A Causal Classification of Orthography Errors in Web Texts

Mirko Tavosanis
Università di Pisa
Dipartimento di Studi italianistici
Via del Collegio Ricci 10
I-56126 Pisa PI Italy
tavosanis@ital.unipi.it

Abstract

Errors, even at the spelling level, can provide useful insight into the nature of a written text. This paper presents a classification of spelling errors in Web texts based on their causes (misspellings, typos and intentional deviations), linking them to the attitudes of their authors and the circumstances of their writing. Examples are drawn from blog and forum entries in English and Italian.

1. Introduction

Blogs and other genres of Web texts are often considered hastily written and 'full of errors'. However, this claim should probably be qualified. Preliminary findings [Tavosanis, 2006] show, for example, that at least in some situations, real misspellings in blogs are only as frequent as in online newspapers edited by professional journalists. The 'noise' in many Web texts can, in fact, often be ascribed to stylistic choices, rather than to any particular shabbiness of writing. It may also be concentrated in specific kinds of text, while other, more professional types may be almost noise-free.

As a preliminary step to further analysis, the present paper aims to classify the deviations from received orthography found in blogs and electronic texts in Latin -alphabet languages. Such deviations follow different patterns and have different causes; the paper groups them into three general causal categories: misspellings, typos (further broken down in three subcategories) and intentional deviations (broken down into three subcategories). Each kind of deviation is assigned a code, and a sample XML-TEI encoding of text errors is proposed.

Although the classification has yet to be applied to a proper corpus, preliminary samplings do hint at its relevance to real-world situations.

1.1 Current classifications

It is interesting to note that such a causal classification involves categories slightly different from those used in many current classifications.

The traditional distinction between typographic errors, cognitive errors and phonetic errors is recalled in [Kukich 1992: 387], while [Ringlstetter et al. 2006: 297] describes four error classes: typing errors, spelling errors, errors resulting from inadequate character encoding, and OCR errors. Both classifications exclude intentional deviations (though a brief discussion of the relation between errors and special vocabulary is provided in [Ringlstetter *et al*., 2006: 297 and 311]) and both take into account only the mechanical causes of typos, excluding the psychological ones.

Reconsidering the overall classification of errors could therefore have interesting consequences in the assessment of error percentages in different kinds of texts.

1.2 Criteria

Regarding the samples discussed in the paper, unless otherwise indicated, the errors have been drawn from English and Italian blog and forum entries included in the iJCai-2007 Weblog data collection.

All deviations have been considered, irrespective of their automatic recognizability (incorrect word forms coinciding with correct forms of another word are still difficult to detect for both human beings and computers). Moreover, features of punctuation and the use of uppercase / lowercase were not taken into account.

2. Misspellings: errors committed through ignorance of orthography

Most languages with an alphabetic writing system have an orthography: a 'correct' way of writing words (the problem of concurrent orthographies of entire languages or single words will be not dealt with here; any word spelling will be assumed to be 'correct', if acknowledged as such by at least one authoritative source, such as a dictionary, even if other sources classify it as a misspelling).

In traditional learning models, it is assumed that learners know how to speak the language that they must write. If the writing standard of every language had a one to one correspondence between phonemes and graphemes (or graphemes + diacritics), spelling would represent a trivial task: knowing the correct pronunciation of a word and the graphemes of an alphabet, it would be intuitive and easy to infer the correct spelling of any word (for a critical discussion of this general idea, see anyway [Harris, 1986, 1995,] and in particular [Harris, 2000]).

However, only a small group of written languages can boast such a total or near total correspondence between phonemes and graphemes (such scripts may use Latin-based alphabets, such as in Swahili and Tagalog, or other alphabets, and spelling errors are reported as quite rare in them). Other languages (e.g. Italian or German) have a good, but incomplete correspondence; many spelling errors by beginners are therefore concentrated in the areas of reduced correspondence. Several languages, including English, have a largely conventional orthography. Given the many different ways of writing the same sound, knowing the correct pronunciation of a word often does not give speakers enough clues as to how to write it. This is a first source of errors in orthography and is particularly relevant for the English language (code: *misspelling*).

