Children Online: A survey of child language and CMC corpora

Alistair Baron, Paul Rayson, Phil Greenwood, James Walkerdine and Awais Rashid
Lancaster University

Introduction

This survey is part of a wider investigation into child language and computer-mediated communication (CMC)[1] corpora. Its aim is to assess the availability of relevant corpora which can be used to build representative samples of the language of children online. The Isis project[2], of which the survey is a part, carries out research in the area of online child protection. Corpora of child and CMC language serve two main purposes in our research. First, they enable us to build age- and gender-based standard profiles of such language for comparison purposes. One of the key aims of the Isis project is to construct a toolkit to help law enforcement officers identify adults that are masquerading as children in the online environment. Hence, reference datasets with validated age and gender information for contributors are vital to our work. These reference datasets act as proxies for real police case data. We need to collect data from children (under 12s) and teenagers (ages 13–19) due to the law enforcement scenarios involved. To some extent, the requirement to have verifiable age and gender information for training data for our systems is imposed on us so that we have traceable evidence which can be used in building legal cases. Second, the child and CMC language corpora allow us to retrain our existing Natural Language Processing (NLP) tools to deal with such language. In general, NLP tools are trained on the genres, text types and domains included in existing reference corpora such as the British National Corpus or other large corpora derived from published sources. Such corpora tend to consist of the language of adults from a wide range of traditional genres such as fiction, non-fiction, newspapers, magazines, professional writing and in some cases conversational spoken language. However, these standard corpora do not contain examples of the new online genres (Mehler et al. 2011). Notable exceptions are those collected using the Web as Corpus paradigm such as the Wacky corpus collections (Baroni et al. 2009).

There is a growing awareness in the NLP community of the need to adapt existing NLP tools to new domains (Daumé III et al. 2010), but we are not aware of any previous work in the area of retraining NLP tools to deal with language of children online. Our hypothesis is that this is due to the lack of available corpus data. Although, as we shall see, there is a vast amount of research on child language and a growing quantity of research into online language varieties such as chat rooms, blogging, micro-blogging, social networking sites and instant messaging (IM), we have found that very few such datasets are made available as corpora. In turn, this may be due to the ethical and legal issues associated with collecting and redistributing data from children and online sources. The use of online data is a grey area for language researchers. Arguably, it can be considered acceptable to collect a corpus for a study consisting of data that is already publicly available in online forums (Seale et al. 2006). However, to package up this data and redistribute it without seeking permission from the individuals and/or copyright holders would not be. Initiatives using web-derived corpora usually make available lists of URLs rather than actual copies of data (Sharoff 2006), or they restrict access via a web interface so that full texts cannot be downloaded (Baker 2009; Davies 2009). As for child language data, the requirements of university ethics committees mean that collections are made with ethical consent for direct use by the researchers only and not for secondary analysis or redistribution more widely.

Practical issues may also be a factor in the lack of child language corpora. As with adult spoken data which is more expensive to collect and transcribe than written data, the costs associated with collecting and converting written child language data will be higher, for example, transcription is needed to convert handwriting into text form. This may also be more time consuming if the handwriting is difficult to read.

Collecting a corpus of online forum and chat room data is technically possible, but without validated ages and gender information, the text is of little use in our particular research for building the standard profiles mentioned above. Validated ages and genders are also a prerequisite for other types of analysis, e.g. language acquisition research.

There are a number of reasons motivating the collection and release of corpora of child language and language representing the online world. As with other corpora, it is important to share datasets in order to support "the scientific method", i.e. for reasons of replicability, falsifiability, completeness and objectivity (Leech 1992). Corpora of child language can be used for language acquisition and development research. The analysis of corpora of online and CMC language is growing significantly as this type of communication has become ubiquitous since the development of the World Wide Web.

There are a number of other published surveys of linguistic corpora. Typically, these appear as appendices to corpus linguistic textbooks (McEnery and Wilson 1996; Biber et al. 1998) or online reference lists that are kept up to date (e.g. David Lee's bookmarks for corpus-based linguists[3], the corpus resource database maintained at VARIENG in Finland[4], and the survey carried out by the EU CLARIN project[5]). Surveys of well-known corpora (Edwards 1993; Xiao 2008; Lee 2010) and specific types of corpora (Pravec 2002) have been published. There are also surveys of corpora for specific languages (Wang 2001; Nikkhou and Choukri 2005; Xiao-jun 2006) and groups of languages (Pusch 2002). However, none of these surveys consider the language of children or online varieties as we do here.

In order to train NLP tools and build language frequency profiles for different age and gender groups, including fine-grained age ranges for children and teenagers, we require bodies of text which can be attributed to specific ages and genders. Additional metadata about the source of the text (i.e. the writer or speaker) would also be beneficial, this may include gender and social class markers. Existing corpora were surveyed to assess their suitability for the research, and we concluded that the following corpus properties need to be taken into account:

- The mode of the corpus, e.g. written or spoken. There have been many discussions on the differences between spoken and written language, see for example Perera (1986) and Chafe and Tannen (1987).
- The range of writers/speakers available, especially with regards to age and gender. See, for example, Schler et al. (2006) for the "effects of age and gender on blogging".

