Introducing Mini-McCALL: A pilot version of the Mid-Sweden Corpus of Computer-Assisted Language Learning

Mats Deutschmann, Annelie Ädel, Gregory Garretson and Terry Walker
Mid-Sweden University

1 Introduction

In this paper, we present Mini-McCALL, a 1.3-million-word corpus of computer-mediated communication in the context of online English university courses.[1] The data consist of three types of written communication – both between students and between students and teachers – in English: discussion forum messages, e-mail messages, and documents handed in as assignments. This pilot corpus comprises the first stage of a proposed 10-million-word corpus of computer-assisted language learning based on the online English courses offered by the Department of Humanities at Mid-Sweden University (Mittuniversitetet).

In what follows, we first consider e-learning – online, off-campus study, where the medium of instruction and communication involves computer technology – from a theoretical perspective, and the need for such a corpus as ours to facilitate research into this new learning environment, as well as into the language used in e-learning. We then describe the structure and content of Mini-McCALL and highlight both the research potential of the material and studies currently underway, as well as looking forward to the future development of the full Mid-Sweden Corpus of Computer-Assisted Language Learning (McCALL). Both Mini-McCALL and the ultimate McCALL corpus will be made freely available to the research community.

2 Collaborative e-learning

The demands of an increasingly digitalised global 'learning society' have caused higher education to undergo a number of radical changes in recent decades. The accessibility of higher education has increased, and e-learning, in particular, has played a central role in this development. Internet-based courses,

in which all communication takes place in an online environment, are now available in almost every subject domain.

Parallel to these developments, there has been a move away from behaviouristic models in favour of the 'learning community model' in higher education. This model has at its core situated learning theory and social/cultural learning, with the purpose of enabling students to support each other as well as to collaboratively learn with, and from, one another (see e.g. Wenger 1998; Palloff and Pratt 1999; Olofsson and Lindberg 2005). Again, e-learning has been part of this development. Spurred on by advances in information and communication technology, online learning environments no longer merely offer ready-made educational material to be downloaded – nowadays, these environments enable students to learn together in a social context (see e.g. Koschmann 1996; Bonk and Cunningham 1998; Stephenson 2001; Haythornwaite 2002). As Kern and Warschauer (2000: 11) put it, technology usage has moved on from software that involves "learners' interaction with computers to interaction with other humans via the computer". In computer-supported collaborative learning (CSCL, Salmon 2004), the focus is on student-driven participation and collaboration.

These ideas have been implemented in Internet-based courses in English Studies offered by the Department of Humanities at Mid-Sweden University. Since autumn 2004, in addition to traditional classes, the department has offered full-time undergraduate English courses taught entirely over the Internet with no physical meetings, in which the learning management system WebCT is the primary platform of communication. The courses are designed from a collaborative learning perspective, each one incorporating methods such as peer-reviewing, group discussions, and group problem-solving, resulting in high levels of student-student and teacher-student interaction.

The following exchange between students in a discussion forum, extracted from Mini-McCALL, illustrates collaborative online learning in action. Here the task is to identify, correct and explain language errors in a number of sentences, and post the answers to the discussion forum for peer-review. Example (1) shows portions of a student's discussion-forum post and the peer-review from another student. Note that although the participants' names appear to be genuine, they have in fact been constructed – the anonymisation process is explained in Section 4.3.

(1)

```
Message VT05S.D.242
Time: Saturday, February 19, 2005 09:20
Author: Yasmin Strömvall
Subject: Yasmin Strömvall


I just want to appologize for being late, but I have sick kids here,
and they have been sick for three days now. I have done the
assignment and I will try to get time to comment during the day.

[...]

I have done it as good as I can, but I have problems finding the
exact explanations and the right terms to why I do the changes I do.

[...]

[Example] 3) I won't be neither sad nor surprised.
Double negative. You can't use "not" in front of "neither – nor", but
you can in front
of "neither – or"
I will be neither sad nor surprised
or
I won't be neither sad or surprised.

Hope I understood this whole thing correct :)
Yasmin!!!
_____________________________________________________________


Message VT05S.D.257
Time: Tuesday, February 22, 2005 09:34
Author: Sofia Ullman
Subject: Re: Yasmin Strömvall


Hi!
I've read your answers and first off I must say that I agree with
you, this assignement took some time... I was late too.. Your answers
were great, it's always good to read someone elses answers too
because it makes it easier to understand. The only thing I was
wondering about was the dubble negative. I don't think you can use
neither together with won't, I think it has to be either and or if
you use that. But with will, neither and nor works fine.
See you:)
Sofia
```

Collaborative learning places great demands on social skills. In order for the learning process to be effective, students need to define their roles, create trust, and negotiate shared goals (Palloff and Pratt 1999). This is achieved through language. For instance, we see in (1) mutual displays of consideration for the other – e.g. apology, counter-apology, foregrounding of positive feedback, and hedging of criticism – all of which help to protect face and thus maintain trust. In e-learning, the absence of face-to-face interaction, and therefore the lack of important communication aids such as intonation, eye contact and body lan-

guage, means that non-referential meanings instead have to be embedded in the written text. Also in (1), for example, we see how 'emoticons' can be added as a substitute for facial expressions. We return to these and other strategies in Section 5.2.