In depth study of spelling errors has been conducted particularly with regard to the process of learning English orthography, as in the classification by [Gentry, 1982], which describes five learning stages: precommunicative; semiphonetic; phonetic; transitional; conventional. The transition from the phonetic to the conventional level can however be found in many different languages. In Italian orthography, for example, the *q* letter is not related to a phoneme of its own (it is used to transcribe the /*k*/ phoneme, which is however more often transcribed with the letter *c*); traditional teaching of orthography in Italian primary schools therefore concentrates on the 'correct' (i.e. conventional) use of *q* and *c* in words like *acqua, squarcia, cuore* [Sabò, 2005: 107].

Misspellings of words in blogs and other Web genres are usually committed simply because the writer does not know the 'right way' to spell them. It is interesting to note that such errors are particularly revealing because they are usually *not* committed intentionally (only small subsets of them are committed for stylistic purposes; see § 4).

Misspellings can therefore provide useful indications about the level of formal instruction of a particular writer or a community. In the following samples such deviations from the standard are marked in italics:

```
I'm neither Taoist, Jungian, nor Platonist
enough to ascribe it with equanamity [i.e.,
equanimity] to echoes of any sort of
universal thing or experience
```

```
nel tardo pomeriggio fra scaramuccie [i.e.,
scaramucce] loro e malesseri della madre, un
gruppo di una 40ina di persone del quartiere
e? riuscito a rientrare nella casa
```

2.1 Linguistic interference

A particular case of misspellings stems from linguistic interference. When the authors are not native speakers of the language they are writing, the standards and customs of their mother tongue may appear in their writing. Basic errors, which are easy to avoid for L1 writers, are therefore likely to crop up in texts written by L2 writers with far superior knowledge of other aspects of the language – or, at least, this is a current postulate of contrastive linguistics (see [Bebout 1985]).

Little research is however available at the moment to quantify this phenomenon, which though perhaps significant in global electronic writing, seems reduced in school contexts. Indeed, a survey in a Californian school revealed that "students from five different language backgrounds [including English] demonstrated remarkably similar patterns of [English] spelling development" [Tompkins *et al.*, 1999: 16].

If they actually do occur with significant frequency, anyway, such errors should follow detectable patterns, and their identification should enable, for example, determining the L1 of a particular writer.

2.2 Unspoken words

Similar errors should also appear in words lacking spoken referents. Modern languages have rich written uses, and many words, especially in specialized technical language, are currently used in writing, but seldom spoken aloud. In this case mistakes are likely to stem from simple ignorance of the exact spelling of words, without implications regarding the relation between sounds and letters. However, even in this regard, we have (somewhat surprisingly) little data to substantiate such a postulate.

3. Typos: errors committed for mechanical or psychological reasons

Simple typos are very common in keyboard-entered texts, especially if the text is entered by unskilled typists. Moreover, in electronic texts (especially in real-time communications and email), such errors are generally and explicitly tolerated, and the pressure to correct them is correspondingly low. Such errors therefore provide little useful information about the writers' linguistic skills. They may instead reveal a good deal about typing and editing skills, the time allocated to writing and revision, and the tools used.

Current research indicates that this kind of error is by far the most common: [Ringlstetter *et al.* 2006: 314] attributes 93.5% percentage of the errors in an English corpus to simple typing errors and explains the corresponding figure of 55.7% in a German corpus as due to the addition of mechanical problems (mainly character conversion).

Three subtypes of errors can be included in this category:

1. psychological slips of pen or keyboard (code: *typoPsyc*)
2. mechanical typos (code: *typoMech*)
3. typos stemming from limitations of a technical nature (code: *typoLim*)

3.1 Psychological slips of pen or keyboard

The first kind of errors is due to limits to attention span. Even users with a good knowledge of the orthography of a language may produce many errors due to psychological causes. Such causes are often difficult to identify precisely, as in the following examples:

I got to hang out in my *chones* [i.e. *clothes*] all day today
[the author presumably knows the correct form of the word because the error is basic and other sections of the text reveal a good command of orthography]

what if I had one of those *recharcheable* [i.e. *rechargeable*] heart thingies or something and the battery ran out
[the author presumably knows the correct form of the word because in the next sentence the word *recharge* is spelt correctly]

This category, even in electronic communication, includes the traditional gamut of *lapsus* in manuscript texts, as studied for centuries by philology, physiology and psychology (including one of the most influential and controversial works of the 20th century, Sigmund Freud's *Psychopathology of Everyday Life*). Unintentional duplication or deletion of sections of a text, *saut du même au même*, polar errors, and so on, are among the most common types and can be found in every kind of writing (especially if it involves copying instead of direct composition), including handwriting.