- The variety of English in the corpus; e.g. British or American (Hofland and Johansson 1982; Swan 2005:39–44) and dialect (Trudgill and Chambers 1991; Trudgill 1999; Wales 2000; Anderwald and Szmrecsanyi 2009).
- The metadata available for the writers/speakers in the corpus.
- The compilation date of the corpus and how current the data is within it. Jones and Schieffelin (2009) note differences in (IM) language use between 2003 and 2006.
- The size of the corpus.
- How much cleanup of the text is required? How well is the text formatted, is it straightforward to extract the relevant metadata and texts?
- Availability of the corpus. Is it free to use? Are there any restrictions on its use?

The ideal corpus for use in our current research would be a sufficiently large corpus containing a variety of web-based texts with a wide range of British[6] writers/speakers from different age groups and social classes. This hypothetical corpus would also have been collected recently, with well structured text and rich metadata markup and also be freely available with no restrictions on its use. Unfortunately, a perfect corpus is rarely in existence and made available for use, and typically, data from different sources needs to be combined in order to build a corpus or set of corpora for the purpose of the specific research in question. This paper shall assess existing child language and CMC corpora and other sources to establish whether they could be used for our specific purposes. At the same time this review will provide details to other researchers who wish to use existing child language and CMC corpora in their own studies, whilst also highlighting the gaps in data available for study.

The paper is structured as follows. We begin our review of available corpora which contain some child and teenage language in Section 1. In Section 2 we review the availability of corpora derived from online or CMC sources. We focus, in particular, on sources of online child and teenage language in Section 3. The final section of the paper, Section 4, contains a summary table highlighting the lack of corpora directly relevant to the study of child language online. Section 4 also includes our conclusion on this survey and implications for future work in this area. As this paper contains a large number of references, not all will be listed in the reference list, but the paper is accompanied by an online extended 'Bibliography of Child Language and CMC Corpora' that can be accessed at INSERT URL.

1. Child and Teenage Language Corpora

A surprisingly small number of corpora have been produced which specifically contain child and/or teenage language; earlier studies of such language, especially classroom language, have utilised only a small amount of data using qualitative methods and alternative research methods to corpus linguistics. More recent corpus-based studies of child language focus on children's language development and "first language acquisition" (see Foster-Cohen (1999) for an introduction to research in the area) or school-based language (Schleppegrell 2001; 2004). Most of these studies focus on spoken language whereas sources for child and teenage written language are generally harder to find; Smith et al. (1998:217) state "Corpora of child language do exist [...], but consist almost exclusively of speech". The lack of data is probably due to children's written data being largely hand-written, and this is notoriously difficult to transcribe; Smith et al. (1998) describe the difficulties of transcribing the Lancaster Corpus of Children's Project Writing (detailed below). A small amount of research has looked into spelling errors in child language, see Sofkova Hashemi (2003) and Baron and Rayson (2009).

The lack of corpora available is even more apparent when just looking for representations of teenage language, Stenström et al. (2002:x) state "The dearth of investigations into teenage language is due in part to its under-representation in language corpora." There is much more material available for children and teenagers learning English as a second language (see e.g. Granger 1998; Pravec 2002), although this is not suitable to use as a proxy for native language in our case.

The earliest corpus found offering some potential match to our search criteria is the Polytechnic of Wales (PoW) corpus[7], a relatively small corpus of 65,000 words containing transcripts of recordings of 120 children aged between 6-12 both at play and in interviews. The children were divided equally according to gender, age and socio-economic class and metadata is available to reflect this. Possible downsides of the PoW corpus are its age (collected 1978-1984), the specific dialect of the speakers (all children were from Pontypridd, South Wales) and the large amount of processing likely to be required to cleanup the data. Perera (1986) uses this data in her study looking at the features of children's written language and how it contrasts to their spoken language.

Probably the best-known source of child language is CHILDES[8] (Child Language Data Exchange System) (MacWhinney 2000). Development of CHILDES began in 1984 and the resource is still being augmented. A significant amount of transcribed spoken data from young children is available in a variety of languages, including English (both British and American varieties). The majority of the transcripts contain speech from under 5s. Corpora in CHILDES which contain data from children over 5 years old consists of[9]:

British:

- *Fletcher:* The Reading corpus of transcripts from 72 British children aged 3, 5, and 7 (see Fletcher and Garman 1988).
- *Gathercole/Burns:* Collected from Scottish Children, with 4 children in the group "5- year-olds" (with a mean age of 5 years 0 months), the oldest age is 6 years 4 months (see Gathercole 1986).