3 The call for McCALL

As early as 1995, Dillenbourg *et al*. (1995: 198) pointed to the fact that "research on the pragmatics of communication remains to be exploited in the field of collaborative learning". However, to date, relatively little linguistically-oriented research focusing on collaborative processes in e-learning environments has been carried out. Not much is known about the crucial role played by language in this new mode of education, nor about potential problems posed specifically in a second language and/or cross-cultural context of this kind.

There are indications, however, that such research could be very rewarding, in part with regard to cross-cultural pragmatics, and perhaps even more so as a tool for analysing the processes of collaboration and knowledge construction in a cross-cultural setting. For example, in her analysis of collaborative learning through the medium of e-mail correspondence, Tella (1992) was able to show that phatic use of language was essential to communication. In another early study of language in spoken collaboration, Schegloff (1991) analysed 'next turn proof procedures', which refers to how speakers, by examining the recipient's answer, receive confirmation of his/her understanding of the utterance. Schegloff argued that such signals are an important resource for managing intersubjectivity and joint understanding in talk. Through this and other related practices such as repair, participants continually construct and display their understanding, and negotiate their positions in the interaction. In a more recent study of synchronous and asynchronous computer-mediated discussions, Sotillo (2000) was able to show that the nature of the medium affected the use of discourse functions – categories central to the negotiation of meaning – such as requests, responses, apologies, greetings, complaints, and reprimands. Discussions in asynchronous communication contained significantly fewer such items. Sotillo (2000: 96) also showed that online discussions contained a high frequency of speech acts associated with social interaction, such as greetings, jokes and apologies.

Considering the growing interest in e-learning on the part of teachers, students, and academic institutions, it is becoming increasingly urgent that this type of learning environment be researched – to map and evaluate not only the efficiency of online communication in such contexts, but also the learning pro-

cesses going on in these types of environments. Mini-McCALL, providing a valuable source of data from computer-mediated communication, is an important step forward: not only has there previously been "a disregard for CMC [computer-mediated communication] in corpus building" (Beißwenger and Storrer 2008: 305), but it is also the case that no other corpus resources involving CMC learner data exist to date.

4 The corpus

Mini-McCALL is based on Internet courses taught entirely in English at Mid-Sweden University. The data are collected from a full-time, five-week course on English grammar given to students taking their first term of English Studies at the undergraduate level. The data come from eight individual classes (or sections), two from each term over the course of four terms (autumn 2004 to spring 2006). Because they are from eight sections of the same course, the data from all groups are highly comparable. As mentioned above, Mini-McCALL is the first stage of a larger project, the final product of which will be a much more comprehensive corpus, the Mid-Sweden Corpus of Computer-Assisted Language Learning (McCALL), consisting of the entirety of Mid-Sweden University's online English courses from 2004–2008. This project is described in more detail in Section 6 below.

4.1 The types of text

The classes represented in the corpus took place entirely within the learning management system WebCT. Mini-McCALL contains texts of three types: discussion forum messages, e-mail messages, and documents containing students' assignments. Because these last were sent in as attachments to the e-mail and discussion forum messages, we refer to them as 'attachments'. Currently, the corpus comprises 1,262,775 words, broken down by type of text as shown in Figure 1. All told, the discussion fora contain 1,193 threads consisting of 4,179 messages. The e-mail communication contains 4,603 threads consisting of 5,927 messages. Distributed over the discussion and e-mail messages are 477 attachments.

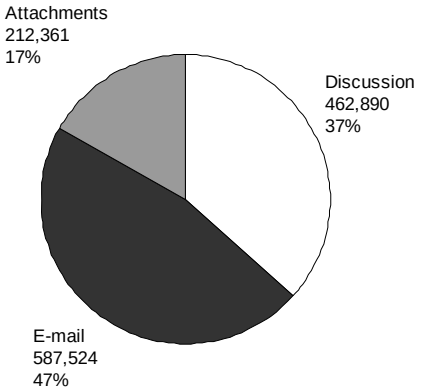

Figure 1: Composition of the corpus by type of text (word counts and percentages)

The discussion forum data consist primarily of peer-review tasks in which students work in groups of four. In each group, every student is required to provide individual solutions to and analyses of grammatical problems, and then comment on the work of the other three students. Most of the communication concerns four obligatory collaborative tasks on grammar dealing with (a) article usage, (b) verb tense and aspect usage, (c) structural levels (word classes and clause elements), and (d) a 'spot the error' task, as in (1) above. In addition, there is one discussion thread connected to a warm-up task entitled "What is your experience of grammar?", in which students write about their prior experiences of the subject. The fact that students are asked not only to give individual solutions, but also to comment on each other's work and to discuss and justify their criticism, leads to high levels of student-student interaction, as the examples in this paper testify.