Interpretation of such errors is particularly difficult and is best approached by framing the circumstances of the text creation (for classifications more strictly linked to writing and linguistic competence see [Fromkin, 1980; Timpanaro, 2002]). Moreover, in a short excerpt of text it is often impossible to establish whether the cause of an error is insufficient knowledge of spelling or a simple slip of the hand (subtype 2), as in the following case:

we are going to focus on how this *inconvieneces* [i.e. *inconveniences*] us, selfish Americans.

It is in any event interesting to note that this kind of error is often omitted in technical discussions (in particular, in [Kukich 1992, Ringlstetter *et al.* 2006]). In [Ringlstetter *et al.* 2006: 297], it is noted that "focusing on garbled standard vocabulary, tokens may be seriously damaged in an 'unexplainable' way", but no samples are provided, and such damaged tokens seem to exclude further evaluation, while, according to the authors, "most [i.e., not all] of the remaining errors can be assigned to one of the four classes" used in the paper, i.e. typing errors, spelling errors, errors resulting from inadequate character encoding and OCR errors.

3.2 Mechanical typos

The second kind of typo is typical of keyboard writing and is simply due to punching the wrong keys, leading to accidental transpositions, deletions, substitutions or insertions of characters (see in particular [Kukich 1992]). Often it involves adjacent keys, as in:

I can not believe that i sat through the *shole* [i.e. *whole*] thing.

The most typical feature of this kind of typo is the unintentional inclusion in a word of numbers, punctuation marks and so on (such errors are explicitly excluded from discussion in [Ringlstetter *et al.*, 2006: 303]), as occurred in this blog post with the number *5* in the word *that*:

and was surprised noone asked any questions in any of the papers in the states (tha5t I could see..granted I live in the UK). http://dilbertblog.typepad.com/the_dilbert_blog/2006/09/appease_on_eart.html

Of course, distinguishing between some such errors and more complex slips of the hand or errors committed through ignorance can often be difficult or altogether impossible. In

any event, such typos area strict related to typing and editing skills and with the time devoted to writing. Their presence outside of "stylish" uses (see § 4) suggests that the text has been hastily written and/or poorly edited.

3.3 Typos stemming from limitations of a technical nature

The third kind of typos consists of unwanted substitutions of characters correctly entered by the writer. Such substitutions are frequent in blogs and web texts: a writer may type an orthographically correct text only to discover that the publishing system used cannot handle the special characters or diacritics of a particular alphabet and cancels them or substitutes them with random characters (for this kind of problem, with particular regard to German diacritics, see [Ringlstetter *et al.* 2006: 307-308]). Generally, writers quickly seem to become aware of such problems and can develop complex strategies to avoid them, often following explicitly developed conventions, which cannot be considered 'errors' (see § 4). In this class of typos we can also include erroneous outcomes of OCR (see in particular [Ringlstetter *et al*. 2005] and [Ringlstetter *et al.* 2006: 305-307]). Such errors are in fact completely external and mechanical and cannot therefore provide any useful insights as to the competencies of the writers or the circumstances of the writing.

4. Intentional deviations

Deviations from standard are often actively sought in online communications. There seem to be three basic causes for this:

1. stylistic requirements (code: *styleDev*)
2. desire to overcome limitations of a technical nature (code: *limDev*)
3. desire to deliberately circumvent or 'fool' automatic indexing mechanisms (code: *fooDev*)

4.1 Stylistic requirements

The use of deliberate deviations from orthography for stylistic purposes is not a new fact. In modern times, many 'misspellings' have become standard ways, for example, to distinguish fictional texts as sub- or non-standard: English spellings such as *tonite,* instead of *tonight,* or the Italian *squola,* in place of *scuola,* are typical examples of this.