American:

- *Carterette and Jones:* Collected from 54 first graders (6-7), 48 third graders (8-9), 48 fifth graders (10-11), and 24 adults. The children's speech was recorded at two different schools in California. See Carterette and Jones (1974) for more details.
- *Evans:* Transcripts of first grade (6-7) children in pairs at indoor play. Recorded in
- Ontario, Canada.
- *Gathercole:* Cross-sectional data of children aged 2 years and 9 months to 6 years and 6 months (see Gathercole 1980).
- *MacWhinney:* Transcripts from MacWhinney's diary study of the development of his two sons, Ross (born 1977) and Mark (born 1979). Ross was recorded between the ages of 6 months and 8 years, Mark between 7 months and 5 years 6 months.
- *Frog Story - Wolf/Hemphill* Transcripts from 30 children whose language was studied from ages 6 to 8, the children were recorded narrating a picture book (Miranda et al. 1992).

The texts are in a particular structure, using the CHAT Transcription Format[10] which is used by the CLAN Program[11], in most cases gender and age metadata are available. The database is freely downloadable, although some restrictions are in place on its use[12].

A further spoken corpus is available named Kids' Speech[13]. The corpus was built in 2001 to train and evaluate speech recognisers, and the majority of the corpus is scripted repeated words and sentences. However, for each child there exists a transcript of about one minute of spontaneous speech. Approximately 100 children at 11 grade levels have been recorded, kindergarten (4-6) to tenth grade (15-16), totalling around 1100 minutes of speech. A free set of samples of the corpus is available from the website. Metadata does not seem to be available, although the texts are in individual grade directories and the website details the split between male and female speakers. The full corpus is available from the Linguistic Data Consortium (LDC) catalog[14].

Schleppegrell (1991) describes her use of a set of transcripts from 14 group interviews with 59 third and sixth grade children. It is unclear whether these transcriptions are available to use; further details are available in Mary Schleppegrell's doctoral dissertation (Schleppegrell 1989).

For spoken teenage British English, COLT[15] - the Bergen Corpus of London Teenage Language (Haslerud and Stenström 1995), is probably the most well-known. Collected in 1993, the corpus contains 500,000 words of transcribed speech of speakers aged between 13 and 17 years from various districts of London. Stenström et al. (2002) discuss the compilation of the corpus as well as some analysis and findings. The corpus is well structured and contains age and gender metadata. COLT is part of the BNC (detailed below). More recently, the Linguistic Innovators Corpus (LIC)[16] (Gabrielatos et al. 2010) has been collected, which can be used to compare teenage London English from 1993 (COLT) to 2005. LIC contains 1.4 Million words of London English, mostly from 16-18 year olds (but also 70+). One shortcoming with COLT and LIC is their restriction to a single dialect, London English, although (Stenström et al. 2002:x) state:

> "[...]since London is one of the world's most 'central' and trendiest cities. Its teenage vernacular, we assumed, must infiltrate the language of teenagers far beyond London's boundaries, and even those of Britain itself."

A corpus containing a wider variety of British dialects may be applicable in the form of the IViE Corpus[17] (Grabe and Post 2002; Grabe 2004; Kochanski et al. 2005). The corpus was built to research intonational variation and consists of recordings of speakers at secondary schools (aged about 16 years old) in Belfast, Bradford, Cambridge, Cardiff, Leeds, Liverpool, London, Dublin and Newcastle - 12–14 speakers from each area, recorded between 1997 and 2000. Because of the purpose of the corpus, only 60 seconds of each child's data is 'free conversation'. The corpus is freely available upon filling out a registration form.

SACODEYL[18] is a corpus of transcribed structured video interviews in a variety of European languages, including English (mainly British), with school pupils aged between 13 and 18 years. The corpus is very recent (collated 2006-2009) and contains 21 ten-minute interviews. The corpus was built for use with foreign language learning in schools (see Hoffstaedter and Kohn 2009), and according to the website's data protection note[19] : "The interviewees and/or the interviewees' parents or tutors have formally consented to the use of their interviews for educational purposes. No other use of such material is allowed." The corpus is well structured with TEI XML encoding. Metadata is available for each speaker and consists of age, gender, role and a brief descriptive note. The corpus does contain some spelling variation 'corrections', however these are very few (22 occurrences) and consist of basic forms such as *gonna* → *going to*[20].

The Child Language Survey (CLS) funded by the Nuffield Foundation in the 1960s gathered a large amount of data on the language of children around 8–15 years old. Spoken and written data was collected from various school types in London, Kent, Sussex and Yorkshire. Its size has been estimated at one million words (80% spoken and 20% written). Unfortunately, the CLS data has been largely unexploited because the vast majority of the data has not been transcribed digitally (Perera (1986) is one of a few researchers who have used the data). However, recent efforts have seen portions of the data digitised, including the LUCY Corpus[21] (Sampson 2003; 2005) and the 'Variability in child language' pilot study at Lancaster University[22] (Pooley et al. 2008). The LUCY Corpus is freely available online with no restrictions although the metadata is not present.