The e-mail data represent student-teacher communication only, that is, e-mails sent from students to teachers or teachers to students. No student-student e-mails have been included in the corpus, out of respect for the students' privacy. The e-mails are concerned with many different topics; as discussed below, there are many announcements sent in bulk by the teachers, but there are also many threads in which students report to the teacher or ask questions, as in (2):

```
(2)   Message HT04N.E.3
      Time: Tuesday, September 7, 2004 19:06
      Author: Lilja Sjögren
      Recipient: George Sederstedt
      Subject: Assignments!


      Hello George!

      I have been in and out of the discussion room a couple of times and
      it seems that noone is there to discuss the assignment that is due on
      friday. I need some tips how to start the discussion. I have never
      done this before. And what do I do if noone is discussing before
      friday?

      Lilja Sjögren
      ____________________________________________________


      Message HT04N.E.5
      Time: Wednesday, September 8, 2004 11:50
      Author: George Sederstedt
      Recipient: Lilja Sjögren
      Subject: Re: Assignments!


      Dear Lilja,

      For the discussion assignments you actually make your comments in the
      discussion forum, which is not a simultanous chat. You simply mail
      your contribution and wait for someone to respond. The mails can be
      read by all. In the chat you can talk more informally. Get back to me
      if this is still unclear.

      George
```

One potentially debatable aspect of Mini-McCALL's compilation is the large number of duplicate e-mails included, resulting from mass-mailings in which a teacher sent the same message to each student in the class. This has two negative effects: artificially inflating the proportion of contributions by the teachers, and causing great redundancy in the language in the corpus. However, we have opted not to delete such duplicate messages, but rather to tag them with an XML attribute ('duplicate'), for two reasons. First, their presence in the corpus allows researchers to keep track of which messages a given student has received. Second, retaining these messages facilitates tracing the various parallel threads of communication in the corpus. For example, it often happens that a single mass-mailing will prompt responses from several different students; by retaining all messages, we keep each thread intact.

The third type of text in the corpus is the attachments, each of which originally accompanied a particular discussion forum message or e-mail message, and most of which were originally MS Word documents. These mainly consist of documents with students' individual solutions to the tasks or answer keys sent

by the teacher. Each such document is clearly cross-referenced with regard to the message it accompanied (and vice versa), making it easy for the corpus user to refer to the corresponding text; in fact, in the HTML version of the corpus (see below), the documents are hyperlinked to each other. Example (3) shows the beginning of a typical attachment document:

(3)
```
Document ID: VT06S.D.16.1
Attached to message: VT06S.D.16
Time: Tuesday, January 31, 2006 15:17
Author: Lina Holmström
Subject: Determiners and pronouns


Discussion 1: Determiners and Pronouns


1. Discuss the use or absence of the article in the following
sentences:


a. I met an interesting chap at a party last night.


Comment: The speaker uses indefinite articles to "an interesting
chap" and "a party last night". This can mean that the referents
(chap and party) are not known to the hearer.


b. Why are you still in bed? You should be at school.


Comment: We can say "the bed" when we mean a particular piece of
furniture. Otherwise it is not combined with the definite article
"the". We say at school when the hearer goes there as a pupil.


c. Life is full of surprises.


Comment: Life (U) in this case required a zero article because here
(c) it is a generic reference. All c) is widely generic.


d. I'm researching the life of Mozart.


Comment: On the other hand in case d the speaker is talking about a
specific person's life, Mozart. It is clear in the situation which
person the speaker means.
[…]
```

4.2 The participants

The corpus represents contributions from 240 individuals: 235 students, three teachers, and two in the category ‘other’ (in fact teachers of other courses, grouped with the teachers in the figures below). Figure 2 shows the breakdown into teacher-produced text, student-produced text, and ‘assignment text’. This

last category includes the text of assignments as repeated in students' own work, as in the sentences prefaced with a letter in (3) above. Another example of students repeating a sentence before discussing it is when they are asked to spot the grammatical errors in a given sentence. By marking up such text as assignment text, we avoid attributing it erroneously to students.

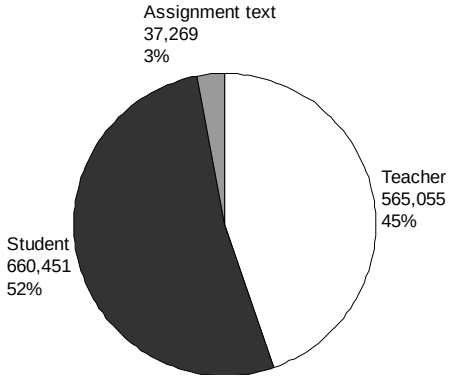

Figure 2: Composition of the corpus by participant role (word counts and percentages)

The most serious disadvantage of the Mini-McCALL corpus – and one which will be rectified in the final McCALL corpus – is the small number of teachers involved. In fact, a single teacher taught four of the eight classes, and in addition was the most prolific communicator. The production of this individual therefore accounts for fully one-quarter of the entire corpus (see the discussion above about mass-mailings). For this reason, the suitability of Mini-McCALL for comparing different teaching styles is limited.