Nowadays, one of the most extreme uses in this direction is the so-called *leetspeak* or "elite speak", still very popular in electronic communication and the online gaming world (for a short description of the linguistic features of leetspeak see [Microsoft, 2006]). This language is distinguished by distortions of the written form of words. The most conspicuous facet of such distortions is the substitution of a letter with a number or a symbol with a similar shape: for instance, the name *leet speak* itself can be written as *1337 5p34k* , where the number *3* substitutes the *E* (a short list of substitutions is given in [Blashki and Nichol, 2005: 80], while [Leet, 2006] provide more exhaustive coverage). This kind of mechanical play is further complicated by the use of different phonetic solutions for English word spellings (see § 2): *you* can be replaced by *joo* or by *j00*. Other features typical of leetspeak include the use of abbreviations, particular suffixes, and the substitution of *–z* for *–s*:

```
I bet a "certain government agency" is
feeling pretty silly at turning away someone
with my l33t sk11lz now, eh?
```

Most interestingly, some common mechanical typos are used as standard forms in leetspeak. Words like *teh* (instead of *the*) and *pwned* (instead of *owned*) are among the few standard features of this kind of written language (see also the Google home page translated into leetspeak: http://www.google.com/intl/xx-hacker/). It is also possible to create closed lists of some of these errors.

However, leetspeak guides include explicitly spontaneous typing errors as one of the features of the language. [Blashki and Nichol, 2005: 83] do suggest, although without a true linguistic analysis, that "many of the words used in Leet and gaming language are originally derived from incorrect spelling generally due to speed of typing, and then deliberately and repeatedly used as incorrect". In other languages seems that typical misspelling and typos are deliberately avoided, and that only intentional deviations from standard are admitted.

In any case, such a technique is, of course, the product not of a lower-than-average knowledge of a language, but of a superior one. It can then be assumed that leetspeak writers have a good command of at least some complex graphic conventions.

4.2 Desire to overcome limitations of a technical nature

The good command of graphic conventions required by leetspeak relates it to the second kind of voluntary deviations, which does not exist in the English language. These deviations stem from the technological limitations inherent in the transcription of special characters or diacritics (see § 3), not included in the restricted ASCII character set.

Wrong handling of those characters is still commonplace and it prompts user to develop various substitution techniques, such as the vowel+apex sequence used in Italian to replace accented letters in email and electronic writing ([Pistolesi, 1997] and, in a more complete way, [Pistolesi,

2004] describe this kind of substitution in electronic communication outside of Web pages, such as in e-mails and chats; for documentation of the situation in English writing, see in particular [Baron, 1998] and [Crystal, 2006]). Moreover, differences in keyboards can preclude composing a particular text at all. For example, American or English keyboards do not have accented letters; this makes it hard or impossible for many users to use them to write a Spanish or French text following standard orthography.

The forum *La meglio gioventù,* published in 2004 by the Web site of the Italian newspaper *La repubblica,* exhibits many examples of substitution techniques. The forum has seen wide participation by Italians living abroad, and many orthographic deficiencies in the texts can therefore be explained by the use of non-Italian keyboards (e.g. keyboards without accented characters), and not by any lack of competency of the writers.

The following quotation is typical of this kind of problem. The original post comes from a writer living in England; all accented letters are replaced by the sequence letter + apex:

Non ho potuto vedere il film *perche'* [i.e. *perché*] non ho accesso ai canali RAI in questi giorni e la cosa mi rattrista molto.Penso che la mia meglio *gioventu'* [i.e. *gioventù*] sia legata al momento in cui ho cominciato a decidere da sola.

Also in such cases, the non-standard solution does not imply ignorance on the part of the writer. It instead hints at a particular competence: knowledge of shortcuts to overcome the limits of the interface, creative solutions to graphical problems and so on.

4.3 Desire to deliberately circumvent or 'fool' automatic indexing mechanisms

Lastly, in some cases non-stylistic and unnecessary deviations from standard are purposefully sought for. Web page developers may try to attract traffic to their sites by including in them spelling mistakes in order to raise the rankings of their pages in search engines results. Here, the purposely introduced misspellings correspond (or are thought to correspond) to common spelling errors committed by users in their queries or to the results of spelling correction routines used by the search engines themselves.