The Lancaster Corpus of Children's Project Writing (LCCPW)[23] is another potential source for children's written English. The corpus is made up of project work by a class of 37 children in a UK school over 3 years in the early 1990s. The children were aged 8–11 years old at the time of data collection. The transcription and encoding of the corpus is described by Smith et al. (1998). The majority of the corpus is available to download from the project website. Metadata does not seem to be available for the age and gender of the writer for a particular text, although this may be available directly from the compilers. Spellings have been regularised in the corpus with XML tags; according to Smith et al. (1998:222), these include the original form as an attribute, although in the transcriptions available to download from the website, only the regularised form is available.

Chipere et al. (2001; 2004) describe studies using a corpus of children's writing. Their corpus consists of 918 narrative essays of at least 50 words, which "were collected from various schools in England by the Research and Evaluation Department of The University of Cambridge Local Examinations Syndicate (UCLES)." The essays are from children aged 7, 11 and 14 years and "also cover seven (out of a possible eight) levels of writing ability as defined by the National Curriculum for English" (Chipere et al. 2004:143). The corpus does not seem to be freely available online.

Milton (1998); Milton and Hyland (1999); Hyland and Milton (1997) discuss studies comparing Hong Kong students' writing and British students' writing. As part of this research a corpus of 770 GCE A-Level General Studies exam scripts were collected, totalling 500,000 transcribed words (Hyland and Milton 1997:187). The majority of the students would have been 18 years old when the exam was taken in 1994. The corpus does not seem to be available for download, although the LUCY Corpus does contain the scripts graded A or B, totalling 13,000 words (Sampson 2005).

British A-Level essays are also available from the Louvain Corpus of Native English Essays (LOCNESS)[24]. The corpus contains British and American students' essays: 60,209 words from British A-Level students, 95,695 words from British university students, and 168,400 words from American university students. The corpus is freely available on request from Sylviane Granger or Sylvie De Cock.

The LUCY Corpus contains further texts which may be relevant. Firstly, five pieces of coursework from students on an access course attached to a computing degree at Sussex University, aged 18-22 and male. Secondly, 24 first-year undergraduate essays from five different courses at Sussex University. Unfortunately, metadata does not seem to be present, so the data could only be attributed to 'young adults'.

Berman and Nir-Sagiv (2007) describe research using a corpus containing written texts and spoken transcriptions from American English-speaking children and adults split into four age groups: 9–10 years, 12–13 years, 16–17 years, and 25–35 years. There were 20 participants in each group, 10 male and 10 female. The corpus forms part of a multi-lingual corpus encompassing seven languages: Dutch, English, French, Hebrew, Icelandic, Spanish and Swedish from the Spencer Foundation funded project entitled "Developing literacy in different contexts and in different languages" (see Berman and Verhoeven 2002). The availability of the English portion of the corpus is unclear.

The SCOTS Corpus[25] (Anderson 2007; Anderson et al. 2007) is a 4 million word written and spoken corpus of the language of current-day Scotland. Included in the corpus are transcripts of oral interactions between young children and caregivers. This data is freely available to download from the corpus website with detailed metadata included.

Various more general corpora were scrutinised to establish whether portions could be extracted to represent child or teenage language. It was found that the majority of the corpora only contained texts written or spoken by adults, there were, however, a few exceptions. The British National Corpus (BNC)[26] includes texts ascribed to children and adolescents, and is metadata rich with writers/speakers marked with a gender and age-group (under 15, 15–24, 24–34, 35–44, 45–59, over 59, and unknown). The total corpus is 100 million words, 90% written and 10% spoken and was collected between 1991 and 1994[27]. The International Corpus of English (ICE)[28] is a series of corpora in worldwide varieties of English. Each corpus is 1 million words of written and spoken texts. All texts are dated 1990 or later; ICE-GB, the British variety of the corpus was collected in 1998. Unfortunately, all authors and speakers are aged 18 and above; worse still, on investigation of ICE-GB, only an age-group is given, 18–25 years being the youngest; hence, the texts could only be associated to 'young adults'. Metadata of writers and speakers is located in separate files, but the corpora are well structured and formatted.

Source	**Positive aspects**	**Potential problems**
PoW Corpus	• Speakers aged 6-12 years • Age, gender and socio-economic class metadata	• Small (65,000 words) • Dated (1978-1984) • Specific Welsh dialect • Large amount of cleanup required.
CHILDES	• Younger ages than from other corpora	• Majority of data from children under 5 years