However, in terms of the students, the corpus is much more well-balanced, being fairly representative of the student population. As is commonly the case in foreign-language programmes, a majority of the students – 79 per cent – are female. Interestingly, their contributions (in terms of number of words) account for 80 per cent of the student data, indicating very even participation by the sexes on average. Not surprisingly, given the discussion above, the situation is very different for the teachers: two of the three are male, and these two taught seven of the eight classes. What is more, these two are each more prolific than the female teacher, so their contributions account for 95 per cent of the teacher data. This means that, overall, 55 per cent of the corpus material is produced by

males. Figure 3 shows a breakdown of the corpus material by both participant role and sex:[2]

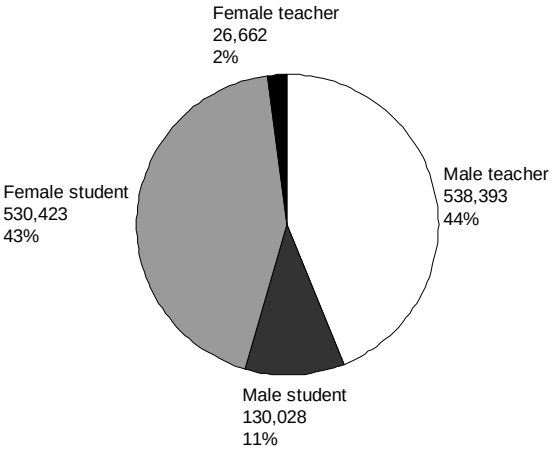

Figure 3: Composition of the corpus by participant role and sex (word counts and percentages)

In terms of student age, the corpus presents a broad range: from 18 to 57. As shown in Figure 4, student age figures are heavily skewed toward the younger end of the scale – the typical age brackets for undergraduates. What is in fact noteworthy about these online courses is the number of individuals aged 25 and older who enroll. These students come from varied backgrounds; some are professionals looking to enhance their career by improving their English skills, while others want to change careers, for example to become teachers. This leads to a greater heterogeneity within the groups than is usual in first-term university courses.

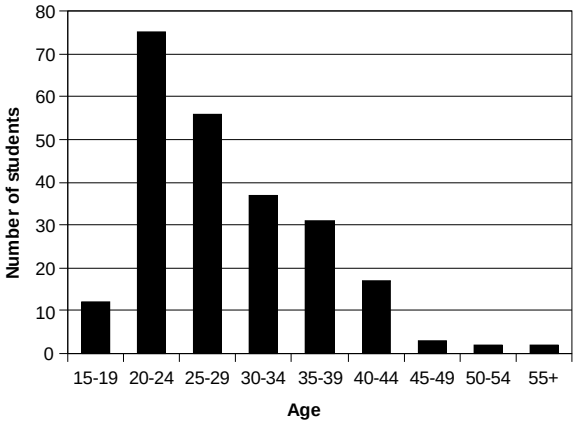

Figure 4: Age distribution of the students in the corpus

As might be expected in the case of a Swedish university, the majority of the students come from a monolingual Swedish background. All of them have studied English for several years in secondary education (a requirement for admission), and several have had further, extensive exposure to English, from living in an English-speaking country. The language backgrounds of the students may be broken down roughly as shown in Figure 5. Note that there are a small number of native speakers of English on the courses, which adds even greater diversity to the corpus. For their part, the teachers are all native speakers of Swedish with extremely high competence in English. The corpus is therefore of special interest to anyone wishing to study the English production of L2 speakers in general, or of Swedes in particular, at different levels of proficiency.

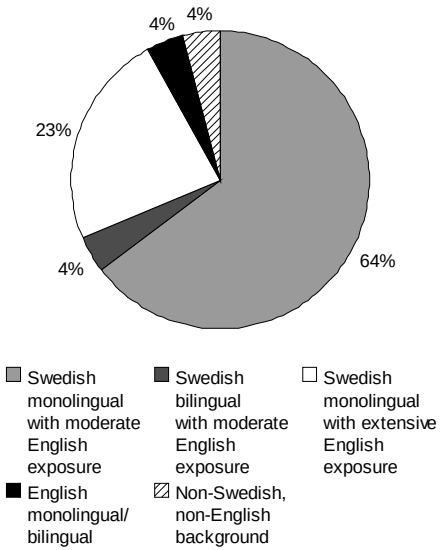

Figure 5: Language background of the 235 students in the corpus

4.3 Construction of the corpus

The primary format of the corpus is XML, though we have also produced an HTML version for convenient browsing of the corpus, and it would be a simple matter to create a plain-text version without mark-up. Example (4), which repeats the content of (2), gives a sample of the raw XML. Currently, the XML application (set of tags) in use is one designed specially for this corpus, though one of the long-term goals of the McCALL compilation project is to adopt a more elaborate tagset in compliance with the guidelines of the Text Encoding Initiative. As for the character encoding, the files are encoded in UTF-8; the use of Unicode means that no sacrifices were necessary regarding non-ASCII characters, which are abundant in Swedish names.

