin (1993) *Writing Science: Literary and Discursive Power.* London: The Falmer Press.

Halliday, M. A. K., A. McIntosh and P. Strevens (1964) *The Linguistic Sciences and Language Teaching.* London: Longman.

Hard af Segerstag, Y. (2002) ‘Use and Adaptation of the Written Language to the Conditions of Computer-Mediated Communication’, *PhD thesis*, University of Goteborg.

Hardie, A. (2007) ‘Part-of-speech ratios in English corpora’ *International Journal of Corpus Linguistics* 12/1: 55-81.

Hardie, A. (2008) *Corpus Linguistics Advanced Research and Training (CLARET).* Lancaster

Harkin, J. (2002) *Mobilisation: the Growing Public Interest in Mobile Technology.* London: Demos.

Harrell, L. (1957) *A Comparison of the Development of Oral and Written Language in School-Age Children.* Lafayette, Ind.: Child Development Publications.

Hasselgren, J., E. Montnemery and M. Svensen (2003) ‘HMS: a predictive text entry method using bigrams’. *Ext. Abstracts 10th Conference of the European Chapter of the Association of Comp. Linguistics*.

Hendricks, W. O. (1971) ‘Review: Investigating English Style’ *Language* 47/4: 990-1000.

Herring, S. (1996) *Computer Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives.* Amsterdam: John Benjamins.

Herring, S. (2004) ‘Slouching towards the ordinary: current trends in CMC’ *New Media and Society* 6/1: 26-36.

Hirotaka, N. (2003) ‘Reassessing current cell phone designs: using thumb input effectively’.

Extended Abstracts of CHI 03.

Ho, C. and P. Teo (2007) *Discourse in the Modern World: Perspectives and Challenges*. Singapore: McGraw-Hill.

Ho, C. Anderson and Leong (eds) (forthcoming) *Transforming Literacies and Language: Innovative Technologies, Integrated Experiences*. Routledge.

Hoey, M. (2004) *Lexical Priming.* London: Routledge.

Hofland, K. and S. Johansson (1982) *Word Frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities.

Holmes, J. and M. Marra (2002) 'Over the edge? Subversive humor between colleagues and friends' *International Journal of Humor* 15: 65-87.

How, Y. (2004) 'Analysis of SMS efficiency'. *Undergraduate thesis*, National University of Singapore.

How, Y. and M.-Y. Kan (2005) 'Optimizing predictive text entry for short message service on mobile phones'. *Computer Human Interfaces International*.

Howden, M. (1984) 'Code and creativity in word formation' *Forum Linguisticum* 8/3: 213-22.

Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, S. (2003) 'Lexis, wordform and complementation pattern: a corpus study' *Functions of Language* 10: 31-60.

Hunston, S. and G. Francis (2000) *Pattern Grammar.* Amsterdam: John Benjamins.

Hunston, S. and J. Sinclair (2000) 'A local grammar of evaluation' in Hunston, S. and G. Thompson (eds), pp. 74-101.

Hunston, S. and G. Thompson (eds) (2000) *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press.

Hutchby, I. and V. Tanna (2008) 'Aspects of sequential organization in text message exchange' *Discourse and Communication* 2/2: 143-164.

Hymes, D. (1968) 'The ethnography of speaking' in Fishman, J. (ed.) pp. 99-138.

Hymes, D. (1981) *In Vain I Tried to Tell You. Essays in Native American Ethno-poetics.* Philadelphia: University of Pennsylvania Press.

Ives, S. (1971) 'A theory of literary dialect' in Willimson, J. and V. Burke (eds), pp. 145-177.

Jacobson, S. (1966) *Unorthodox Spelling in American Trademarks.* Stockholm: Almqvist & Wiksell.

Jacquith, J. (1976) 'Diagraphia and advertising: the public as guinea pig' *Visible Language* 10: 295-308.

Jaffe, A. (2000) 'Introduction: non-standard orthography and non-standard speech' *Journal of Sociolinguistics* 4/4: 497-513.

Jaffe, A. and S. Walton (2000) 'The voices people read: orthography and the representation of non-standard speech' *Journal of Sociolinguistics* 4/4: 561-587.

James, G., R. Davison, A. Cheung and S. Deerwester (1994) *English in Computer Science:*

a corpus-based lexical analysis. Hong Kong: Longman, for the Language Centre, Hong Kong University of Science and Technology.

Jellinek, D. (2003) 'How to make those mobiles count' in *The Guardian*, 8th October.

Johnson, L. (1975) 'Dread Beat an Blood'. Accessible: http://www.last.fm/music/Linton+Kwesi+Johnson/_/Dread+Beat+An%27+Blood. Accessed: 14th November 2008.

Jucker, A.H. (1992) *Social Stylistics: Syntactic Variation in British Newspapers*. Berlin: Mouton de Gruyter.

Kachru, Y. (2003) 'On definite reference in world Englishes' *World Englishes* 22/4: 497-510.

Kasesniemi, E.-L. and P. Rautiainen (2002) 'Mobile culture of children and teenagers in Finland' in Katz, J. and M. Aakhus (eds), pp. 10-192.

Katz, J. and P. Rautiainen (eds) (2002) *Perpetual Contact: Mobile Communication, Private Talk, Public Performance.* Cambridge: Cambridge University Press.

Keenan, E. O. and T. Bennett (eds) (1977) *Discourse across Time and Space.* Los Angeles: USC.

Kesseler, A. and A. Bergs (2003) 'Literacy and the new media: vita brevis, lingua brevis' in Aitchison, J. and D. Lewis (eds), pp. 75-84.

Kress, G. (2000) *Early spelling: between Convention and Creativity.* London: Routledge.

Krippendorf, K. (1980) *Content Analysis: an Introduction to its Methodology.* Newbury Park, CA: Sage.

Kroll, B. (1977) 'Ways communicators encode propositions in spoken and written English: a look at subordination and coordination' in Keenan, E. O. and T. Bennett (eds), pp. 69-108.

Kucera, H. and W. Francis (1967) *Computational Analysis of Present-day American English.* Providence, BI: Brown University Press.

Kuiper, K. (2004) 'Formulaic performance in conventionalised varieties of speech' in Schmitt, N. (ed.), pp. 37-54.

Lam, P. Y. W. (2007) 'A corpus-driven lexico-grammatical analysis of English tourism industry texts and the study of its pedagogical implications in English for Specific Purposes' *Language and Computers* 61: 71-89.

Lee (2001) 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle' *Language Learning and Technology* 5/3: 37-72.

Lee, M. (2006) 'Methodological issues in building a corpus of doctor-patient dialogues for metaphor'. *Cognitive-linguistic approaches: what can we gain by computational treatment of data? A theme session at DGKL-06/GCLA-06 (Meeting of the German Cognitive Linguistics Association)*. Munich, Germany: 7th October.

Leech, G. (1998) 'The Special Grammar of Conversation' in *The Longman Language Review*, 5: 9-14.

Leech, G., P. Rayson and A. Wilson (2001) *Word Frequency in Written and Spoken English: Based on the British National Corpus.* London: Longman.

Leech, G., M. Hundt, C. Mair and N. Smith (2009) *Change in Contemporary English: a Grammatical Study.* Cambridge: Cambridge University Press.

Leith, D. (1983) *A Social History of English.* London: Routledge.

Lesher, G., B. Moulton and D. Higginbotham (1998) 'Optimal character arrangements for ambiguous keyboards' *IEEE Trans on Rehabilitation Engineering* 6: 415-423.

Lewis, D. (1993) *The Lexical Approach.* Brighton: Language Teaching Publications.

Ling, R. (2004) *The Mobile Connection: the Cell Phone's Impact on Society.* San Francisco: Morgan Kaufmann.

Ling, R. (2005) 'Mobile communications vis-a-vis teen emancipation, peer group integration and deviance' in Harper, R., A. Taylor and L. Palen (eds), pp. 175-189.

Ling, R. (2007) 'The length of text messages and use of predictive texting: who uses it and how much do they have to say?' *American University TESOL Working Papers* 4. Available: http://www.american.edu/tesol/CMCLingFinal.pdf. Accessed: 12th November 2008.

Ling, R. (2008) *New Tech, New Ties: How Mobile Communication is Reshaping Social Cohesion.* London: MIT Press.

Ling, R. and N. Baron (2007) 'Text Messaging and IM: linguistic comparison of American college data' *Journal of Language and Social Psychology* 26/3: 291-298.

Ling, R. and B. Yttri (2002) 'Hyper-coordination via mobile phones in Norway' in Katz, J. and M. Aakhus (eds), pp. 139-169.

Louw, B. (1993) 'Irony in the text or insincereity in the writer? - the diagnostic potential of semantic prosody' in Baker, M., G. Francis and E. Tognini-Bonelli (eds), pp. 157-176.

Ludeling, A and M. Kyto (2008) *Corpus Linguistics: an International Handbook.* Berlin/New York: Mouton de Gruyter.

MacKenzie, I., H. Kober, D. Smith, T. Jones and E. Skepner (2001) 'Letterwise: prefix disambiguation for mobile text input. Technical report, Eatoni'. Accessible: http://www/eatoni.com/research/lw-mt.pdf. Accessed: 4th March 2006.

Madell, D. and S. Muncer (2004) 'Back from the beach but hanging on the telephone? English adolescents' attitudes and experiences of mobile phones and the Internet' *CyberPsychology* 7/3: 359-367.

Mair, C., M. Hundt, G. Leech and P. Smith (2002) 'Short term diachronic shifts in part-of-speech frequencies' *International Journal of Corpus Linguistics* 7/2: 245-264.

Malinowski, B. (1922) *Argonauts of the Western Pacific.*

Mann, C. and F. Stewart (eds) (2000) *Internet Communication and Qualitative Research: a Handbook for Researching Online.* London: Sage Publications Ltd.

Marcinkeviciene, R. (1998) 'Hapax Legomena as a Platform for Text Alignment' *Proceedings of the Third European Seminar 'Translation Equivalence*', Montecatini Terme, Italy, October 16-18, 1997.

Markell, N. (ed.) (1969) *Psycholinguistics: an Introduction to the Study of Speech and Personality.* London: The Dorsey Press.

Mason, O. and S. Hunston (2004) 'The automatic recognition of verb patterns: a feasibility

study' *International Journal of Corpus Linguistics* 9/2: 253-270.

Mason, O. (2007) 'Review: Linear Unit Grammar, Integrating speech and writing' *TESL-EJ* 11/2: 1-4.

Mason, O. (2008) 'Stringing together a sentence: linearity and the lexis-syntax interface' *Language and Computers* 64: 231-248.

Matthiessen, C. (2006) 'Frequency profiles of some basic grammatical systems: an interim report' in Thompson, G. and S. Hunston (eds), pp. 103-142.

Maybin, J. and J. Swann (eds) (2006) *The Art of English: Everyday Creativity.* Basingstoke: Palgrave Macmillan.

Maybin, J. and J. Swann (2007) 'Everyday creativity in language: textuality, contextuality, and critique' *Applied Linguistics* 28/4: 497-517.

McCarthy, M. (1988) 'Some vocabulary patterns in conversation' in Carter, R. and M. McCarthy (eds), pp. 181-200.

McCarthy, M. (1993) 'Spoken discourse markers in written text' in Sinclair, J., M. Hoey and G. Fox (eds), pp. 170-182.