Most of these deviations are hidden in Web pages metatags or in sections of the text made invisible to human readers. Here is a particularly elaborated example of this:

eneric ggeneric geeneric genneric geneeric generric generiic genericc generic viagra vviagra viiagra viaagra viaggra viagrra viagraa viagra generic eneric gneric geeric genric geneic generc generi generic viagra iagra vagra vigra viara viaga viagr viagra

generic g eneric ge neric gen eric gene ric gener ic generi c generic generic viagra v iagra vi agra via gra viag ra viagr a viagra viagra generic egneric gneeric geenric genreic geneirc generci generic generic viagra ivagra vaigra vigara viarga viagar viagra viagra

Lips ,tongue ,or troleandomycin TAO br middot nasal congestion br middot an antifungal medication such as alprostadil Caverject ,Muse ,Edex or yohimbine Yocon ,Yodoxin ,others ,isosorbide dinitrate Dilatrate-SR ,Isordil ,Sorbitrate ,and swelling of the lips ,**generic viagra**tongue ,or you may read .Doctor .Do not take **generic viagra** Viagra ,tell your doctor .P p br What happens if I miss a dose .generic viagraP p p p br What other drugs will affect Viagra ?Br Your pharmacist has additional information about Viagra ?Br Viagra is used to treat impotence ,such as Peyronie's disease br middot temporary blue tint in vision generic viagra or other vision abnormalities or br middot have a history of heart failure br middot the HIV medications amprenavir Agenerase ,delavirdine Rescriptor ,indinavir Crixivan ,nelfinavir Viracept ,ritonavir Norvir ,or you may require generic viagra a dosage adjustment or special monitoring during treatment if you are taking any of the lips ,tongue

http://www.fecaltransfusionfoundation.org/anyboard9/forum/uploads/generic-viagra.html

Conversely, writers may try to disguise the true nature of their text in order to avoid censoring and filtering. Techniques for this latter type of deviation often recall leetspeak solutions and seem particularly widespread in email spam, as in this case, taken from the web archive of a mailing list:

Buy Xâ.NâXmg 30 tâblets for only $119.95 37% DìSCOUNT - Overnìght!
pPHENTERMìNE for weìght loss [âppetìte suppressânt].
We got Generjc Vvìâgrâ™ 100 mg wìth 55% SâVìNGS (Lìmìted Supply âvâìlâble).
http://bmrc.berkeley.edu/mhonarc/openmash-cvs/msg03688.html

Such deviations are also of methodological interest because they aim to circumvent some particular mechanism of automatic indexing, but at the same time try to be easily readable by human beings.

5. TEI encoding of orthographic errors and deviations: a preliminary proposal

The importance and diffusion of the Text Encoding Initiative standard suggest that the creation of a TEI-compliant corpus of errors and deviations could be an useful step in the study of the orthography of Web texts (such a project is now in planning phase at the University of Pisa). Pending full definition of TEI P5 (see [TEI, 2006]), TEI – P4 provides native support for encoding some categories of errors. However, standard TEI – P4 elements are only partially adequate to the task of encoding variations in orthography. The <corr> element is e.g. supposed to include "text reproduced although apparently incorrect or inaccurate"; the <abbr> element includes "an abbreviation of any sort". However, only some of the variations discussed here are abbreviations (such as *2morrow* for *tomorrow* in leetspeak, and so on). On the other hand, a text may be incorrect or inaccurate even if it contains no deviations from standard orthography: to write "pencil" instead of "paper" disrupts the sense of a text, but not its orthography.

It thus seems useful, in order to properly encode the entire range of errors discussed, to use the TEI element <w>, which "represents a grammatical (not necessarily orthographic) word".