	• Gender and age metadata available	• Only small amount of data is British English • Dated (1974-1992)
Kids' Speech	• Large range of ages (4-16) • Fairly recent (2001) • Straightforward cleanup	• Entirely American English • Lack of metadata (although texts split into grades)
COLT	• Fairly large (500,000 words) • Well structured • Age and gender metadata available • Largely covers teenage years (13-17)	• Specific dialect (London English) • Quite dated (1993) • Overlap with BNC
Linguistic Innovators Corpus (LIC)	• Large (1.4 million words) • Recent (2005)	• Most speakers 16-18 years old • Specific dialect (London English)
IViE Corpus	• Range of British dialects	• Specific age group (~16 years old) • Relatively small amount of data • Unclear exactly what data (and metadata) is available
SACODEYL	• Very recent (2006-2009) • Largely covers teenage years (13-18) • Well structured • Age and gender metadata available	• Unclear whether data would be available
Child Language Survey (Lancaster Project)	• Recently collected • Age and gender metadata available • Written data	• Small (50,000 words)
LUCY Corpus	• Different age groups (8-15, 18+) • Written data	• Some texts very dated (1960s) • Metadata does not appear to be present • Cleanup would be required to use the data
LCCPW	• Data from the same children over 3 years • Age and gender metadata • Well structured • Written data	• Fairly dated (Early 1990s)
LOCNESS	• Fairly large (324,000 words) • Written data	• Only has data from late-teens • Half of the data is American English • Unclear on structure of texts and metadata
SCOTS	• Detailed metadata available • Well structured • Child data is mainly recent (2000s), although the whole corpus covers from 1949-present	• Precise age not available, only decade of birth • Specific dialect (Scottish) • Small dataset, relevant data covers 59 texts
BNC	• Large (10 million spoken words) • Age, gender and socio-economic	• Fairly dated (early 1990s) • Precise age not always available

	class metadata available • Well structured • General corpus of British English (should contain good spread of language varieties)	• Despite size, only small portion is from children/teenagers • Overlap with COLT
ICE	• Different global varieties of English • Large (1 million words for each variety) • Gender and age (group) metadata available • Well structured	• Youngest age is 18 • No precise age, youngest age group is 18-25

Table 1 - Summary of corpora available containing child/teenage language.

2. Computer Mediated Communication (CMC)

Computer Mediated Communication (CMC), defined by December (1996), is the sending and receiving of messages via the internet, that is rather than face-to-face communication or traditional written communication. Examples include email, forums and bulletin boards, newsgroups (Usenet), chat rooms (IRC), instant messaging (IM), blogs and social networking sites. More recently SMS has come to be included as a form of CMC; Herring (2007:1) defines CMC as "predominantly text-based human-human interaction mediated by networked computers or mobile telephony". Herring (2003; 2007) also describes a specific area of CMC, Computer-Mediated Discourse (CMD) which is "distinguished by its focus on language and language use in computer networked environments, and by its use of methods of discourse analysis to address that focus" (Herring 2003:1).

Various authors (e.g. Crystal 2001; Beißwenger and Storrer 2009) have compared CMC to the spoken and written modes of communication, others have compared specific forms of CMC to the spoken and written modes, for example, email (Yates 1996; Baron 1998; 2003), Chat (Schulze 1999; Greenfield and Subrahmanyam 2003; Zitzen and Stein 2004; Al-Sa'di and Hamdan 2005) and IM: (Ferrara et al. 1991; Voida et al. 2002; Tagliamonte and Denis 2008; Jones and Schieffelin 2009). The general consensus seems to be that CMC has features from both spoken and written forms, but that the language is distinct enough from both to be considered as a separate mode of communication. Furthermore, the various forms of CMC have their "own usage conditions and therefore, each needs to be analysed in its own right" (Baron 2004); for instance, CMC can be in two forms, synchronous (chat rooms and instant messaging) or asynchronous (email and SMS), Honeycutt (2001), Sotillo (2000), af Segerstad (2002:57) and Rettie (2003) examine the difference between these forms in more detail, Grinter et al. (2006) and Ling and Baron (2007) compare SMS and IM use (for teenagers).

Beißwenger and Storrer (2009) discuss various issues in collecting CMC data, such as formatting and acquiring metadata, whilst Androutsopoulos and Beißwenger (2008) give an overview of data collection and methodology issues in Computer-Mediated Discourse Analysis (CMDA). One important issue is the question of research ethics, especially with regards to privacy concerns. In other words, is it ethical to obtain freely available online data and use it for research without first obtaining permission from the website maintainer and/or individual author? See Cherny (1999: Chapter 7), Jacobson (1999), Eysenbach and Till (2001), Hudson and Bruckman (2004) and Seale et al. (2006) for detailed discussions. Partly due to these issues, freely available corpora of CMC data for general use are few and far between, a problem recently highlighted by Beißwenger and Storrer (2009:4):

> "Research on Computer-Mediated Communication is presently conducted for the most part with project-related corpora of raw data. This is due to the fact that CMC research is a relatively new field and that CMC genres currently are not at all or only marginally represented in large balanced corpora. Thus, at the present time, the

> assortment of large accessible corpora that were exclusively designed for analysing CMC phenomena is rather unsatisfactory. Therefore, for empirical studies, corpora often have to be acquired from the Internet or obtained from users of CMC facilities."