McCarthy, M. and R. Carter (1994) *Language as Discourse: Perspectives for Language Teachers*. Harlow: Longman.

McEnery, A. and A. Wilson (1996) *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, A., R. Xiao and Y. Tono (2006) *Corpus-Based Language Studies: an Advanced Resource Book.* London: Routledge.

McEnery, A. and C. Gabrielatos (2006) 'English Corpus Linguistics' in Aarts, B. and A. McMahon (eds), pp. 33-71.

McLoughlin, L. (2000) *The Language of Magazines.* London: Routledge.

Miller, G. (1956) 'The magical number seven, plus or minus two: some limits on our capacity for processing information' *Psychological Review* 63: 81-97.

Milligan, C. (1999) Accessible: http://www.lancs.ac.uk/fss/resources/ethics/copyright.htm. Accessed: 19th October 2007.

Mobile Data Association (regularly updated) 'Text it: the UK's official guide to messaging'. Accessible: www.text.it/mediacentre/.

Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Oxford University Press.

Moon, R. (2008) 'Lexicography and linguistic creativity' *Lexikos* 18: 1-23.

Moran, C. and G. Hawisher (1998) 'The rhetorics and languages of electronic mail' in Snyder, I. (ed.) pp. 80-101.

North, S. (2006) 'Making connections with new technologies' in Maybin, J. and J. Swann (eds), pp. 209-260.

North, S. (2007) ''The voices, the voices': creativity in online conversation' *Applied Linguistics* 28/4: 209-260.

Nunan, D. (2008) 'Exploring genre and register in contemporary English' *English Today* 24: 56-61.

O'Keeffe, A. (2007) 'Keynote speech'. *Corpus Linguistics Conference 2007*. University of Birmingham: 27th-30th July.

O'Keeffe, A., M. McCarthy and R. Carter (2007) *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Oakey, D. (2009) 'The lexical bundle revisited: isolexical and isotextual comparisons'. *English Language Research (ELR) seminar: Corpus Linguistics and Discourse*. University of Birmingham: 10th February.

Oksman, V. and J. Turtianen (2004) 'Mobile communication as a social stage: meanings of mobile communication in everyday life among teenagers in Finland' *New Media and Society* 6/3: 319-339.

Olohan, M. (2004) *Introducing Corpora in Translation Studies.* London: Routledge.

Osborne, P. and L. Segal (1993) 'Judith Butler Interview'. Accessible: http://www.bbgunengine.com/butler.html. Accessed: 31st March 2007.

Oxford Advanced Learners Dictionary (2000) Oxford: Oxford University Press.

Partington, A. (2006) *The Linguistics of Laughter: a Corpus-Assisted Study of Laughter-Talk.* London: Routledge.

Partridge, A. (1964) *Orthography in Shakespeare and Elizabethan Drama.* Lincoln: University of Nebraska Press.

Pawley, A. and F.H. Syder (1983) 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency' in Richards J.C and R.W. Schmidt (eds), pp 191-226.

Pennycook, A. (2007) ''The rotation gets thick. The constraints get thin': creativity, recontextualisation, and difference' *Applied Linguistics* 28/4: 579-696.

Petrie (1999) 'Writing in Cyberspace: a study of the uses, style and content of E-mail'. Accessible: http://www.netinvestigations.net. Accessed: 4th March 2005.

Poesio, M. and R. Vieira (1998) 'A corpus-based investigation of definite description use' *Computational Linguistics* 24/2: 183-216.

Pound, L. (1925) 'The Kraze for 'K'' *American Speech* 1/1: 43-4.

Praninskas, J. (1968) *Trade Name Creation: Processes and Patterns.* The Hague: Mouton.

Preston, D. (2000) 'Mowr and mowr bayud spellin': confessions of a sociolinguist' *Journal of Sociolinguistics* 4/4: 615-621.

Prinz, W., M. Jarke, Y. Rogers, K. Schmidt and V. Wulf (2001) 'Proceedings of the Seventh European Conference on Computer-Supported Cooperative Work ECSCW '01'. *The Seventh European Conference on Computer-Supported Cooperative Work ECSCW '01*. Bonn, Germany.

Quirk, R. (1969) 'Foreward' in Crystal, D. and D. Davy (eds).

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985) *A Grammar of Contemporary English.* London: Longman.

Rayson, P. (2008) 'From key words to key semantic domains' *International Journal of Corpus Linguistics* 13/4: 519-549.

Read, J. and P. Nation (2004) 'Measurement of formulaic sequences' in Schmitt, N. (ed.) pp. 23-35.

Reppen, R., S. Fitzmaurice and D. Biber (eds) (2002) *Using Corpora to Explore Linguistic*

Variation. Amsterdam: John Benjamins.

Richards, J.C. and N. Schmidt (eds) (1993) *Language and Communication*. New York: Longman.

Rock, F. (2001) 'Policy and practice in the anonymisation of linguistic data' *International Journal of Corpus Linguistics* 6/1: 1-26.

Saied, N. (2001) 'Fast, ful text entry using a physical or virtual 12-button keypad. Technical report'. Accessible: http://www.exideas.com/ME/whitepaper.pdf. Accessed: 4th March 2006.

Sampson, G. (1999) 'Correspondence collected from the Corpora Mailing List' in Rock, F. (ed.).

Saussure, F. d. (1916/1966) *Course in General Linguistics.* New York: McGraw-Hill.

Schiffrin, D. (1987) *Discourse Markers.* Cambridge: Cambridge University Press.

Schmitt, N. (ed.) (2004) *Formulaic Sequences: Acquisition, Processing, and Use.* Amsterdam: John Benjamins.

Schmitt, N. and R. Carter (2004) 'Formulaic sequences in action' in Schmitt, N. (ed.) pp. 1-.

Schmitt, N., S. Grandage and S. Adolphs (2004) 'Are corpus-derived recurrent clusters psychologically valid?' in Schmitt, N. (ed.) pp. 127-152.

Scott, M. (1996) *Wordsmith Tools.* Oxford: Oxford University Press.

Scott, M. (1997) 'PC analysis of key words - and key key words' *System* 25/2: 233-245.

Scott, M. (2001) 'Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suit of computer programs' in Ghadessy, M., A. Henry and R. Roseberry (eds), pp. 47-70.

Scragg, D. (1974) *A History of English Spelling.* New York: Barnes and Noble Books.

Sealey, A. (2008) 'Getting the measure of the data: identifying linguistic and social patterns in an oral history corpus'. *41st BAAL Annual Meeting*. Swansea University: 11th-13th September.

Sebba, M. (2007) *Spelling and Society.* Cambridge: Cambridge University Press.

Seidlhofer, B. (2004) 'Research perspectives on teaching English as a lingua franca' *Annual Review of Applied Linguistics* 24: 209-239.

Semino, E. (2004) *Corpus Stylistics.* London: Routledge.

Shortis, T. (2001) *The Language of ICT: Information Communication Technology.*

Shortis, T. (2007a) ''Revoicing Txt: spelling, vernacular orthography and 'unregimented writing'' in Posteguillo, S., M. J. Esteve and M. L. Gea (eds).

Shortis, T. (2007b) 'Gr8 Txtpectations: the creativity of text spelling' *English Drama Media Journal* 8: 21-26.

Sinclair, J. (ed.) (1987) *Looking Up: An Account of the COBUILD Project.* London: HarperCollins.

Sinclair, J. (1991) *Corpus Concordance Collocation.* Oxford: Oxford University Press.

Sinclair, J. (1992) 'The automatic analysis of corpora'. *Proceedings of Nobel Symposium 82*. Stockholm

Sinclair, J. (2000) 'Lexical Grammar' *Naujoji Metodolojija* 24: 191-203.

Sinclair, J. (2001) 'Preface' in Ghadessy, M., A. Henry and R. Roseberry (eds), pp. vii-xvi.

Sinclair, J. (2004) *Trust the Text.* London: Routledge.

Sinclair, J. and M. Coulthard (1975) *Towards an Analysis of Discourse: the English Used by Teachers and Pupils.* Oxford: Oxford University Press.

Sinclair, J. and M. Couthard (1992) 'Towards an analysis of discourse' in Coulthard, M. (ed.) pp. 1-34.

Sinclair, J., M. Hoey and G. Fox (1993) *Techniques of Description: Spoken and Written Discourse*. London: Routledge.

Sinclair, J. and A. Mauranen (2007) *Linear Unit Grammar.* Amsterdam: John Benjamins.

Sinclair, J. and A. Renouf (1991) 'Collocational frameworks in English' in Aijmer, K. and B. Altenberg (eds), pp. 128-143.

Snyder, I. (ed) *Page to Screen: Taking Literacy into the Electronic Era*. London: Routledge.

Steele, R. and T. Threadgold (eds) (1987) *Language Topics: Essays in Honour of Michael Halliday. Volume III.* Amsterdam: John Benjamins.

Stubbs, M. (1988) *Educational Linguistics*. Oxford: Basil Blackwell.

Stubbs, M. (1994) 'Grammar, text and ideology: computer-assisted methods in the linguistics of representation' *Applied Linguistics* 15/2: 201-223.

Stubbs, M. (1998) 'German loan words and cultural stereotypes' *English Today* 14: 19-26.

Stubbs, M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell.

Sutherland, J. (2002) 'Cn u txt?' in *The Guardian*, 11th November.

Swales, J. (1990) *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Swann, J. (2006) 'The art of the everyday' in Maybin, J. and J. Swann (eds), pp. 3-53.

Tagg, C. (2007) 'A corpus based analysis of SMS text messaging' in Teo, P. and C. Ho (eds), pp 267-284.

Tagg, C. (forthcoming) 'Wot did he say or could u not c him 4 dust? Written and spoken creativity in text messaging' in Ho, C. Anderson and Leong (eds).

Tannen, D. (1982) *Spoken and Written Language.* Norwood, NJ: Ablex.

Tannen, D. (1989) *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse.* Cambridge: Cambridge University Press.

Taylor, A. and R. Harper (2003) 'The gift of the gab?: a design oriented sociology of young people's use of mobiles' *Journal of Computer Supported Cooperative Work* 12(3): 267-296.

Terkourafi, M. (2003) 'Book review: Formulaic Language and the Lexicon' *Journal of Pragmatics* 35: 1721-1726.

Thompson, G. and S. Hunston (eds) (2006) *System and Corpus: Exploring Connections.* London: Equinox.

Thurlow, C. and A. Brown (2003) 'Generation Txt? Exposing the sociolinguistics of young

people's text-messaging' *Discourse Analysis Online* 1/1.

Thurlow, C. (2006) 'From Statistical to Moral Panic: the metadiscursive construction and popular exaggeration of new media language in the print media' *Journal of Computer-Mediated Communication* 11/3: 667-701.

Todd, R. (1999) 'Correspondence collected from the Corpora Mailing List' in Rock, F. (2001.).

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work.* Amsterdam: John Benjamins.

Tribble, C. (2000) 'Genres, keywords, teaching: towards a pedagogic account of the language of project proposals' in Burnard, L. and T. McEnery (eds), pp 75-90.

University of Leicester (2006) 'Press Release: Texts to reveal 'whodunnit''. Accessible: http://www2.le.ac.uk/ebulletin/news/press-releases/2000-2009/2006/08/nparticle.2006-08-09.0473409952. Accessed: 7th February 2008.