(4)
```
<thread>
  <message id="HT04N.E.3">
    <header>
      <time>Tuesday, September 7, 2004 19:06</time>
      <author id="liljsjög">Lilja Sjögren</author>
      <recipient id="georsede">George Sederstedt</recipient>
      <subject> Assignments! </subject>
    </header>
```

```
      <body>
       Hello George!<lb/>
       I have been in and out of the discussion room a couple of
       times and it seems that noone<lb/> is there to discuss the
       assignment that is due on friday. I need some tips how to
       start the<lb/> discussion. I have never done this before. And
       what do I do if noone is discussing before<lb/> friday?<lb/>
       <lb/>
       Lilja Sjögren<lb/>
      </body>
    </message>
    <message id="HT04N.E.5" parent="HT04N.E.3">
      <header>
        <time>Wednesday, September 8, 2004 11:50</time>
        <author id="georsede">George Sederstedt</author>
        <recipient id="liljsjög">Lilja Sjögren</recipient>
        <subject>Re: Assignments!</subject>
      </header>
      <body>
       Dear Lilja,<lb/>
       For the discussion assignments you actually make your
       comments in the discussion<lb/> forum, which is not a
       simultanous chat. You simply mail your contribution and wait
       for<lb/> someone to respond. The mails can be read by all. In
       the chat you can talk more<lb/> informally. Get back to me if
       this is still unclear.<lb/>
       George<lb/>
      </body>
    </message>
    […]
  </thread>
```

One of the noteworthy aspects of the corpus – though one that is deliberately not salient – is the thorough process of anonymisation that has been applied to the data. All of the participants agreed to have their data recorded, on condition that their real names were withheld. Anonymisation may be performed in various ways, for example by replacing each name with a placeholder such as <NAME>, or by simply removing the surname, as was done in the British National Corpus. Example (5) shows a text fragment from the BNC in two versions: (a) in XML, and (b) in plain text, with the tags removed. Note the <gap> element that has replaced the name, highlighted here in bold text.

(5) a. <w c5="CJC" hw="and" pos="CONJ">And</w><c c5="PUN">, </c><w c5="AV0" hw="then" pos="ADV">then </w><w c5="PNP" hw="she" pos="PRON">she </w><trunc><w c5="UNC" hw="sa" pos="UNC">sa </w></trunc><w c5="NP0" hw="mrs" pos="SUBST">Mrs </w>**<gap desc="name" reason="anonymization"/>**<w c5="VVD" hw="say" pos="VERB">said </w><w c5="PRP" hw="to" pos="PREP">to </w><w c5="PNP" hw="i" pos="PRON">me</w><c c5="PUN"> […]

b. And, then she sa Mrs said to me, not Mrs but Mrs , she helps out, and she organized the Burns supper thing, she said

(BNC, text KE4)

For Mini-McCALL, the removal of the *full* name was required to ensure anonymity, while at the same time it was deemed necessary to enable corpus users to easily distinguish individual participants from one another, unhampered by possibly intrusive coding. Mini-McCALL contains a large amount of dialogue involving numerous individuals, and we anticipate that the corpus will be used in studies of group communication; hence we have opted for a solution that required a more laborious process but has certain advantages. We have anonymised the corpus by replacing every individual's name with a plausible-sounding, randomly generated name that retains the correct sex. Both first and last names were replaced (independently) with names drawn from Swedish census data, with the result that the new names are essentially similar in nature to the old names. Example (6), which shows a message after anonymisation, gives a sense of the non-intrusiveness of this process:

(6) Message HT04S.D.110
Time: Sunday, September 19, 2004 20:29
Author: Tua Skoog
Subject: Tua's input on the dialogue Tense and Aspect
[includes attachment]

Hi Melinda, Yara and Ajdin! (I hope I haven't forgotten anyone!)
Attached you'll find my answers.

I only have two contributions to this grammatical debate. First I'd like to join Yara in sharing her view-point concerning 2 C. Both the Present and the Progressive is possible. However I'd prefer the phrase expressed using the Present Simple tense, in opponent to the Progressive (-ing) form preferred by Ajdin and Melinda.
[…]

The result is discourse that, while protecting the identities of the participants, feels quite natural. Because the names are changed consistently throughout, it should be as easy *after* anonymisation as before to identify the referent of a noun phrase.

One difficulty we faced in this process was that a number of individuals had non-Swedish names. We could have opted to replace their names with equally foreign names, but we realized that this was likely to be insufficient to guarantee anonymity. Therefore, as it stands, an individual named Nigel Jones might receive a new name such as Kjell Lindström. This could potentially mislead researchers, which is certainly a disadvantage. On the other hand, those inter-

ested in the details of the participants' backgrounds will find extensive metadata available in the corpus, as described in the next section.

4.4 Corpus metadata

Mini-McCALL is relatively rich in metadata, consisting primarily of three types: (1) sociolinguistic information on the language users represented, (2) pedagogical information on the course and activities involved, and (3) information on the texts included. Some of this information may be read directly from the texts, while some is available from the metadata that accompany the texts, and some comes from supplementary materials that will be distributed with the corpus. Table 1 presents several types of metadata that are available:

Table 1: Metadata included in the corpus

Type	**Variable**	**Description**
Participant information	Group membership	List of participants in each of the 8 classes
	Sex	Male or female
	Age	In years, at beginning of course
	Language background	See categories in Figure 5 above
	Role	Student, teacher, other
	Student grade	Final grade for the course
	Activity level	Number of messages, number of words
	Teacher experience	Teaching background and experience
Pedagogical information	Course and task descriptions	Description of the course and the various tasks assigned to the students, including instructions
	Course evaluations	Average evaluation scores for each of the 8 classes, relating to instructions, technical aspects, feedback, reading materials, exam, course aims, etc.
Textual information	Type of text	Discussion message, e-mail message or attachment
	Date of submission	Each message is dated, allowing researchers to chart course activity levels
	Sequencing information	Messages are grouped into threads and ordered chronologically, explicitly showing the communicative exchanges occurring

The inclusion of course evaluation data and student grades enables student satisfaction and achievement to be cross-referenced with other corpus data, which should be of value for pedagogical research. Furthermore, the temporal information on the messages enables researchers to follow the dynamics of the course – that is, when and in connection with which tasks students tended to be most or least active.