One prevalent feature of CMC language is its distance from standard written English, particularly in terms of Orthography. Crystal (2001) discusses the "language of the internet", highlighting its features and peculiarities, these include features such as: abbreviations ("prob"), acronyms ("lol") and other shortcuts ("2mora"), case irregularities ("i"), emotion indicators (":-)", "haha", "sooooo") and other spelling issues, e.g. typing errors ("hpuse") and misspellings ("wierd"). There are various "dictionaries" of these forms, on websites[29] and in books (normally as appendices) Crystal (2001; 2004; 2008), see also Thurlow (2003) who lists non-standard forms found in a small corpus of teenage text messages. These peculiarities make corpus linguistic analysis difficult; Ooi (2001) examines the issues in the part-of-speech annotation of chat room data. Tavosanis (2007) looks at classifying different types of spelling 'errors', particularly in blogs, Varnhagen et al. (2009) performs a detailed categorisation of spellings in Instant Messaging, Myslin and Gries (2010) carry out an exploratory and descriptive study of Spanish internet orthography, and Driscoll (2002) in her study of "Gamer chat" notes a series of such features, including shortenings, acronyms, alternative spellings and new meanings for standard words. There has also been some research into normalising spelling in CMC data (Clark 2003; Ringlstetter et al. 2006), and specifically for SMS (Aw et al. 2006; Choudhury et al. 2007; Kobus et al. 2008; Acharyya et al. 2009; Cook and Stevenson 2009; Yvon 2010; Beaufort et al. 2010), chat (Wong et al. 2006; 2008), emails (Sproat et al. 2001; Agarwal et al. 2007) and newsgroups (Agarwal et al. 2007; Zhu et al. 2007). Furthermore, recent NLP and Information Retrieval workshops have focussed on research in the area of CMC language (Karlgren 2006; Lopresti et al. 2008).

Many studies have analysed the effect of gender in CMC use and language, see for example Danet (1998), Sussman and Tyson (2000) and Yates (2001). Studies have focused on how females and males use CMC differently (e.g. avatar usage (Kang and Yang 2006; Nowak and Rauh 2006) - and also how assigned avatars effect gender-based language (Palomares and Lee 2009)), but also language use has been investigated, such as the use of emoticons (Witmer and Katzman 2006) and exclamation marks (Waseleski 2006). Some studies have also looked at age differences, such as Grinter and Palen (2002) Kang and Yang (2006), Schler et al. (2006) and Tagliamonte and Denis (2008). Thelwall (2008b) considered differences in swearing across age, gender and country (UK-US) variables. Further studies which analyse gender and age differences in specific CMC forms are included in the extended bibliography.

These age and gender differences have been taken further in recent studies, transferring author profiling and authorship attribution research on non-CMC language (for example, Singh (2001) discusses detecting gender through "lexical richness methods") into the domain of CMC. Authorship attribution aims to identify the author of a text by comparing it to other texts by the same author, studies have looked at emails (de Vel et al. 2001; Tsuboi and Matsumoto 2002; Corney 2003) and newsgroups (Argamon et al. 2003). Author profiling aims to discern certain aspects of an author's identity, such as gender or age, without necessarily having any other texts by that author, studies have looked at emails (Thomson and Murachver 2001; Corney et al. 2002; de Vel et al. 2002), blogs (Schler et al. 2006) and chat (Lin 2007).

The remainder of this section will consider the individual forms of CMC in turn, listing studies of each form and possible corpus sources. SMS (2.1), instant messaging (2.2), chat rooms (2.3), emails (2.4) are covered individually with other forms such as newsgroups, forums/bulletin boards, blogs and social network sites covered in Section 2.5. A summary table of CMC corpora is given in Section 2.6.

2.1. SMS

There has been a lot of press and research activity recently in the area of text messaging, for general SMS discussions see Crystal (2008) and Thurlow and Poff (Forthcoming). SMS language, or "textspeak" has received particular attention (see e.g. Shortis 2007), for an SMS glossary (emoticons, abbreviations, etc) see Crystal (2004). As well as the corpora studies detailed below, various SMS studies use methods such as surveys, ethnography and focus groups without using SMS corpora (these are included in the extended bibliography). There is also an informal bibliography of SMS related papers available[30]. The remainder of this section will detail potential SMS corpora discovered.

CorTxt (Tagg 2009) is a corpus of 11,067 text messages (SMS), totalling 190,516 words. It was collected between 2004 and 2007 for Caroline Tagg's PhD research at Birmingham University. The vast majority of messages were collected from Caroline's friends and family, a small number of messages (441) were collected from the AOL anonymous online public forum (Tagg 2009:67). Participants in the study were aged 19–68 and were mainly British English speakers.

Choudhury et al. (2007) use a corpus of English SMS texts downloaded from Treasuremytext[31] to evaluate an SMS normalisation technique. The SMS texts were manually translated into a Standard English form and automatic alignment was used to produce pairs of SMS spellings and their standard equivalents. The corpus is available online[32] and contains 854 individual messages. There are some problems with the corpus: initial analysis of the standardisation reveals some clear mistakes, automatic alignment between the SMS texts and Standard English texts is only 80% accurate (Choudhury et al. 2007:162), and no metadata is provided related to who wrote the text messages. As for the source of the corpus, Treasuremytext, there does not appear to be a simple method available for extracting large amounts of publicly available texts, and you are required to know the username of a public profile before reading their texts. It is possible that the site has changed in its structure since Choudhury's corpus was built in 2007. Cook and Stevenson (2009) also use this corpus, in addition to their own (which does not appear to be publicly available), to evaluate their own SMS normalisation technique.