Van Dijk, T.A. (1985) *Discourse and Literature*. Amsterdam and Philadelphia: John Benjamins.

Weber, R. (1986) 'Variation in spelling and the special case of colloquial contractions' *Visible Language* 20/4: 413-426.

Werry, C. (1996) 'Linguistic and interactional features of Internet Relay Chat' in Herring, S. (ed.) pp. 47-64.

Willans, G. (1953) *Down with the skool! A Guide to School Life for Tiny Pupils and their Parents*. London: Pavilion.

Williamson, J.V. and V.M. Burke (eds) (1971) *A Various Language*. New York: Holt, Reinehart and Winston.

Willis, D. (1990) *The Lexical Syllabus.* Oxford: Macmillan Heinemann.

Wilson, A., P. Rayson and A. McEnery (2003) *Corpus Linguistics by the Lune: a Festschrift for Geoffrey Leech*. Frankfurt: Peter Lang.

Woolls, D. and M. Coulthard (1999) 'Tools for the trade' *Forensic Linguistics [now Journal of Speech, Language and the Law]* 5: 33-57.

Wray, A. (2002) *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008) *Formulaic Language: Pushing the Boundaries.* Oxford: Oxford University Press.

Xiao, R. and A. McEnery (2005) 'Two approaches to genre analysis: three genres in modern American English' *Journal of English Linguistics* 33/1: 62-82.

Yates, S. (1996) 'Oral and written aspects of computer conferencing' in Herring, S. (ed.) pp. 13-28.

Zaff, B. (1995) 'Designing with affordances in mind' in Flach, J., P. Hancock and K. Vicente (eds), pp. 238-272.

ZiCorp (2002) 'eZiText. Technical report'. Accessible: http://www.zicorp.com. Accessed: 4th March 2006.

APPENDICES

Appendix 4.1 *Emailed request for participation in the text message project*

Dear All

I'm writing to request your help in my PhD research into SMS text messages. Please don't feel you have to, but I would be very grateful if you felt you could.

I need to collect a database of text messages in order to examine we use language when we text. I will be looking at things like spellings, linguistic creativity, and how SMS is different from speech and from writing.

All I am asking you to do is to save messages you send and receive, to copy them into the attached Excel template and send them to me (if your phone does not save the messages you send, the received messages are enough). Alternatively, if you already save your messages, you could retrieve ones already received and sent, and submit those. I also need to know:

- the date when each message was sent;
- the time when the messages were sent;
- the name/initials of the sender (or some way of distinguishing between them and so you can remember who they are);

I realise that copying out the messages takes time and for this I am doubly grateful for your effort.

I will also have to ask you to complete a short personal profile and to sign a form giving me consent to use the messages. All the people who have sent you messages which you then send me must also complete this consent form. This can all be done electronically, so should be fairly straightforward. We can arrange this after you have sent me the messages, but please could you tell the people who have sent you messages what you are doing with their messages, and check that this is ok.

The messages will be completely anonymised, with names etc changed so that none of the parties involved can be identified. The personal profiles will be used to describe the corpus, or sections of it, as a whole and not for individual identification of anyone involved.

I am interested in determining how people text on a day-to-day basis and so no text is too boring! Please don't pick out 'interesting' or 'funny' texts - I need the mundane ones too! Having said that, if you have private messages or ones that you don't feel comfortable sending, then of course don't.

If you can help, please let me know. I can then explain further if you would like me to. A big thank you in advance to all those who feel they can help. It is much appreciated and very very useful.

Hope to hear from you soon,

Caroline

Appendix 4.2 *Consent form and personal information questionnaire*

User data User: _____

Please could you complete the following. This information will be held separately from the text messages collected, and will not be made available to any party other than the researcher.

Date of Birth: _____

Gender: M / F (delete as appropriate)

Profession: _____

Native language: _____

Make of phone: _____

Do you generally use predictive text devices (that is, where the phone predicts the word you are typing)?

yes no sometimes

How long have you used SMS text messaging? _____ years _____ months

How often do you send SMS? Please mark with a cross where you would place yourself on the scale below, assuming that:

1 = Infrequent user (sending less than one text message a day);
3 = Daily user (sending one or two messages a day);
5 = Frequent user (sending more than one or two messages a day).

1 2 3 4 5
_____

Consent to use text messages for research

I consent to the contents of my SMS text messages between myself and others participating in the research project organised by Ms Caroline Tagg, sent during the period November 2005 until October 2007, being used by Ms Tagg for the purpose of linguistic analysis as part of her research activities at the University of Birmingham. I also consent to the use of any of my personal data necessary for the purpose of analysis in connection with this research. I understand that linguistic findings pertaining to the text messages may be used for a commercial purpose, but that individual messages or personal data will not.

I give my consent on the understanding that all such personal data relating to myself shall be processed in accordance with the Data Protection Act 1998 by Ms Tagg, that the personal data shall be destroyed by the end of 2010, and that, if the contents of the messages are ever published in a dissertation, thesis or other format for any purpose, all such content will be anonymized by Ms Tagg such that neither myself or my correspondents can be identified from that content.

SIGNATURE DATE

NAME

Appendix 5.1 *The top two hundred and fifty most frequent headwords/groups in the text messaging corpus (including those without respellings).*

	Headword	Freq	%	Variant forms
1	YOU	7,884	4.14	u(3043),ya(256),yer(14),ye(9),uu(2)
2	TO	4,976	2.61	2(690),ot(3)
3	I	4,257	2.23	
4	X	3,689	1.94	xx(833),xxx(635),xxxx(11),xxxxxx(3),xxxxxxx(2), xoxox(1),xxxxx(42)
5	A	3,580	1.88	
6	THE	3,553	1.86	d(21),da(6),th(8),hte(3),te(2),ze(2)
7	AND	3,171	1.66	n(182),an(19),adn(10),amd(2),annd(2)
8	IN	2,387	1.25	iin(2)
9	FOR	2,057	1.08	4(357),fer(2)
10	IT	2,020	1.06	
11	HAVE	1,993	1.05	av(8),hve(6),ave(5),hav(106)
12	IS	1,577	0.83	
13	BE	1,567	0.82	b(375)
14	ME	1,555	0.82	
15	ON	1,523	0.8	
16	ARE	1,478	0.78	r(422), ar(2)
17	OF	1,420	0.75	
18	AT	1,393	0.73	
19	MY	1,285	0.67	
20	GOOD	1,265	0.66	gud(40),gd(25),goodo(3)
21	BUT	1,257	0.66	
22	SEE	1,255	0.66	c(248)
23	JUST	1,240	0.65	jus(18),jst(6),jurt(2)
24	I'M	1,216	0.64	im(280)
25	SO	1,159	0.61	soo(1),sooo(5),soooo(1)
26	IF	1,125	0.59	
27	WILL	1,124	0.59	wil(12)
28	THAT	1,118	0.59	tht(1),dat(4),tha(2)
30	YOUR	1,111	0.58	ur(286),yor(2),yer(9),yr(13),u'r(2),ure(1)
29	NOT	1,106	0.58	nt(4)
31	DO	1,061	0.56	d'(12),du(5)
32	OK	1,040	0.55	okay(75),okey(27),k(33),okej(1),okie(12)
33	UP	1,032	0.54	
34	WITH	1,002	0.53	wiv(82),wid(4),wiht(3)
35	WAS	952	0.5	woz(4),ws(3)
36	YES	952	0.5	yep(69),yeh(22),yea(15),yeah(382),yup(13)
37	GET	946	0.5	
38	WE	931	0.49	
39	CAN	894	0.47	cn(1)
40	TOMORROW	888	0.47	tomoz(9),tomorro(6),tomorow(4),tomora(3),tomo(361),mor row(6),mora(1),tom(24),2mora(14),tomoro(10),2morrow(9) ,tmw(9),2morow(4),2morro(4),2mrrw(4),2moz(3),2mrw(2), amoro(2),tomorrrow(2),2moro(42)
41	WHAT	839	0.44	wot(148),wat(37)
42	ALL	788	0.41	
43	KNOW	776	0.41	no(57),knw(1),kno(2)
44	ABOUT	768	0.4	bout(123),abt(4)
45	HOPE	746	0.39	hpe(2)
46	OUT	739	0.39	
47	BACK	733	0.38	bak(38),bac(7),bck(3),bk(11),bek(2)
48	GOT	715	0.37	
49	NO	714	0.37	nope(19),nah(18)

50	NOW	713	0.37	nw(9)
51	GOING	710	0.37	goin(60)
52	ITS	696	0.37	it's(240),tis(21)
53	HOW	692	0.36	hw(2)
54	DON'T	671	0.35	dont(198),dnt(4)
55	TIME	663	0.35	
56	AM	649	0.34	
57	I'LL	628	0.33	ill(46),il(21),i'l(30)
58	THIS	598	0.31	dis(6)
59	WHEN	589	0.31	wen(75)
60	THERE	588	0.31	ther(2)
61	NIGHT	586	0.31	nite(105)
62	AS	582	0.31	
63	OR	579	0.3	
64	HI	569	0.3	
65	FROM	568	0.3	frm(9),fm(3)
66	THEN	564	0.3	
67	HAD	562	0.3	hd(1),ad(2)
68	TOO	554	0.26	2(61)
69	GO	542	0.28	
70	WANT	446	0.28	
71	COME	537	0.28	cum(8),com(2)
72	WORK	526	0.28	
73	WELL	520	0.27	wel(5)
74	ONE	514	0.27	1(88)
75	HEY	512	0.27	
76	DAY	508	0.27	
77	THINK	503	0.26	
78	SORRY	502	0.26	sori(3),sos(12)
79	SOON	499	0.26	sn(6)
80	HOME	496	0.26	
81	HE	464	0.24	
82	STILL	461	0.24	stil(14)
83	DID	458	0.24	
84	BEEN	446	0.23	bin(22)
85	LATER	437	0.23	l8r(20),lata(5),l8er(2)
86	TONIGHT	431	0.23	2nite(45),tonite(10),2night(12),2nigt(3)
87	HELLO	419	0.22	hallo(1),ello(9),allo(5),helo(6),herro(3)
88	THANKS	412	0.22	thanx(32),thx(5),sanks(2),thnx(2),thks(1),thanxs(1),tnx(2)
89	WEEKEND	393	0.21	wknd(2),wkend(48),w'end(10),w'kend(3)
90	LIKE	387	0.2	
91	COULD	380	0.2	cud(48),cld(19)
92	VERY	380	0.2	v(126)
93	WEEK	380	0.2	wk(60),weeek(2)
94	YOU'RE	376	0.19	ur(177),yr(4),u'r(1),ure(1),you'r(2),youre(6)
95	SOME	373	0.2	sum(10)
96	TEXT	369	0.19	txt(94),tx(6),tex(2)
97	LOVE	368	0.19	luv(46),lv(4),lov(4)
98	HERE	367	0.19	ere(3)
99	WOULD	366	0.19	wud(38),wld(22)
100	CAN'T	363	0.19	cant(121),cannot(3)
101	LET	361	0.19	
102	NEED	361	0.19	
103	TODAY	349	0.18	2day(44),dey(2)
104	I'VE	342	0.18	ive(37),iv(5),i'v(3)
105	THO	332	0.17	though(145),although(8),thou(1),altho(2)
106	HAPPY	330	0.17	hapy(2),appy(1)
107	REALLY	329	0.17	realy(5)