5 Linguistic and pedagogical research based on Mini-McCALL

5.1 Potential research areas of Mini-McCALL

The corpus is likely to be of value for both linguistic and pedagogical research. With respect to pedagogical research, the corpus makes it possible to study teacher-student communication from a pedagogical perspective in a systematic fashion. As mentioned above, the corpus also includes a range of variables useful to researchers in pedagogy, such as proficiency level, sex, and achievement. Furthermore, the supplementary course materials make it possible to conduct pedagogically-oriented research on the effects of course materials such as task instructions. Ideally, we would have liked to include control group data from a classroom version of the same course, but this has not been possible.

Mini-McCALL allows for a special focus on communicative aspects of computer-supported collaborative learning. It allows systematic study of what we may call the 'language of collaborative learning', which has been a relatively unexplored area. As Dillenbourg *et al.* (1995: 198) point out, a "promising possibility for collaborative learning research is to exploit selective branches of linguistics research on models of conversation, discourse or dialogue". Such research could shed light on the communicative strategies used in the negotiation of meaning, and may lead both to new insights into the social processes involved in collaborative learning and to an increase in the educational standards of CSCL environments.

The design of Mini-McCALL makes possible a range of linguistic investigations. The data not only are rich in (frequently overlooked) interactional features of communication, but also offer the possibility of comparison across a variety of linguistically relevant variables. These are *sociolinguistic* variables – such as participant role (teacher or student), age, and sex – and *discourse* variables – such as type of text (discussion forum messages, e-mails, assignments). A brief selection of potential research areas, both theoretical and applied, might include the following:

- *Adaptation of language to new technologies*: The corpus can be used for several types of investigations in computer-mediated communication, for example on ways in which language users adapt their language to the medium or technology (e.g. e-mailing or discussion fora). Most research in this area (e.g. Herring 2001; Hård af Segerstad 2002) has focused on native-speaker language, so the current corpus data, dominated by non-native speakers of English, should prove to be of great interest.
- *Second language learning*: The corpus also makes a contribution to the available data on student writing in L2 English. Corpus-based learner language research has mushroomed over the course of the past decade (as summarised in Granger 2004). The corpus not only extends the amount of currently available material, but also includes previously unexplored types of text.
- *Interpersonal communication and communicative strategies*: 'Collaborative linguistic behaviour' is of the utmost importance in collaborative learning contexts. The corpus data provide a unique opportunity to study interpersonal aspects of both teacher-student and student-student communication, e.g. speech acts such as apologising, complaining, or asking for favours. The data also enable comparison across teachers and course modules, as well as across private (e-mail) and more public communication (discussion fora). Another highly relevant research topic concerns metadiscursive strategies used by students and teachers, covering research questions such as the following: How do the participants correct errors, ask for clarification, explain terminology, and give positive/negative feedback? What appear to be the most felicitous metadiscursive strategies?
- *Discourse identity construction*: A final example of an interesting research area for which Mini-McCALL would be suitable is the examination of discourse identity construction, for example strategies for construing oneself as a good/bad student, or as a friendly/competent teacher. This is a highly important aspect of social competence in general and of academic socialisation in particular.

5.2 Current research on Mini-McCALL

Several studies based on the pilot corpus are already underway, while one study, an analysis of the effects on student activity of the quantity and quality of teacher communication in the courses, has already been published (Deutschmann and Lundmark 2008). The results of this study suggest that there is a posi-

tive correlation between the quantity of teacher-student communication and the quantity of student-student communication in a course. In other words, active teachers influence the communicative culture in a course and foster student activity. In addition, the study showed that teachers who used less formal, more involved language were more successful in creating active communicative environments in the courses. Finally, the study suggested that courses in which students communicated extensively displayed better throughput than courses in which students were less active.

Popaditch (2009) is concerned with how students deliver criticism in peer-review situations. The study focuses on the use of hedging and other linguistic strategies related to indirectness that seem to be typical of such discourse in the corpus. Some of the interesting features found are illustrated in (7) and (8), which are taken from the discussion fora. In the task discussed, the students are asked to decide whether the progressive or the simple present tense is appropriate in a number of sample sentences. They are also asked to justify their answers; one such justification is shown in (7). In (8), we see a peer-review and the response to the review:

(7) Message HT04N.D.109
Time: Thursday, September 16, 2004 17:21
Author: Ulrika Cederblad
Subject: Re: Tina Kock

Hi Jonna, Tina and everyone!

About 1c, I think it should be the form "going to + base form" because the speaker assumes that there already have been made plans for the weekend before the weekend.