With a research focus of optimising predictive text entry for SMS, How and Kan (2005) produced the NUS SMS corpus - a collection of 10,117 SMS messages mainly from Singaporean university students. The corpus is freely available from the Internet[33] in a structured XML format, no standardisation of the messages is provided. Unfortunately, little metadata is given - age and gender can not be attributed to individual messages. Another issue is the language variation introduced through nearly all 'texters' being non-native English speakers. Acharyya et al. (2009) also use this corpus, in addition to their own (which does not appear to be publicly available), to evaluate their own SMS normalisation technique. More recently, the project has been resurrected as a "live corpus project"[34]. As of February 2012, 41,208 English SMS messages have been collected with contributions from various sources with detailed metadata for each message included. Further examination is required to assess whether this growing corpus would be useful for our purposes in terms of how many of the messages are native (British) English and how many of these messages contain adequate metadata.

The HKU SMS Corpus[35] contains 853 messages, totalling 6787 words. It was created as part of the project 'Linguistic Features of Mobile Phone Communication' at The University of Hong Kong. The corpus contains some spelling standardisation, but contains no metadata. The corpus is a mix of Chinese and non-native English; like the NUS corpus, the language variation introduced by this may be problematic for comparisons to British SMS data.

A further small SMS corpus exists on the website[36] to accompany Tim Shortis' (2001) textbook *The Language of ICT: Information and Communication Technology*. Few details are available about the corpus other than there being 202 messages, which were collected in the UK around the year 2000. Bieswanger (2007) used the corpus in his study comparing English and German SMS language.

Psychology researchers (Tim Grant & Kim Drake) at University of Leicester (2006) aim to collect text messages with metadata from volunteers through a web form[37]. No results have been released as yet, and it is unclear whether their data will be made available.

The sms4science project[38] aims to create an international corpus of text messages. It is coordinated by the Centre for Natural Language Processing (CENTAL) at the Université Catholique de Louvain in Belgium. At the time of writing, no English component to the corpus was available.

There are various further publications which discuss using a corpus of SMS texts in their research; however, often the corpus they have used is not freely available or details of its source are lacking. For example, Aw et al. (2006) discuss their use of an SMS corpus to test a normalisation technique with the final aim of English-to-Chinese SMS translation. Their corpus “consists of 55,000 messages collected from two sources, a SMS chat room and correspondences between university students” (Aw et al. 2006:35). It is unclear if their corpus is available for external use, whether metadata is included, or what background the ‘texters’ in their corpus have. Other publications which discuss a potential source of SMS texts but do not indicate whether the data is externally available can be found in the extended bibliography.

2.2. Instant Messaging

Instant Messaging is an area which has seen a lot of research - with and without corpora, however, searches revealed not a single available corpus of messages at the time of writing. There are many studies which do not release their corpora or do not indicate whether their corpora are available, these are included in the extended bibliography. Privacy issues may be impinging on these corpora being freely available for research.

2.3. Chat Rooms

Chat rooms (e.g. Internet Relay Chat (IRC) (Werry 1996)) vary in different ways; for instance, whether they are free or require a subscription fee, the age range they target or whether they are orientated to a certain theme. These features are likely to affect the range of participants in the chat rooms and the language used. One particular area which has received some research attention is the difference between monitored and unmonitored chat rooms (see Tynes et al. 2004; Subrahmanyam et al. 2006).

Chat rooms are different to other forms of CMC in that conversations are generally between multiple participants, rather than generally being one-to-one. This has led to a particular thread of research which has focused on discourse analysis and thread structure of chat logs, problems exist because it is not always clear to whom a user’s message is directed. Holmer (2008) describes *ChatLine*, a piece of software which can be used to view and analyse chat logs’ discourse structure. Further research in the area of discourse analysis and chat logs comes from Forsyth (2007) and Forsyth and Martell (2007).

Users have a very large number of software platforms and rooms to choose from when deciding where to chat, with every topic imaginable being discussed somewhere. Detecting the topic of chat rooms is an area which has received research attention, both for security purposes and to help users find relevant chat rooms. Bengel et al. (2004) introduce *ChatTrack*, a tool for use by the “intelligence community” to build profiles for certain chat rooms or users which are compared to concept profiles in order to classify the topics discussed in the chat rooms or by a specific user. Meanwhile, Van Dyke et al. (1999) present *Butterfly*, an IRC based tool for suggesting channels (chat rooms) for given topics based on keywords in the chat logs. Further studies on topic detection in chat rooms include Tuulos and Tirri (2004), Çamtepe et al. (2004) and Adams and Martell (2008).