108	OFF	328	0.17	
109	BECAUSE	306	0.16	cos(216),coz(24),cause(4),cuz(3),cs(2),cus(2),cz(2)
110	ANY	305	0.16	ne(8)
111	COMING	301	0.16	comin(23), cumin(6)
112	NEXT	300	0.16	nxt(29)
113	BIT	295	0.15	
114	THAT'S	295	0.15	thats(129)
115	MUCH	292	0.15	
116	OH	292	0.15	o(30)
117	BY	284	0.15	
118	LAST	284	0.15	lst(2)
119	SHE	282	0.15	
120	NEW	279	0.15	
121	SURE	279	0.15	
122	SHOULD	277	0.15	shud(11),shld(7),shd(3)
123	RING	275	0.14	
124	AN	271	0.14	
125	PHONE	268	0.14	fone(16)
126	WHERE	266	0.14	wher(4)
127	AGAIN	262	0.14	
128	YET	255	0.13	
129	WAY	253	0.13	
130	DOING	252	0.13	doin(33)
131	AFTER	251	0.13	afta(1)
132	PLEASE	250	0.13	pleasey(4),plez(3),pls(35),plse(3),plz(4),pse(2)
133	DIDN'T	249	0.13	didnt(42)
134	GREAT	249	0.13	gr8(3),g8(3)
135	HAS	245	0.13	
136	THEY	245	0.13	
137	HIM	243	0.13	
138	FINE	241	0.13	
139	CALL	237	0.12	cal(3)
140	COOL	235	0.12	
141	BIRTHDAY	232	0.12	bday(28),b'day(10)
142	GIVE	231	0.12	giv(10),gv(1)
143	HAVING	230	0.12	havin(24),avin(6)
144	DONE	227	0.12	dun(1)
145	EVENING	226	0.12	eve(58),evo(3)
146	NICE	224	0.12	
147	SAT	224	0.12	saturday(86),satdy(3)
148	SAY	222	0.12	
149	THEM	222	0.12	em(25)
150	UNTIL	221	0.12	til(104),till(70),untill(1)
151	GONA	220	0.12	gunna(11),gona(6)
152	FRIDAY	220	0.12	fri(79)
153	FUN	220	0.12	
154	PROB	219	0.11	problem(37),probably(85),probaly(3)
155	HOWS	217	0.11	how's(102),howz(8)
156	MORNING	216	0.11	mornin(23),morn(18)
157	MEET	215	0.11	
158	CHRISTMAS	214	0.11	xmas(68),crimbo(2),chrimbo(2)
159	HER	214	0.11	
160	MIGHT	214	0.11	mite(9),myt(6)
161	MAKE	209	0.11	
162	ONLY	207	0.11	
163	US	207	0.11	
164	SPEAK	206	0.11	spk(8)
165	DOWN	204	0.11	

166	MORE	203	0.11	
167	OVER	199	0.1	ova(1)
168	MATE	193	0.1	m8(10)
169	BEFORE	191	0.1	b4(29),befor(2)
170	BETTER	191	0.1	betta(2),beta(6),bettr(1)
171	NUMBER	189	0.07	num(7),numba(3),no(50)
172	TAKE	187	0.1	
173	SOMETHING	186	0.1	sth(9),somethin(8),summort(3),sumfing(2),summat(2),summing(2)
174	FANCY	180	0.09	
175	AROUND	179	0.09	
176	HAVEN'T	178	0.09	havent(36),have'nt(2),havn't(2),havnt(8)
177	LATE	174	0.09	
178	SOUNDS	169	0.09	
179	WORKING	169	0.09	workin(22)
180	YEAR	169	0.09	yr(4)
181	ANYTHING	167	0.09	anythin(3)
181	WERE	163	0.09	
182	LOVELY	161	0.08	
183	CHEERS	157	0.08	
184	THOUGHT	157	0.08	thort(5)
185	SAID	156	0.08	
186	TRY	155	0.08	
187	WE'RE	155	0.08	we'r(3),were (27))
188	SEND	154	0.08	snd(1)
189	GETTING	152	0.08	gettin(17),getin(1),geting(1)
190	TELL	152	0.08	
191	SHALL	151	0.08	
192	WHO	151	0.08	wo(2)
193	HOUSE	148	0.08	
194	CAR	145	0.08	
195	MUM	145	0.08	
196	WON'T	145	0.08	wont(45)
197	FREE	144	0.08	
198	LOOKING	143	0.08	lookin(18)
199	SUNDAY	142	0.07	sun(54)
200	WENT	140	0.07	wnt(1)
201	ROUND	139	0.07	
203	FEEL	138	0.07	fil(1)
204	MAYBE	137	0.07	mayb(3)
205	THANK	136	0.07	tank(2)
209	ANYWAY	135	0.07	
206	FIRST	135	0.07	1st(26)
207	EARLY	134	0.07	
208	THINGS	134	0.07	
210	TRAIN	132	0.07	
211	WHAT'S	132	0.07	whats(48),wots(17),wats(7),wot's(3)
212	KEEP	131	0.07	
213	HIS	130	0.07	hs(2)
214	MONDAY	130	0.07	mon(37)
215	LONG	126	0.07	
216	THURS	126	0.07	thursday(48),thrus(1),thu(6),thur(2)
217	BOTH	125	0.07	
218	HALF	125	0.07	
219	ALREADY	123	0.06	
220	THING	123	0.06	thiing(2)
221	BEING	122	0.06	bein(11)
222	BED	121	0.06	

223	INTO	121	0.06	in2(3)
224	LITTLE	121	0.06	liddle(1)
225	LOOK	121	0.06	lk(1)
226	LEFT	120	0.06	
227	OTHER	120	0.06	
228	MEETING	119	0.06	meetin(7)
229	OUR	119	0.06	
230	LOTS	118	0.06	
231	PICK	118	0.06	
232	RIGHT	118	0.06	rite(3)
233	DRINK	117	0.06	
234	ENJOY	117	0.06	
235	TA	116	0.06	
236	FINISHED	114	0.06	finshed(2),finnished(2)
237	MO	114	0.06	moment(41)
238	TOWN	114	0.06	
239	WHICH	114	0.06	wich(2)
240	THINKING	112	0.06	thinkin(14)
241	WHY	112	0.07	y(7)
242	AFTERNOON	110	0.06	avo(4),pm(25),arvo(3)
243	LEAVE	110	0.06	
244	MAY	110	0.06	
245	MESSAGE	110	0.06	msg(7),mess(2)
246	BUSY	109	0.06	
247	PUB	109	0.06	
248	WATCHING	109	0.06	watchin(17)
249	BIG	107	0.06	
250	JOB	107	0.06	

Appendix 5.2 *The 150 most frequent spelling groups (excluding words which are not respelt)*

	Headword	Freq	%	Variant forms
1	YOU	7,884	4.14	u(3043),ya(256),yer(14),ye(9),uu(2)
2	TO	4,976	2.61	2(690),ot(3)
3	X	3,689	1.94	xx(833),xxx(635),xxxx(11),xxxxxx(3),xxxxxxx(2),xoxox(1),xxxxx(42)
4	THE	3,553	1.86	d(21),da(6),th(8),hte(3),te(2),ze(2)
5	AND	3,171	1.66	n(182),an(19),adn(10),amd(2),annd(2)
6	IN	2,387	1.25	iin(2)
7	FOR	2,057	1.08	4(357),fer(2)
8	HAVE	1,993	1.05	av(8),hve(6),ave(5),hav(106)
9	BE	1,567	0.82	b(375)
10	ARE	1,478	0.78	r(422), ar(2)
12	GOOD	1,265	0.66	gud(40),gd(25),goodo(3)
11	SEE	1,255	0.66	c(248)
13	JUST	1,240	0.65	jus(18),jst(6),jurt(2)
14	I'M	1,216	0.64	im(280)
15	SO	1,159	0.61	soo(1),sooo(5),soooo(1)
16	WILL	1,124	0.59	wil(12)
17	THAT	1,118	0.59	tht(1),dat(4),tha(2)
18	YOUR	1,111	0.58	ur(286),yor(2),yer(9),yr(13),u'r(2),ure(1)
19	NOT	1,106	0.58	nt(4)
20	DO	1,061	0.56	d'(12),du(5)
21	OK	1,040	0.55	okay(75),okey(27),k(33),okej(1),okie(12)
22	WITH	1,002	0.53	wiv(82),wid(4),wiht(3)
23	WAS	952	0.5	woz(4),ws(3)
24	YES	952	0.5	yep(69),yeh(22),yea(15),yeah(382),yup(13)
25	CAN	894	0.47	cn(1)
26	TOMORROW	888	0.47	tomoz(9),tomorro(6),tomorow(4),tomora(3), tomo(361),morrow(6),mora(1), tom(24),2mora(14), tomoro(10),2morrow(9),tmw(9),2morow(4),2morro(4), 2mrrw(4),2moz(3),2mrw(2),amoro(2),tomorrrow(2),2moro(42)
27	WHAT	839	0.44	wot(148),wat(37)
28	KNOW	776	0.41	no(57),knw(1),kno(2)
29	ABOUT	768	0.4	bout(123),abt(4)
30	HOPE	746	0.39	hpe(2)
31	BACK	733	0.38	bak(38),bac(7),bck(3),bk(11),bek(2)
32	NO	714	0.37	nope(19),nah(18)
33	NOW	713	0.37	nw(9)
34	GOING	710	0.37	goin(60)
35	ITS	696	0.37	it's(240),tis(21)
36	HOW	692	0.36	hw(2)
37	DON'T	671	0.35	dont(198),dnt(4)
38	I'LL	628	0.33	ill(46),il(21),i'l(30)
39	THIS	598	0.31	dis(6)
40	WHEN	589	0.31	wen(75)
41	THERE	588	0.31	ther(2)
42	NIGHT	586	0.31	nite(105)
43	FROM	568	0.3	frm(9),fm(3)
44	HAD	562	0.3	hd(1),ad(2)
45	TOO	554	0.26	2(61)
46	COME	537	0.28	cum(8),com(2)
47	WELL	520	0.27	wel(5)
48	ONE	514	0.27	1(88)
49	SORRY	502	0.26	sori(3),sos(12)