Using <w>, the standard attribute *type* can be applied to describe the class of the deviation, following the classification given above (standardized values could be: *misspelling*, *typoPsyc*, *typoMech*, *typoLim*, *styleDev*, *limDev, fooDev*). Another useful attribute is the *lemma,* which identifies the word's lemma. A possible TEI encoding of some of the errors discussed so far could thus be:

```
nel tardo pomeriggio fra <w
type="misspelling"
lemma="scaramuccia">scaramuccie</w> loro e
malesseri della madre, un gruppo di una
40ina di persone del quartiere e? riuscito a
rientrare nella casa

what if I had one of those <w
type="typoPsyc"
lemma="rechargeable">recharcheable</w> heart
thingies or something and the battery ran out

I can not believe that i sat through the <w
type="typoMech" lemma="whole">shole</w> thing

nel tardo pomeriggio fra scaramuccie loro e
malesseri della madre, un gruppo di una
40ina di persone del quartiere <w
type="typoLim" lemma="essere">e? </w>
riuscito a rientrare nella casa

I bet a "certain government agency" is
feeling pretty silly at turning away someone
with my <w type="styleDev"
lemma="elite">l33t</w> <w type="styleDev"
lemma="skill">sk11lz</w> now, eh?

Penso che la mia meglio <w type="limDev"
lemma="gioventù">gioventu'</w> sia legata al
momento in cui ho cominciato a decidere da
sola

<w type="fooDev"
lemma="Phentermine">pPHENTERMìNE</w> for <w
type="fooDev" lemma="weight">weight</w> loss
[<w type="fooDev"
lemma="appetite">âppetìte</w> <w
type="fooDev"
lemma="suppressant">suppressânt</w>]
```

For particular kinds of processing, the <w> element could be given an additional attribute, *reg,* for the regularized form of the word (such as *skills* for *sk11lz*), and the values of the *type* attribute could be inserted into the TEI DTD, following the TEI rules for such extensions, as in:

```
reg CDATA #IMPLIED
type (misspelling | typoPsyc | typoMech |
typoLim | styleDev | limDev | fooDev)
#IMPLIED
```

6. Conclusions and future developments

'Errors' are not a homogeneous set. Some types of errors are linked to limitations in the author's skills or knowledge; others to stylistic choices, or time constraints on composition, and so on. Analyzing and understanding misspellings, typos and deviations can provide useful insights into the true nature of Web texts.

In many cases there is of course no automatic means to ascribe a particular deviation to one of the three categories indicated. The insertion of a letter within a word could be either a mechanical typo (category 2) or a conscious choice (category 3). An individual word may be misspelled because the writer does not know its correct form (category 1), or due to a simple slip of the hand (category 2). Such judgements (when possible) have been made by human beings since the beginnings of modern philology, and to date there seems no alternative to this, admittedly imperfect, procedure. However, the availability of large corpora tagged by human beings (such as the project hinted at in § 5) could help lead to the development of more sophisticated automatic tools to this end.

As a further development, detailed descriptions of the particular features of recently developed text types, such as blogs, could be exploited to better characterize and identify them. This should, in particular, allow for identifying new feature set components useful for automatic genre

classification. Although lexical features are already a significant component of such classifications (see, in particular, [Santini, 2006a-f], [Santini *et al*., 2006]), consideration of the role of errors may provide a sounder basis for determining the true genre of any given text.

References

[Baron, 1998] N. S. Baron. Letters by phone or speech by other means: the linguistic of email. *Language & Communication*, 18:133-170, 1998.

[Bebout 1985] Linda Bebout. An error analysis of misspellings made by learners of English as a first and as a second language. *Journal of Psycholinguistic Research*, (14)6:569-593, 2005.

[Blashki and Nichol, 2005] Katherine Blashki and Sophie Nichol. Game Geek's Goss: Linguistic Creativity in Young Males Within An Online University Forum (94/\/\3 933k'5 9055oneone). *Australian Journal of Emerging Technologies and Society*, (3)2:77-86, 2005.

[Crystal, 2006] David Crystal. *Language and the Internet* (second edition). Cambridge University Press, Cambridge, 2006

[Fromkin, 1980] *Errors in linguistic performance. Slips of the Tongue, Ear, Pen, and Hand*. Edited by Victoria A. Fromkin. Academic Press, New York *et al.*, 1980.

[Gentry, 1982] Richard J. Gentry. Developmental Spelling: Assessment. *Diagnostique* (8)1:52-61, 1982.

[Harris, 1986] Roy Harris. *The Origin of Writing*. Duckworth, London, 2000.

[Harris, 1995] Roy Harris. *Signs of Writing*. Routledge, London & New York, 1995.

[Harris, 2000] Roy Harris. *Rethinking Writing*. The Athlone Press, London, 2000.

[Kukich, 1992] Karen Kukich. Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, 24(4):377-439, 1992.