The way I have understod it is that you use this form if you already made the plan in the past before the conversation. Could that be right?

In 1e the speaker assumes that plans for the weekend have not been made yet. But I have been thinking and I have another theory aswell :)and I wounder if you might use this form when the speaker is saying this during the weekend and not before the weekend. Since the person probobly is doing something at the time its an activity and therefor its Present progressive.

Im not sure about this so I would be very glad if somebody could tell if Im right or wrong.

Im a bit confused ...
Ulrika

(8) Message HT04N.D.118
Time: Thursday, September 16, 2004 20:03
Author: Tina Kock
Subject: Re: Discussion 2 - Jonna Österberg

Hello Jonna!

I have an opinion about 1d: "I'll have done it by tonight."
You wrote that will + base form is used here, but "have done" is not the base form of the verb. That's why I think the tense is Future Perfect. The sentence is about an activity that will be completed before a specific time.

Am I right?

/Tina

Message HT04N.D.147
Time: Saturday, September 18, 2004 10:30
Author: Melinda Jensen
Subject: Re: Discussion 2 - Jonna Österberg

OK.... I think you're right! (I wonder why I thought it was will + base form?), now when you have explained it to me everything seems so clear.... :-)

bye
Melinda

These examples, much like (1) above, are in many respects typical of peer-review dialogue, at least the way it is realised in the present material. They represent a genre full of markers of politeness/indirectness, such as the use of modals and hedges, where opinions are suggested rather than stated. There is also a frequent use of interactive markers and signals to establish rapport, such as encouragement, praise, expressions of gratitude, informal language and use of emoticons (underlined in the examples above). As we also saw in (1), put in a learning-theoretical context, the language conforms to socio-cultural models of learning, which maintain that the establishment of a social context and group identity are central to the learning experience.

6 Looking ahead: The real McCALL

Mini-McCALL involves data from only one of the Internet courses (English grammar) taught at Mid-Sweden University, over a limited period of time (four terms). The long-term goal, however, is to create a comprehensive corpus of online language learning – the Mid-Sweden Corpus of Computer-Assisted Language Learning (McCALL) – containing all of the university's online English

courses from 2004-2008. This will be a unique resource, containing several million words of not only written but also transcribed spoken English, from both teachers and students, at various levels, and in various genres. The programme includes courses in language (grammar, essay writing and oral proficiency), cultural studies, literature, and linguistics; the subjects covered are relatively typical of undergraduate courses in English Studies at Swedish universities.

The full material will include over 100 course modules taught by 16 teachers to over 900 students, with all communication taking place in English. The estimated size of the final corpus is over 10 million words of written text and over 100,000 words of transcribed spoken material.

The material to be included in McCALL is quite diverse in character. Four communicative modes will be represented: synchronous writing, asynchronous writing, synchronous speech and asynchronous speech. With respect to the written components, we are particularly pleased to be able to include drafts of many of the written essays, a very rare type of data which will enable interesting research into the writing process. This is particularly relevant to models of peer-reviewing, which are commonly used in teaching academic writing. We not only have drafts of student essays, but also teacher comments on a large portion of the written student production, which means that we will be able to document this 'chain of genres' (Swales 2004) quite thoroughly. Another chain of genres into which this corpus will allow research is the transfer of a written essay to a spoken presentation. Given the design of the corpus, we will be able to observe the development of a first-term student essay through several drafts to the finished essay, and then to a spoken presentation of the same content.

In the planned McCALL material, the primarily text-based learning management system (WebCT) is complemented by other tools such as online video conferencing (Marratech). The online video conferencing sessions that have been recorded supply interesting synchronous speech data. Asynchronous speech data are also available, primarily in the form of audio-commented slide presentations produced both by students and by teachers. Ideally, we would like to distribute not only transcriptions of these but also the sound files, which would also require anonymisation.

7 Conclusion

Here we have described Mini-McCALL, a pilot corpus containing data from three types of written communication – discussion forum messages, e-mail, and written assignments – from an online learning environment in English Studies at Mid-Sweden University. We predict that this will be the first of many corpora of

online learning to come, making it possible to study empirically both the linguistic and the pedagogical aspects of such environments. We have also outlined our proposed next step in the McCALL project, which will lead to a comprehensive four-year snapshot of all of the various types of communication that take place in an online learning environment.

With a resource like Mini-McCALL, we hope to be able to get a clearer picture of the communicative processes involved in an e-learning environment, both at the individual level and at the group level. The corpus will allow researchers to link linguistic findings to social, pedagogical and textual metadata and thus will enable us to pose a number of interesting questions with direct application to e-learning and collaborative learning: What characterises the language of (un-)successful peer-reviewing? How do students negotiate their differences, and to what extent is a good command of the L2 important here? And ultimately, how do the various variables investigated contribute to a student's success or failure in a course?