50	SOON	499	0.26	sn(6)
	STILL	461	0.24	stil(14)
	BEEN	446	0.23	bin(22)
	LATER	437	0.23	l8r(20),lata(5),l8er(2)
	TONIGHT	431	0.23	2nite(45),tonite(10),2night(12),2nigt(3)
	HELLO	419	0.22	hallo(1),ello(9),allo(5),helo(6),herro(3)
	THANKS	412	0.22	thanx(32),thx(5),sanks(2),thnx(2),thks(1),thanxs(1),tnx(2)
	WEEKEND	393	0.21	wknd(2),wkend(48),w'end(10),w'kend(3)
	COULD	380	0.2	cud(48),cld(19)
	VERY	380	0.2	v(126)
	WEEK	380	0.2	wk(60),weeek(2)
	YOU'RE	376	0.19	ur(177),yr(4),u'r(1),ure(1),you'r(2),youre(6)
	SOME	373	0.2	sum(10)
	TEXT	369	0.19	txt(94),tx(6),tex(2)
	LOVE	368	0.19	luv(46),lv(4),lov(4)
	HERE	367	0.19	ere(3)
	WOULD	366	0.19	wud(38),wld(22)
	CAN'T	363	0.19	cant(121),cannot(3)
	TODAY	349	0.18	2day(44),dey(2)
	I'VE	342	0.18	ive(37),iv(5),i'v(3)
	THO	332	0.17	though(145),although(8),thou(1),altho(2)
	HAPPY	330	0.17	hapy(2),appy(1)
	REALLY	329	0.17	realy(5)
	BECAUSE	306	0.16	cos(226),coz(24),cause(4),cuz(3),cs(2),cus(2),cz(2)
	ANY	305	0.16	ne(8)
	COMING	301	0.16	comin(23), cumin(6)
	NEXT	300	0.16	nxt(29)
	THAT'S	295	0.15	thats(129)
	OH	292	0.15	o(30)
	LAST	284	0.15	lst(2)
	SHOULD	277	0.15	shud(11),shld(7),shd(3)
	PHONE	268	0.14	fone(16)
	WHERE	267	0.14	wher(4), wer(1)
	DOING	252	0.13	doin(33)
	AFTER	251	0.13	afta(1)
	PLEASE	250	0.13	pleasey(4),plez(3),pls(35),plse(3),plz(4),pse(2)
	DIDN'T	249	0.13	didnt(42)
	GREAT	249	0.13	gr8(3),g8(3)
	CALL	237	0.12	cal(3)
	BIRTHDAY	232	0.12	bday(28),b'day(10)
	GIVE	231	0.12	giv(10),gv(1)
	HAVING	230	0.12	havin(24),avin(6)
	DONE	227	0.12	dun(1)
	SAT	224	0.12	saturday(86),satdy(3)
	THEM	222	0.12	em(25)
	EVENING	221	0.12	eve(54),evo(3)
	UNTIL	221	0.12	til(104),till(70),untill(1)
	GONNA	220	0.12	gunna(11),gona(6)
	FRIDAY	220	0.12	fri(79)
	HOWS	217	0.11	how's(102),howz(8)
	MORNING	216	0.11	mornin(23),morn(18)
	CHRISTMAS	214	0.11	xmas(68),crimbo(2),chrimbo(2)
	MIGHT	214	0.11	mite(9),myt(6)
	SPEAK	206	0.11	spk(8)
	OVER	199	0.1	ova(1)
	MATE	193	0.1	m8(10)
	BEFORE	191	0.1	b4(29),befor(2)
	BETTER	191	0.1	betta(2),beta(6),bettr(1)

	NUMBER	189	0.1	num(7),numba(3),no(50)
	SOMETHING	186	0.1	sth(9),somethin(8),summort(3),sumfing(2),summat(2),summing (2)
	HAVEN'T	178	0.09	havent(36),have'nt(2),havn't(2),havnt(8)
	WORKING	169	0.09	workin(22)
	YEAR	169	0.09	yr(4)
	ANYTHING	167	0.09	anythin(3)
	THOUGHT	157	0.08	thort(5)
	WE'RE	155	0.08	we'r(3),were (27)
	SEND	154	0.08	snd(1)
	GETTING	152	0.08	gettin(17),getin(1),geting(1)
	WHO	151	0.08	wo(2)
	WON'T	145	0.08	wont(45)
	LOOKING	143	0.08	lookin(18)
	SUNDAY	142	0.07	sun(54)
	WENT	140	0.07	wnt(1)
	FEEL	138	0.07	fil(1)
	MAYBE	137	0.07	mayb(3)
	THANK	136	0.07	tank(2)
	FIRST	135	0.07	1st(26)
	PROBABLY	133	0.07	prob(45),probaly(3)
	WHAT'S	132	0.07	whats(48),wots(17),wats(7),wot's(3)
	MONDAY	130	0.07	mon(37)
	HIS	129	0.07	hs(2)
	THURS	126	0.07	thursday(48),thrus(1),thu(6),thur(2)
	THING	123	0.06	thiing(2)
	BEING	122	0.06	bein(11)
	INTO	121	0.06	in2(3)
	LITTLE	121	0.06	liddle(1)
	LOOK	121	0.06	lk(1)
	MEETING	119	0.06	meetin(7)
	RIGHT	118	0.06	rite(3)
	FINISHED	114	0.06	finshed(2),finnished(2)
	MO	114	0.06	moment(41)
	WHICH	114	0.06	wich(2)
	THINKING	112	0.06	thinkin(14)
	WHY	112	0.06	y(7)
	AFTERNOON	110	0.06	avo(4),pm(25),arvo(3)
	MESSAGE	110	0.06	msg(7),mess(2)
	WATCHING	109	0.06	watchin(17)
	NEVER	107	0.06	neva(2)
	HOUR	106	0.06	hr(11)
	STUFF	105	0.06	stuf(1)
	FORWARD	104	0.05	fwd(5),4ward(11)

Appendix 5.3 *Respelling rules or patterns in CorTxt (alphabetical order, with all realisations of each rule)*

Rule	**Realisations**
	okie(12)
	yep(69)
	yeh(22)
	yea(15)
	yeah(382)
	yup(13)
	dey(2)
1 for one (/wʌn/)	1(88)
1st for first	1st(26)
2 for to (tu:)	2(690), 2mora(14), 2morrow(9), 2morow(4), 2morro(4), 2mrrw(4), 2moz(3), 2mrw(2), 2moro(42), 2nite(45), 2night(12), 2nigt(3), 2day(44), in2(3)
2 for too (tu:)	2(61)
4 for fore (/fɔ:/)	b4(29)
4 for for (/fɔ:/)	4(357), 4wd(11)
8 for at / ɛɪt/	l8er(2)
8 for ate / ɛɪt/	l8r(20) m8(10)
8 for eat /ɛɪt/	g8(3), gr8(3)
a for e (/ə/)	da(6)
a for er (/ə/)	afta(1), ova(1), betta(2), beta(6), beta(6), numba(3), neva(2), lata(5), neva(2)
a for ou (/ə/)	ya(256)
a for ow (/ə/)	2mora(14), tomora(3), mora(1)
a for to	amoro(2)
ah for o (/æ/, /ə/)	nah(18)
apostrophe omission	im(280), its(435), dont(198), ill(46), cant(121), ive(37), iv(5) thats(129), didnt(42), hows(107), havent(36), were (27), wont(45), whats(48), il(21), havnt(8), wots(17), dnt(4),youre(6),ure(1)
appellation of o	evo(3)
appellation of pe (/p/)	nope(19)
appellation of y	pleasey(4)
ar for a (/ɑ:/)	arvo(3)
at for thing	summat(2)
b for be (/bi:/)	b(375)
c for see (/si:/)	c(248)
c omission from ck	bak(38), bk(11), bek(2)
d for th (/ð/)	d(21), da(6), dat(4), wid(4), dis(6)
dd for tt	liddle(1)
double letter reduction	geting(1), getin(1), beta(6), 2mrw(2), tomorow(4), tomoro(10) 2morow(4), sori(3), mora(1), hapy(2), gona(6), il(21), i'l(30) wel(5), stil(14), wil(12), helo(6), realy(5), cal(3), 2mora(14), 2moro(42), amoro(2), tomora(3), stuf(1)
doubled letter	summort(3), summat(2), summing(2), untill(1), till(70)
e for a	bek(2)

e for ea (/i:/)	plez(3)
e for ou (/ə/)	ye(9)
er for or (/fə/)	fer(2)
er for or/our (/ə/ or /ɜ:/)	yer(23)
f for ph (/f/)	fone(16)
f for th	sumfing(2)
final letter omission (e)	d(21), th(8), hav(106), ar(2), ther(2), com(2), lov(4), i'v(3), iv(5), wher(4), giv(10), befor(2), we'r(3), you'r(2), mayb(3) cos(226), coz(24), cuz(3), cs(2), cus(2), cz(2), cum(8), dun(1), pls(35), plz(4), plez(3), sum(10), luv(46)
final letter omission (g)	comin(23) avin(6), cumin(6), goin(60), doin(33), havin(24), mornin(23), somethin(8), workin(22), anythin(3), gettin(17), getin(1), lookin(18), bein(11), meetin(7), thinkin(14), watchin(17),
final letter omission (t)	jus(18), tha(2), tex(2), tx(6)
final letters omission (misc)	thou(1), kno(2), thur(2), thu(6), d'(12), bac(7), o(30), an(19), n(182)
final syllables omission	tom(24), arvo(3), sat(115), tomo(361), eve(58), evo(3), fri(79), prob(37), prob, morn(18), sun(54), num(7), mon(37), thurs(69), thu(6), thur(2), mess(8), mo(73), avo(4), sth(9)
first letter omission (th, h)	em(25),tis(21), av(8), ave(5), ad(2), ello(9), allo(5), ere(3), appy(1), avin(6)
first syllable omission	morrow(6), mora(1), k(33), bout(123), though(145), thou(1) cos(226), coz(24), cause(4), cuz(3), cs(2), cus(2), cz(2) till(70), til(104)
i for ee (/ɪ/)	bin(22), fil(1)
i for y	sori(3)
ite for ight (/aɪt/)	mite(9), rite(3), 2nite(45), nite(105), tonite(10)
j for y	okej(1)
letter addition (o)	soo(1), sooo(5), soooo(1)
letter variants	xx(833), xxx(635), xxxx(11), xxxxxx (2), xoxox(1) xxxxx(3), xxxxxxx (42) okay(75), okey(27), hallo(1)
mid letter omission (h)	te(2), 2nigt(3), wat(37), wen(75), wo(2), wots(17) wats(7), wot's(3), wich(2)
mid letter(s) omission	bday(28), b'day(10), thx(5), summing(2), abt(4), thks(1) pse(2), shd(3), finshed(2), tnx(2), w'end(10), w'kend(3), wkend(48), bettr(1), havn't(2), summort(3), summat(2) summing(2), probaly(3), hve(6), havnt(8), fm(3)
mid syllable omission (ur)	satdy(3)
n for an;	n(182)
ne for any (ɛnɪ)	ne(8)
no for know (/nəʊ/)	no(57)

o appellation	avo(4), arvo(3)
o for a (/ɒ/)	wot(148), wots(17) wot's(3), woz(4)
o for au (/ɒ/)	cos(226), coz(24)
o for ough (/əʊ/)	tho(176), altho(2)
o for ow (/əʊ/)	2morro(4), 2moro(42), amoro(2), tomoro(10), tomorro(6)
or for ough (/ɔ:/)	thort(5)
or for our (/ɔ:/ /ə/)	yor(2)
ort for thing	summort(3)
r for are (/ɑ:/)	r(422)
rr for ll	herro(3)
s for th (/θ/ /s/)	sanks(2)
sos for sorry	sos(12)
space omission	tis(21)
t for th (/θ/ /t/) or mid letter reduction (h)?	tank(2)
typo: doubled letter	uu(2), finnished(2), weeek(2), annd(2), iin(2), tomorrrow(2) thiing(2)
typo: substitution	jurt(2), amd(2)
typo: transposition	ot(3), hte(3), adn(10), wiht(3), have'nt(2), thrus(1)
u for o (/ɒ/)	gunna(11), cuz(3), cus(2)
u for o (/u:/)	du(5)
u for o (/ʌ/)	cumin(6), cum(8), luv(46), dun(1), sum(10), summort(3) summat(2), summing(2)
u for oo (/ʊ/)	gud(40)
u for oul (/ʊ/)	cud(48), wud(38), shud(11)
u for you	u(3043)
ur for you're (/jɔ:/)	ur(177)
ur for your (/jɔ:/)	ur (286)
v for f	arvo(3)
v for f (/f/)	avo(4)
v for th (/ð/)	wiv(82)
v for very	v(126)
vowel addition (o)	goodo(3)
vowel omission (a)	tht(1), cn(1), bck(3), hd(1), thx(5), thnx(2), thks(1) tnx(2), lst(2), ws(3), bk(11)
vowel omission (e)	wknd(2), txt(94), tx(6), nxt(29), snd(1), wnt(1)