[Leet, 2006] *Leet* voice in Wikipedia (http://en.wikipedia.org/wiki/Leet), 2006

[Microsoft, 2006] *Leetspeak: A parent's primer to computer slang. Understand how your kids communicate online* in the Microsoft web site (http://www.microsoft.com/athome/security/children/leetspeak.mspx), 2006.

[Pistolesi, 1997] Elena Pistolesi. Il visibile parlare di IRC (Internet Relay Chat). *Quaderni del Dipartimento di linguistica – Università di Firenze*, :213-246, 1997.

[Pistolesi, 2004] Elena Pistolesi. *Il parlar spedito. L'italiano di chat, e-mail, SMS*. Esedra, Padova, 2004.

[Ringlstetter *et al.*, 2005] Christoph Ringlstetter, Klaus U. Schulz, Stoyan Mihov and Katerina Louka. The Same is Not The Same - Postcorrection of Alphabet Confusion Errors in Mixed-Alphabet OCR Recognition. In *Proceedings of the 8th International Conference on Document Analysis and Recognition ICDAR'05*, pages 406-410, 2005.

[Ringlstetter *et al.*, 2006] Christoph Ringlstetter, Klaus U. Schulz, and Stoyan Mihov. Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. *Computational Linguistics*, 32(3):295-340, 2006.

[Sabò, 2005] Lisa Sabò. *La competenza linguistica al termine della scuola dell'obbligo: analisi ortografica e linguistica (il sistema dei pronomi)*, Degree Dissertation, Pisa University, 2005.

[Santini, 2006a] Marina Santini. Common Criteria for Genre Classification: Annotation and Granularity. In *Workshop of Text-based Information Retrieval*, In Conjunction with ECAI 2006, Riva del Garda, Italy - Aug 29th, 2006.

[Santini, 2006b] Marina Santini, Some issues in Automatic Genre Classification of Web Pages. In *JADT 2006 - 8èmes Journées internationales d'analyse statistique des données textuelles* du 19 au 21 avril 2006 à l'université de Besançon (France).

[Santini, 2006c] Marina Santini, Identifying Genres of Web Pages. In *TALN 2006 - Traitement Automatique des Langues Naturelles*: du 10 au 12 avril 2006 à Leuven (Belgique) / *Natural Language Processing*: April 10-12, 2006 in Leuven (Belgium).

[Santini, 2006d] Marina Santini, Interpreting Genre Evolution on the Web. In *EACL 2006 Workshop: NEW TEXT - Wikis and blogs and other dynamic text sources*, Trento, 4th of April 2006, pages 32-40, 2006.

[Santini, 2006e] Marina Santini, Web pages, text types, and linguistic features: Some issues. *ICAME Journal*, 30, 2006.

[Santini, 2006f] Marina Santini, From Biberian text types to genres of web pages: An overview of studies on automatic genre identification. In *GENRE TEXTUEL/DOMAINE/ACTIVITÉ*. Toulouse, 30 et 31 mars 2006, Journées d'étude organisées par l'opération «Sémantique et Corpus» / *TEXTUAL GENRE/FIELDS/ACTIVITY*. March 2006, 30th and 31th - Toulouse, France - Workshop organised by the «Sémantique et Corpus» group

[Santini *et al.*, 2006] Marina Santini, Richard Power, Roger Evans. Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages. In *COLING - ACL 2006*, Sydney (Australia) 17-21 July, 2006 - Poster Session.

[Tavosanis, 2006] Mirko Tavosanis. Are Blogs edited? A linguistic survey of Italian blogs using search engines. In *Proceedings of the Computational Approaches to analyzing weblogs Conference*, pages 211-213, Stanford,

March 2006. AAAI.

[TEI, 2006] The Text Encoding Initiative Web site (http://www.tei-c.org/), 2006.

[Timpanaro, 2002] Sebastiano Timpanaro. *Il lapsus freudiano. Psicanalisi e critica testuale*. Edited by Fabio Stok. Bollati Boringhieri, Torino, 2002.

[Tompkins *et al.,* 1999] Gail Tompkins, Shareen Abramson, and Robert H. Pritchard. A multilingual perspective on spelling development in third and fourth grades. *Multicultural Education*, 6(3):12–1, 1999.