While the number of Internet-based courses in higher education is steadily increasing, it is also becoming increasingly clear that e-learning environments tend to produce lower throughput rates than traditional classroom environments (Westerberg and Mårald 2006). Therefore, there is an immediate need for a better understanding of those processes which are the key to successful learning in this type of environment. We believe that it is crucial to study the types of communication that take place in such courses in order to achieve this understanding. We offer Mini-McCALL to the research community in hopes of fostering research in these important areas.[3]

Notes

1. We gratefully acknowledge funding for Mini-McCALL from the Department of Humanities at Mid-Sweden University. We also extend our thanks to Nick Sheppard for his assistance in collecting the corpus metadata.
2. Note that the 'assignment text' data are excluded from these calculations, but the 'duplicate text' data are included.
3. Readers interested in obtaining Mini-McCALL should contact Dr. Mats Deutschmann at mats.deutschmann@miun.se.

References

Beißwenger, Michael and Angelika Storrer. 2008. Corpora of computer-mediated communication. In A. Lüdeling and M. Kytö (eds.). *Corpus linguis-*

tics: An international handbook (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29). Vol. 1, 292–309. Berlin and New York: Walter de Gruyter.

BNC. 2007. The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Bonk, Curtis J. and Donald J. Cunningham. 1998. Searching for learner-centered, constructivist, and sociocultural components of collaborative educational learning tools. In C. J. Bonk and K. S. King (eds.). *Electronic collaborators. Learner-centered technologies for literacy, apprenticeship, and discourse*, 25–50. Mahwah, NJ: Lawrence Erlbaum.

Deutschmann, Mats and Carita Lundmark. 2008. 'Let's keep it informal, guys': A study of the effects of teacher communicative strategies on student activity and collaborative learning in internet-based English courses. *Tidskrift för lärarutbildning och forskning* [Journal of Teacher Education and Research] 2: 39–57.

Dillenbourg, Pierre, Michael J. Baker, Agnès Blaye and Claire O'Malley. 1995. The evolution of research on collaborative learning. In E. Spada and P. Reiman (eds.). *Learning in humans and machine: Towards an interdisciplinary learning science*, 189–211. Oxford: Elsevier.

Granger, Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In U. Connor and T. A. Upton (eds.). *Applied corpus linguistics: A multidimensional perspective* (Language and Computers 52), 123–145. Amsterdam: Rodopi.

Hård af Segerstad, Ylva. 2002. *Use and adaptation of written language to the conditions of computer-mediated communication*. Gothenburg: Gothenburg University.

Haythornwaite, Caroline. 2002. Building social networks via computer networks: Creating and sustaining distributed learning communities. In K. A. Renninger and W. Shumar (eds.). *Building virtual communities: Learning and change in cyberspace*, 159–190. Cambridge: Cambridge University Press.

Herring, Susan. 2001. Computer-mediated discourse. In D. Schiffrin, D. Tannen and H. Hamilton (eds.). *The handbook of discourse analysis*, 612–634. Oxford: Blackwell.

Kern, Richard and Mark Warschauer. 2000. Theory and practice of network-based language teaching. In M. Warschauer and R. Kern (eds.). *Network-*

based language teaching: Concepts and practice, 1–19. Port Chester NY: Cambridge University Press.

Koschmann, Timothy. 1996. Paradigm shifts and instructional technology. In T. Koschmann (ed.). *CSCL: Theory and practice of an emerging paradigm*, 1–23. Mahwah, NJ: Lawrence Erlbaum.

Olofsson, Anders D. and Ola Lindberg. 2005. The learning community: An understanding of web-based education? Paper presented at the UniZon Conference *Crossborder NetWorking and Learning*, Vaasa, Finland, 18–19 April 2005.

Palloff, Rena M. and Keith Pratt. 1999. *Building learning communities in cyberspace: Effective strategies for the online classroom*. San Francisco, CA: Jossey-Bass.

Popaditch, Irina. 2009. But, a "tricky little word": The language of cooperation and disagreement in a Swedish online language learning environment. Seminar paper presented on January 27 in the Department of Humanities, Mid-Sweden University.

Salmon, Gilly. 2004. *E-moderating: The key to teaching and learning online*. 2nd ed. London: Routledge.

Schegloff, Emanuel A. 1991. Conversation analysis and socially shared cognition. In L. B. Resnick, J. M. Levine and S. D. Teasley (eds.). *Perspectives on socially shared cognition*, 150–171. Washington: American Psychological Association.

Sotillo, Susana M. 2000. Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning and Technology* 4 (1): 82–119.

Stephenson, John (ed.). 2001. *Teaching and learning online: Pedagogies for new technologies*. London: Routledge.

Swales, John M. 2004. *Research genres: Explorations and applications.* Cambridge: Cambridge University Press.

Tella, Seppo. 1992. *Talking shop via e-mail: A thematic and linguistic analysis of electronic mail communication* (Research Report 99). Helsinki: University of Helsinki, Department of Teacher Education.

Wenger, Etienne. 1998. *Communities of practice: Learning, meaning, and identity.* Cambridge: Cambridge University Press.

Westerberg, Pernilla and Gunilla Mårald. 2006. *Avbrott på nätutbildningar – en studie av när och varför studenter hoppar av alternativt fullföljer IT-stödda distanskurser* [Interrupted net courses: A study of when and why students

drop out or finish IT-supported distance courses]. Umeå: Umeå University, Umeå Centre for Evaluation Research.