vowel omission (ea)	pls(35), plse(3), plz(4), pse(2), spk(8), yr(4)
vowel omission (e-a-e)	msg(7)
vowel omission (ee)	wk(60)
vowel omission (i – e)	gv(1)
vowel omission (i)	hs(2)
vowel omission (o)	2mrrw(4), 2mrw(2), dnt(4), nt(4), knw(1), hpe(2), nw(9) hw(2), frm(9), fm(3), cz(2), tmw(9)
vowel omission (o, e)	cs(2), lv(4)
vowel omission (oo)	gd(25), sn(6), lk(1)
vowel omission (or-ar)	fwd(5), 4wd(11)
vowel omission (ou)	cld(19), wld(22), shld(7), shd(3), hr(11), yr(13)
vowel omission (u)	jst(6)
x for christ	xmas(68)
x for k	thanxs(1)
x for ks (/ks/)	thx(5), thnx(2), tnx(2) thanx(32)
y for why (/waɪ /)	y(29)
yt for ight (/aɪt/)	myt(6)
z for rrow	2moz(3), tomoz(9)
z for s (/z/)	coz(24), cuz(3), cz(2) plz(4), plez(3), howz(8) woz(4)
z for th (ð)	ze(2)

Appendix 5.4 *The extent to which variants occur in previous texts*

Respelling form and function	**Extent to which reflects or extends existing patterns**
abbreviations	
a) Standard abbreviations (pm, no, 1st)	Y
b) apostrophe omission (im, its, dont ill)	N Occurs as competence or performance error and particularly confusion over *its* and plurals (Carney, 1994:80-81; Crystal, 2003: 203).
c) double letter reduction (2morow, wil, beta, gona, hapy)	Y brandnames (Carney, 1994:447) printers' abbreviation (Crystal, 2003:56; Baron, 2000)
d) mid letter(s) omission	
i. h (nigt, wat, wich, wen, tnx)	N although reflected in printers' abbreviation (Crystal, 2003:56; Baron, 2000)
ii. c from ck (bak)	Y reduction of consonant blends (Weber, 1986:417) or cluster simplification (Carney, 1994:447) in advertising;
iii. n (thx, thks)	Y – as above
iv. e (havnt), a (hve, yeh), ou (abt), i (finshed)	N although reflected in printers' abbreviation (Crystal, 2003:56; Baron, 2000) and by consonant writing
v. ee (wkend), eek (w'end),	Y
vi. lea (pse)	N
vii. shd	N
viii. bday, b'day	Y
e) final letter(s) omission	
i. e (sum, luv, ar)	Y advertising (Carney, 1994:447) printers' abbreviation (Crystal, 2003:56; Baron, 2000)
ii. h (o), k (bac), s (thur), w (kno)	Y bac = reduction of consonant clusters (Weber, 1986:417); kno can be compared with Tru-Blu (Carney, 1994:447)
iii. rs (thu), gh (thou)	N (only as extension of clipping practices)
f) final syllable(s) omission (fri, num, mess, eve, tomo, tom)	Y/N, although some are, others (num, mess, tomo, tom) are extensions of clipping practices) (Weber, 1986:415; keeping the first part is the 'commoner type' although the other 'chief type' is clipping the beginning' Crystal 2003:120)
g) initialisms (v, sth)	'The vast majority of abbreviations [across texts] fall into this category' (Crystal, 2003:120)
h) substitutions	
i. soz/sos for sorry	Y, although sos is variant of soz, recent online abbreviation (Urban Dictionary)
ii. z for rrow (tomoz)	Y, recent online abbreviation (Urban Dictionary)
i) standard abbreviations (pm, xmas, 1st)	Y
9. consonant writing	
a) a (tht, bck, hd, lst)	recognizes the information value of consonants Crystal (2003: 425) unique to text messaging? (Hard af Segersteg, 2002)
b) e (txt,next, snd, wnt)	
c) ea (pls, spk, msg,yr)	

d) i (gv, hs)	
e) o (2mrw, dnt, nt, hpe, nw)	
f) oo (gd, sn, lk)	
g) ou (cld, wld, shld, shd,hr)	
10. phonetic spelling	
a) number homophones	
i. 1 for one (/wʌn/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
ii. 2 for to (/tu:/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38) in advertising (Carney, 1994:448)
iii. 4 for fore (/fɔ:/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
iv. 8 for at, ate, eat / ɛɪt/	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38) in advertising (Carney, 1994:448)
b) letter homophones	
i. b for be (/bi:/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
ii. c for see (/si:/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
iii. r for are (/ɑ:/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
iv. u for you	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38) in advertising (Carney, 1994:448)
v. y for why (/waɪ /)	N (extension of existing practice) in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
vi. f for ph (/f/)	Y Androutsopoulous (2000:522) gives the example of 'phat' for 'fat'.
vii. ne for any (ɛnɪ)	N (extension of existing practice) in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
c) ur for you're and your (/jɔ:/) – not phonetic	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
d) other homphones	
i. ite for ight (2nite, rite, mite)	Y in advertising (Carney, 1994:449)
ii. yt for ight (/aɪt/) (myt)	N
iii. no for know (/nəʊ/)	Y in graffiti and other 'marginal practices' including CMC and text messaging (Sebba, 2007:38)
e) schwa represented	
i. a for er (/ə/) (lata, afta, ova, betta, numba)	Y <a> as schwa as common practice since Early English (Weber, 1986:424)

f) letter o	
i. o for a (/ɒ /) (wot, wots, wot's, woz, coz, cos)	Y in advertising and graffiti (Crystal, 2003; Sebba, 2007: 34)
ii. o for ow (/əʊ/) (tomoro)	N although o is alternative spelling of /əʊ/ in final position: compare *tomorrow* and *mango* (Carney, 1994:171)
g) letter u	
i. u for au (/ɒ/) (cuz)	Y Weber (1986) suggests 'wuz' (423) and 'frum' (418)
ii. u for o (/ɒ/) (gunna)	Y Weber (1986) suggests 'wuz' (423) and 'frum' (418) – not clear if phonetic or colloquialism.
iii. u for o (/u:/) (du)	Y Weber suggests *tu* and *yu* (1986:418)
iv. u for o (/ʌ/) (cumin, cum, luv, dun)	Y Weber (1986:418) suggests *cum*; Kessler and Bergs (2003) note *luv* in Victorian letters
v. u for oo or oul (/ʊ/) (gud, cud, wud, shud)	Y gud is noted as recent respelling in Urban Dictionary; but the others extend this pattern.
h) other vowel sounds	
i. ar for a (/ɑ:/) (arvo)	Y ar corresponds with /ɑ:/, while 'arvo' is an Australian colloquial contraction of afternoon (Urban Dictionary).
ii. or for ough (/ɔ:/) (thort)	Y mentioned by Sebba (2007:34)
iii. or for our (/ɔ:/ /ə/) (yor)	N not mentioned by Weber (1986) but or corresponds with /ɔ:/
i) consonants	
i. x for ks (/ks/) (thx)	Y in fanzines (Androutsopoulous, 2000)
ii. z for s (/z/) (coz, cuz, plz)	Y in advertising (Crystal, 2003:; Carney, 1994:447) graffiti (Sebba, 2007:34)
iii. doubled letter (m) (summat, which shortens the previous vowel sound)	Y
j) eye dialect or clipping? (tho, altho)	Y tho is a recent, mainly online abbreviation (Urban Dictionary); altho an extension of this.
11. visual morphemes	
a) xx, xxx, xxxx, xoxo	Y Noted in Victorian love letters by Kessler and Bergs (2003)
12. colloquial contractions	
a) first letter omitted	
i. h (ad, ell, ere, avin)	Y in novel dialogue, magazines and songs as reflected in personal pronouns (Weber, 1986:)
ii. th (em)	Y in novel dialogue, magazines and songs as reflected in personal pronouns (Weber, 1986:)
iii. a (n)	Y

b)	first syllable omitted (though, cos, till, bout, k, morrow, mora)	Y in novel dialogue, magazines and songs Weber (1986:423) mentions 'cause and 'bout.
c)	mid syllables omitted (satdy)	Y *satdy* is an Australian colloquial contraction (Urban Dictionary)
d)	mid letters omitted (probaly)	Y recent colloquial contraction (Urban Dictionary)
e)	weak sounds represented	
	i. shwa (/ə/)	
	1. a for e (/ə/) (da)	Y common practice since Early English (Weber, 1986:424)
	2. a for ou (/ə/) (ya)	Y in novel dialogue, magazines and songs common practice since Early English (Weber, 1986:424)
	3. a for ow (/ə/) (tomora)	Y in novel dialogue, magazines and songs common practice since Early English (Weber, 1986:424)
	4. a for to (amora)	Y
	5. d for th (/ð/) (d, dat, wid)	Y Sebba (2007: 37) mentions *di* for *the* in Creole graffiti: 'runnin' tings as usual like di original Rude Gal does!' Caribbean poetry (eg, Linton Kwesi Johnson, 1975)
	6. e for ou (/ə/) (ye)	Y in novel dialogue, magazines and songs Earlier spelling of schwa, predating a and the most frequent spelling of you (Weber, 1986:424)
	7. er for our (/ɔ:/ /ə/) (yer)	Y in novel dialogue, magazines and songs <er> is
	ii. (/ɪ /)	
	1. i for ee (/ɪ /) (fil, bin)	Y in novel dialogue, magazines and songs
f)	f for ph (sumfing)	Y reflects pronunciation in SE England, eg: Crancher, S (1993) *Dija wanna say sumfing?: a guide to Estuary English*.
g)	t for th (/θ/ /t/) (tank)	Y Sebba (2007:37) mentions *tings* for *things* as Creole respelling Irish or American colloquial contraction (Urban Dictionary)
h)	z for th (ð) (ze)	Y – as a reflection of a French accent
i)	final letter omitted	
	i. d (an, n)	Y in novel dialogue, magazines and songs <d> omitted after <n> (Weber, 1986:423)
	ii. e (th, d)	Y in novel dialogue, magazines and songs
	iii. g (avin, goin, mornin, somethin)	Y in novel dialogue, magazines and songs (Weber, 1986:422)
	iv. t (jus, tha, tex)	Y similar to <d> omission mentioned by Weber see extract from *The Snapper* (Doyle 1993)
j)	final syllable omitted (prob, morn, mo, avo,	Y *avo* and *arvo* are colloquial Australian contractions

arvo)	(Urban Dictionary)
k) tis	Y Archaic or poetic form ‘adopted in verse to fit meter’ (Weber, 1986:419)
l) something (summort, summat, summing)	Y *summat* is a colloquial Northern English (Yorkshire?) contraction and *summort* a West Country version; no evidence found of *summing*. ‘There are few definite rules for suggesting dialect through spelling’ (Weber, 1987:420)
13. colloquial respelling	
a) letter addition	
i. p (yep, nope)	Y
ii. o (soo, goodo)	Y
iii. y (pleasey)	Y
b) ah for o (/æ/, /ə/) (nah)	Y
c) er for or (/fə/) (fer)	Y in novel dialogue, magazines and songs
d) dd for tt (liddle)	Y
e) rr for ll (herro)	Y reflection of South East Asian pronunciation of English (Urban Dictionary)

Appendix 5.5 *Combined forms in CorTxt*

Headword	**Respelling**	**Forms**	**Functions**
THE	d(21)	d for th (/ð/) final letter omission (e)	colloquial contraction + colloquial contraction
	da(6)	d for th (/ð/) a for e (/ə/)	colloquial contraction + colloquial spelling
HAVE	av(8)	first letter omission (h) final letter omission (e)	colloquial contraction + abbreviation
TOMORROW	tomora(3)	double letter reduction a for ow (/ə/)	abbreviation + colloquial contraction
	mora(1)	first syllable omission double letter reduction (r) a for ow (/ə/)	colloquial contraction + abbreviation + colloquial contraction
	2mora(14)	2 for to (/tu:/) double letter reduction a for ow (/ə/)	phonetic spelling + abbreviation + colloquial contraction
	tomoro(10)	double letter reduction (r) o for ow (/əʊ/)	abbreviation + phonetic spelling
	2morow(4)	2 for to (/tu:/) double letter reduction	phonetic spelling + abbreviation
	2morro(4)	2 for to (/tu:/) o for ow (/əʊ/)	phonetic spelling + phonetic spelling
	2mrrw(4)	2 for to (/tu:/) vowel omission (o)	phonetic spelling + consonant writing
	2moz(3)	2 for to (/tu:/) z for rrow	phonetic spelling + abbreviation
	2mrw(2)	2 for to (/tu:/) vowel omission (o) double letter reduction (r)	phonetic spelling + consonant writing + abbreviation
	amoro(2)	a for to double letter reduction o for ow (/əʊ/)	colloquial contraction + abbreviation + phonetic spelling
	2moro(42)	2 for to (/tu:/) double letter reduction o for ow (/əʊ/)	phonetic spelling + abbreviation + phonetic spelling
DON'T	dnt(4)	vowel omission apostrophe omission	consonant writing + apostrophe omission
I'LL	il(21)	double letter reduction (l) apostrophe omission	abbreviation + apostrophe omission
SORRY	sori(3)	double letter reduction (r) i for y	abbreviation + colloquial spelling
TONIGHT	2nite(45)	2 for to (/tu:/) ite for ight	phonetic spelling + phonetic spelling
	2nigt(3)	2 for to (/tu:/) mid letter omission (h)	phonetic spelling + abbreviation
THANKS	thx(5)	vowel omission (a) mid letter omission(n) x for ks (/ks/)	consonant writing + abbreviation + phonetic spelling
	thnx(2)	vowel omission (a) x for ks (/ks/)	consonant writing + phonetic spelling
	thks(1)	vowel omission (a) mid letter omission (n)	consonant writing + abbreviation

	tnx(2)	vowel omission (a) mid letter omission (h) x for ks (/ks/)	consonant writing + abbreviation + phonetic spelling
TEXT	tx(6)	vowel omission (e) final letter omission (t)	consonant writing + colloquial contraction?
I'VE	iv(5)	apostrophe omission final letter omission (e)	apostrophe omission + abbreviation
ALTHOUGH	tho(176)	first syllable omission o for ough (/əʊ/)	colloquial contraction + phonetic spelling
	thou(1)	first syllable omission final letters omission (gh)	colloquial contraction + abbreviation
BECAUSE	cos(226)	first syllable omission o for au (/ɒ/) final letter omission (e)	colloquial contraction + phonetic spelling
	coz(24)	first syllable omission o for au (/ɒ/) final letter omission (e) z for s (/z/)	colloquial contraction + phonetic spelling + phonetic spelling
	cuz(3)	first syllable omission u for au (/ɒ/) final letter omission (e) z for s (/z/)	colloquial contraction + phonetic spelling/colloquial spelling + phonetic spelling
	cs(2)	first syllable omission vowel omission (o, e) final letter omission (e)	colloquial contraction + consonant writing

Appendix 5.6 Competing variants in CorTxt

Headword	Competing variants
you	u(3043), ya(256), ye(9), yer(14)
the	d(21),da(6),th(8),hte(3),te(2),ze(2)
for	4(431), fer(2),
have	av(8), hve(6), ave(5), hav(106),
are	r(422), ar(2),
you're	ur(177), yr(4), u'r (1), ure(1), you'r(2), youre (6)
your	ur (286), yer(23), yor(2), yr(13), u'r(2), ure(1)
good	gud(40), gd(25), goodo(3),
just	jus(18), jst(6),
that	tht(1), dat(4), tha(2),
do	d'(12), du(5),
was	woz(4), ws(3),
tomorrow	tomoz(9), tomorro(6), tomorow(4), tomora(3), tomo(361), morrow(6), mora(1), tom(24), 2mora(14), tomoro(10), 2morrow(9), tmw(9), 2morow(4), 2morro(4), 2mrrw(4), 2moz(3), 2mrw(2), amoro(2), 2moro(42),
what	wot(148), wat(37),
know	no(57), knw(1), kno(2),
about	bout(123), abt(4),
it's	its(435), tis(21),
had	hd(1), ad(2),
come	cum(8), com(2),
tonight	2nite(45), tonite(10), 2night(12), 2nigt(3),
hello	hallo(1), ello(9), allo (5), helo(6), herro(3),
thanks	thanx(32), thx(5), tank(2), sanks(2), thnx(2), thks(1), thanxs(1), tnx(2),
weekend	wknd(2), wkend(48), w'end(10), w'kend(3),
could	cud(48), cld(19),
text	txt(94), tx(6), tex(2),
love	luv(46), lv(4), lov(4),
would	wud(38), wld(22),
I've	ive(37), iv(5), i'v(3),
happy	hapy(2), appy(1),
because	cos(226), coz(24), cause(4), cuz(3), cs(2), cus(2), cz(2),
coming	comin(23), cumin(6),
should	shud(11), shld(7), shd(3),
please	pleasey(4), plez(3), pls(35), plse(3), plz(4), pse(2),
great	gr8(3), g8(3),
give	giv(10), gv(1),
Saturday	sat(115), satdy(3),
gonna	gunna(11), gona(6),
Christmas	xmas(68), crimbo(2), chrimbo(2),
before	b4(29), befor(2),
better	betta(2), beta(6), bettr(1),
something	sth(9), somethin(8), summort(3), sumfing(2), summat(2), summing(2),
haven't	havent(36), havn't(2), havnt(8),
we're	we'r(3), were (27),
getting	gettin(17), getin(1), geting(1),
number	num(7), numba(3),
What's	whats(48), wots(17), wats(7), wot's(3),
message	msg(7), mess(8),
afternoon	avo(4), pm(25), arvo(3),
the	d(21), da(6), th(8), te(2), ze(2)
and	n(182), an(19),
having	havin(24), avin(6),
morning	mornin(23), morn(18),

later	l8r(20), lata(5), l8er(2),
might	mite(9), myt(6),
Thursday	thurs(69), thu(6), thur(2),

Appendix 7.1 *The non-grouped wordlist*

N	Word	Freq.	N	Word	Freq.
1	YOU	4,560	14	ME	1,555
2	TO	4,283	15	ON	1,523
3	I	4,257	16	OF	1,420
4	A	3,580	17	AT	1,393
5	THE	3,511	18	MY	1,285
6	U	3,043	19	BUT	1,257
7	AND	2,956	20	JUST	1,214
8	IN	2,385	21	GOOD	1,197
9	X	2,161	22	BE	1,192
10	IT	2,020	23	SO	1,152
11	HAVE	1,868	24	IF	1,125
12	FOR	1,698	25	WILL	1,112
3	IS	1,577			

Appendix 8.1 *Clause combination**

Clause types: Free (F), Bound (B), Non-clausal element (X)

Sentence types: Simple (S), Compound (CP), Complex (CX)

	Texts from 1400 to 1435**	**Clause**	**Sentence**	
ONE ONE-CLAUSE UNITS (no relationships between clauses) 19/100	Bad light stops play	F	S	simple only
exclamation	Sorry sent that 2 wrong person!	X F	 S	simple only
question	Where r u?	F	S	simple only
e	Have you seen who's back at Holby?!	F	S	simple only
e	I have lost 10 kilos as of today!	F	S	simple only
q	are u still at work?	F	S	simple only
statement	Pow@830	X		x
e	Ooh thank you! Xx	X X X		x
q	How is he?	F	S	simple only
s	We're here. Xxx	F X	S	simple only
e	Now thats going to ruin your thesis!	F	S	simple only
e	sort it out!	F	S	simple only
e	power to the people, sister!	F	S	simple only
e	gud mornin !!!	F	S	simple only
s	Yes, my reg is REGNUMBER. Ciao!	X F X	 S	simple only
s	Ok, see then.	X F	 S	simple only
s	Currently test driving electric car	F	S	simple only
e	Great!	X		x
q	Have you been paid for the aspire day yet boyo x	F X X	S	simple only
TEXT-SPECIFIC 38/100	Have you tried any other companies? We have just emailed one 4 a quote.	F F	S S	simple only
	Bugger me. not again. are there not any cheaper alternative	F ? F	S S	simple only
	Got extra back dated payment pf £1000. Yippee!	F X	S	simple only
	I am back home alone and all sad! Are you back yet?	F F	S S	simple only
	Its bloody snowing. Wey hey	F X	S	simple only
	Its bloody snowing. Wey hey	F X	S	simple only
	Its bloody snowing.	F	S	simple only

	Wey hey	X		
	That's a shame! Maybe cld meet for few hrs tomo?	F F	S S	simple only
	fabtastic. i'll be round for coffee at quarter to half past 0 oclock in the afternoon.	X F	 S	simple only
	Monk yogarht, pasta in fridge. Enjoy!	? F	 S	simple only
	Sorry. mail? POSTALADDRESS	X F F	 S S	simple only
	Wow v v impressed. Hav fun shopping!	X ? F	 S	simple only
	No c him last nite at work txt u later x	X F F X	 S S	simple only
	Ok, that's fine! See ya later. xx	X F F X	 S S	simple only
	Not thought bout it... Drink in tap & spile at seven. Is that pub on gas st off broad st by canal. Ok?	F F F X	S S S	simple only
	lhey, did you have a good time in poragué? I'm in ireland at the mo with the in laws! Going ok .\ yes love being on hol 2!	X F F F X F	 S S S S	simple only
	Hi. Do u want to join me with sts later? Meeting them at five. Call u after class.	X F F F	 S S S	simple only
	No i'm not gonna be able to. too late notice. i'll be home in a few weeks anyway. what are the plans	X F np F F	 S S S	simple only
	Bout ye chief. How many euro's you taking for recreational expenses?! Getting some today. Was thinking 200..	? F F F	 S S S	simple only
	having lunch with NAME78 and baby NAME51 . You'll have a nice time when you eventually get there. See you next week	F F B F	S CX2 S	simple and complex
	House OK! Ours since june 9th... Looking 4ward 2 C-ing U guys... Cheers	F F F B X	S S CX	simple and complex

	my friend	X		
	Hope ur head doesn't hurt 2 much ! Am ploughing my way through a pile of ironing ! Staying in with a chinky tonight come round if you like.	F B F F F B	 CX2 S CX2	simple and complex
	Hi. Hope ur day ok. Table booked for 8.15. Call u later. F x	X F B F F X	 CX2 S S	simple and complex
	just seen another bridgy weirdo, carrying his portable stero down his trousers! this place is full of em! hope ur havin fun x	F B F F B X	 CX2 S CX2	simple and complex
	Don't worry I'll tell her you'll be late. Too late to change. She's at work.	F F B F F	 CX3 S S	simple and complex
	Hello! All well hope you are ok too. Looking forward to seeing you . R u driving or training?	X F F B F F	 CX3 S S	simple and complex
	i'm not sure how to break this to you. there's no easy way to put it.... i can't make the friday-night fun. sorry. however, feel free to text me during the evening if there are any lulls in conversation. anyway, hope ur exotic trip goes well. see u next term. sv	F B F B F X F B B F B F X	 CX2 CX3 CX3 CX2 S	simple and complex
	Hello NAME39, what time r u thinking about leaving? Happy 2 come 2 NAME12's not going to the yard though. R u sure u r going back later?	X F F B F? F B	 S CX3 CX2	simple and complex
	Well in that case i'm busy. Only joking! I'm Very funny... Yep i'm coming –	X F F F X F	 S S S 	simple and complex

	it'll be good to catch up. what's the plan stan?	F B F	 CX3 S	
	Its not quite that good yet. We are going snowboarding in about 5 weeks though. Can't bloody wait. Whats the weather like where you are	F F F F B	S S S CX2	simple and complex
	Good news indeed. But Prodi doesnt seem that good too. In England,its pretty bad too. Blair is a complete idiot. Im gonna b more and more involved in political issues. Im already a unionist,but i think next year i'll b a member in a political party. Its called Ligue Communiste Revolutionnaire. I'll tell u more about it when we'll c each other. Xx	X F F F F F F B F F B X	 S S S S CX3 S CX2	simple and complex
	Old library, no food today. Meet outside front. Any suggestions where we can go	np np F F B	 S CX2	simple and complex
	you shouldn't be looking for excuses! Hasn't that been the pattern recently crap weekends? Am on my way to show dad my photos. Now get on with that marking young lady!	F F F B? F	S S CX2 S	simple and complex
	God it's really getting me down just hanging around. I hate not knowing. What did NAME39 say? How would she know? They'd be crazy not to take u. I'm worried x	X F B F F F F B F X	 CX2 S S S CX2 S	simple and complex
	Sounds good – you can help me look around for nina's pressie – lucky you! And clothes, obviously! See you xx	F F B X np F X	 CX3 S	simple and complex
	Hope you had a good time last night .\ Bet NAME65 is ill. Why didn't NAME209 go? Just a text	F B F B F F	 CX2 CX2 S	simple and complex

	to ask you who did your bathroom and how much did it cost?	B B B	 CX4	
	I've got a lovely new phone. Its cool. might actually be the first one that i truly like. also it connects to my computer and allows me to take my texts off. **but** i haven't worked out how to get them from the software to excel. still trying though	F F F B F F F B F	S S CX2 CP2 CX2 S	simple and complex
	I've got a lovely new phone. Its cool. might actually be the first one that i truly like. also it connects to my computer and allows me to take my texts off. but i haven't worked out how to get them from the software to excel. still trying though	F F F B F F F B F	S S CX2 CP2 CX2 S	simple and complex
Combining features of text- and speech-like units 21/100	Dad's address is POSTALADDRESS. Must get some cards myself spose xx	F F B X	S CX2? 	simple and complex
	Hi- how's the analysis going? Typical- we've been sweltering at school all week- but today is cold and wet! Pooh- that means no excuse not to get on with marking!	X F X F F X F B	 S CP2 CX2	simple complex and compound
	What's the date of NAME209's birthday? Have been sale shoppin ran into NAME84!..... 30 mins later I escape!	F F F F	S CP3	simple and compound
	A little girl is lost and crying in asda! The security man asks "Wats ur mum like?" She replies "big cocks and vodka!"	F F F F F	S S S S S	simple only
	5000 israeli troops have just entered jordan. early reports say she is tired and her xxx's a bit sore.	F F B B	S CX3	simple and complex
	Don'tworry "I just phoned to say I love you .."	F F B F	S CX2 S	simple and complex
	yes... staff-room can be pretty unbearable sometimes!	X F F	 S S	simple only

	c ya next week NAME242. Sx	X		
	Yeah i'm good. All the better for knowing that your back in the country.	X F F B	 S CX2	simple and complex
	tomorow 1230am other wise i think i will be onnit	? F B	 CX2	complex
	sorry, no, have got few things to do. may be in pub later.	X[1] X F F	 S S	simple only
	Hi. Yes interested, perhaps later time but meet bit before for drink?	X X adj F F	 CP2	compound
	NAME78, NAME416, NAME366 and NAME417 picking them up from various points going 2 yeovil and they will do the motor project 4 3 hours and then u take them home. 12 2 5.30 max. Very easy x	X F F F F ? adj X	 CP4	compound
	Do u want an afternoons work on monday, very easy and whilst they r doing the activity u can read a book, let me no thanks x	F adj B F F X X	 CX4	complex
	don't worry I'll add a bit of bling. What u wearing?	F F F	 CP2 S	simple and compound
	Bloody answer ur phone! Been to Cardiff and got the rest of my stuff. Just wondering what will happen with my bond, etc. Have a good week. X	F F F F B F	S CP2 CX2 S	simple complex and compound
	Argh! We're running late, lunch taking a while!	X F F	 CP2	simple and compound

[1] X = Elements outside the clause structure, including pragmatic markers:

- *discourse markers* which are 'words or phrases indicating boundaries in the discourse or words that monitor the state of the discourse in some way', such as *well* or *you know*;
- responses and response tokens (*yes, certainly* or *fine*);
- vocatives;
- *stance markers* such as *sorry, to be honest* and *frankly* (which indicate 'a stance or attitude to a segment or section of discourse') [or are these *viewpoint adjuncts* (592)?]; and
- *interjections* (see 224) such as *hey* and *gosh*.

	Should be there by 3! V sorry, F x	F X X	S	
	me cool u nah, where u been hiding? 2 months left of uni so me well busy wid cswork and ting. but me good. u good?	F F F F F F F	 CP3 CP3 S	simple and compound
	Have hit some hideous traffic no idea why. Will be late.	F F F	 CP2 S	simple and compound
	I am goin gto turkey on friday! Hope u r havin a good holiday did u get chance to pay that money in coz its not in my bank yet?	F F B F B B	S CX2 CX3	simple and complex
	Sorry, only just saw u called – was out walking. Hope party ok yesterday. Fraid have to do things here this afternoon so can't meet, but will be in town tomo if ur free.	X F B F F B F B F F B	 CX3 CX2 CX5	complex
	ooo! Yeah, am ok but temp up again this morn + can't stop coughing. Haven't slept so much in ages! Plan to go in tomo tho. See ya soon. X f	X X F F F F F F X	 CP3 S S S	simple and compound
SPEECH-LIKE 22/100	Why london? We have just had lunch but mum only had coke	F F F	S CP2	simple and compound
	Got stuff i want to get too – tho really shldn't! See u at bull at five. F x	F B B F X	 CX3 S	simple and complex
	Er yeah, i will b there at 15:26, sorry! Just tell me which pub/cafe to sit in and come wen u can	X F X F B F B	 S CX4	simple and complex
	I'm in solihull, do you want anything?	F F	 CP2	compound
	Oh babe	X		complex

	I am sure he is missing u 2 – I miss u loads xxx	F B F X	CX3	
	Dunno, haven't spoke to NAME10 about it yet.	X F	S?	simple only
	It's a toss up between that and going for a balti with cherry blossoms. Not sure what would be more fun.	F F B	S CX2	simple and complex
	Yes, that's fine – was thinking browse round shops or sth but back iin time for me sun dinner. See you tomo. F x	X F F F F X	 CP4 S	simple and compound
	the more holiday photos, anecdotes etc the better, naturally! anyway, i'm slaving away, as per so better stop. see ya soon tagliatelli! SV	F B F F F X	 CX2 CP2 S	simple and complex
	Hi, got the access code to the room, eight more signed up, and i emailed them the questionnaire, shall we meet after my supervision on thursday?	X F F F F	 CP4	compound
	I've ain't BEEN doing much, just working. Yeah We're going to mosley but i'm gonna drive, We're broker than broke.	F B X F F F	 CX2 CP3[1]	complex and compound
	Hurrah it's all over adn we're sitting outside the local. NAME68 says he's not able to stay long though so no plans for this evening as yet.	X F F F B F	 CP2 CX3	complex and compound
	Dude, if you see NAME9 then dont mention that you've spoke to me, she wants me to go to the pub and i cant be bothered and my mom goes on holiday tomorrow.	X B F B F B F F	 CX7	complex
	no reason ... just didn't like to think of u	? F B		complex

[1] Clauses joined by 'comma splice', not by coordinator – 'contact clauses' according to Carter and McCarthy (2006: 556): 'Coordination may occur without the explicit link created by a conjunction. This is particularly true in literary style or for special effect in narrative texts. Coordinated clauses simply placed next to each other without conjunctions are also known as contact clauses'. This includes where 'clauses are coordinated by the use of commas alone'.

	sittin in staff-room alone all afternoon! x	B X	CX3	
	What i mean was i left too early to check, cos i'm working a 9-6.	W F B B	CX3	complex
	Yeah that would be good … I'll phone you tomo lunchtime, shall I, to organise something?	X F F B B	CX4	complex
	Hiya – thinking about weekendand realised I was assuming you would stay Saturday night too – you will , won't you? (tho no prob if you can't).	X F F B B F B B B	CX8	complex
	I think your mentor is NAME71 SURNAME14, but not 100 percent sure.	F F	CP2	compound
	I had to resist the spotty mugs i got you! They've got lovely mugs, lovely coloured seives. x	F B F F X	CX2 CP2	complex and compound
	Am not working but am up to eyes in philosophy so will text u later when a bit more free for chat...	F F F B	CX4	complex
	I 4got NAME67 is playin 2. So u've def got 5 at mo n NAME58 tryin sort a shooter	F B F F	CX2 CP2	complex
	Hi- seems a small group will be going out on fri- i'll book a table for a meal at 7.30- so try to get the earlier train and pack light if possible x	X F B F F F B X	CX6	complex

*The messages above were ranked initially according to whether they were intuitively felt to be speech-like. Having (through this exercise) determined criteria for a speech-like clause combination (e.g., lengthy units and parataxis) and established the criteria for a contrasting category of text-specific messages, the original ranking was adjusted accordingly. Although messages have largely been designated to the text-specific or speech-like category, they are roughly ordered from very text-specific to very speech-like.

**The text messages used in this analysis were chosen by ordering the corpus according to time of day that message was sent. The above text messages were all sent between 1400 and 1435 hours.