Despite a large amount of research into chat rooms, the NPS Chat Corpus[39] (Forsyth and Martell 2007) is the only corpus we discovered that is freely available for academic use. The corpus currently contains 10,567 posts (out of 500,000 posts in total). The messages are from age-specific chat rooms: “teens”, “20s”, “30s”, “40s”, “adults”; although there is no verification of the age and gender of actual posters. Lin (2007) uses the corpus to assess features such as type/token ratio, emoticons used and punctuation to profile authors in chat logs - the results were inconclusive, with it being indicated that this may be due to the small text sizes available. Forsyth (2007) also uses the corpus, introducing XML encoding, POS-tagging and discourse analysis.

Chat logs, IRC logs in particular, can be found quite easily on the Internet using a simple search. A corpus could therefore be built from these logs, however no metadata is available for who is actually chatting and some formatting work would be required to clean up the data and collate logs from different sources.

As with other forms of CMC, many chat corpus based studies were found where the source of the data was not revealed or the corpus used was not externally available, see the extended bibliography. One frequently used method to obtain chat data is to use a “chatbot” which acts as a user on a chat room and silently records all of the chat messages; two such tools have been developed and used by Çamtepe et al. (2004) and Bengel et al. (2004). As mentioned already, there are clearly ethical considerations to take into account here (see for example Hudson and Bruckman 2004).

2.4. Email

The Enron email dataset is probably the best-known corpus of emails available. It contains emails sent by Enron employees (mostly senior management) between 1999 and 2001, the emails were publicly released during the legal investigation of the Enron corporation. The corpus is available online[40] (described by Klimt and Yang 2004) and contains around 500,000 messages from about 150 employees, although no metadata is present. Kalman et al. (2006) use this corpus amongst others in a study of response latencies.

The detection of spam email is quite a well-researched and developed area, with various algorithms developed. The SpamAssassin Public Corpus[41] was built to test spam filtering. The corpus contains around 6,000 emails, 31% of which are spam. The spam and non-spam are distinguished. The data is fairly recent (2002-2006), but the corpus lacks metadata for senders.

Deutschmann et al. (2009) introduce Mini-McCALL, a 1.3 million word corpus of “computer-mediated communication in the context of online English university courses.” The corpus contains forum discussions and emails. The email portion of the corpus consists of nearly 6,000 messages totalling 587,524 words. The corpus is distinguished from most other CMC corpora in that rich metadata for participants is present, including age and gender. One important point to note with the corpus is that its participants are non-native speakers (from Sweden).

There are numerous other studies of large and small email corpora where the corpus is not available for outside use, or it is not made clear whether the corpus is available, these are included in the extended bibliography.

2.5. Other CMC forms

Usenet, which pre-dates the world wide web, is a discussion platform split into categories (newsgroups) where users can read and post public messages. Whilst Usenet has largely been replaced by other forms of CMC such as forums, it is still used today. A very large corpus (over 30 billion words) of English Usenet postings has been created which is freely available to download for academic research (Shaoul and Westbury 2011). Texts less than 500 words or more than 500,000 words were omitted as well as any texts which contained less than 90% English. Unfortunately for our profiling requirements no metadata is available for the authors of individual messages.

A smaller corpus from 20 newsgroups has been created[42] and is generally used for text classification (see e.g. Agarwal et al. 2007). The dataset contains approximately 20,000 documents spread almost evenly across the 20 different newsgroups. The website indicates that it was created by Ken Lang in 1995.

There are further studies with newsgroup data, see Sallis and Kassabova (2000), Argamon et al. (2003)[43], Witmer and Katzman (2006), Zhu et al. (2007) and Hoffmann (2007), whilst Donath (1999) performs an ethnographic study of various features of Usenet communication.

Blogs (or weblogs) are public ‘broadcasted’ online commentaries. Entries are posted by individuals on dedicated sites such as wordpress[44] or blogger[45], with an infinite array of subjects covered. As the blogs are publicly available, it would be quite straightforward to build up a corpus of these blogs; previous studies have done just that (see Huffaker and Calvert 2005; Schler et al. 2006; Stuart 2006; Thelwall and Stuart 2007). Profile information is even available for the ‘bloggers’ such as age and gender, however, the ease in which this data can be falsified leaves its validity at best questionable.

Microblogging is a fairly new phenomenon, on sites such as Twitter[46] users submit short posts ("tweets") updating their "followers", and the world, with what they are doing or what is on their mind. Again, as these posts are publicly broadcasted, it is simple to create a dataset, see Java et al. (2007), who have done so (without, it seems, publicly releasing the data). Petrović et al. (2010) report on the Edinburgh Twitter Corpus which contains 97 million tweets, however, due to a request from Twitter, this dataset is no longer publicly available.

Forums or Bulletin Boards are places where users can discuss topics in an asynchronous fashion, the boards are usually part of a larger website dedicated to a certain subject. The Mini-McCALL corpus mentioned previously for its email content also contains text from forums. In total there are 462,890 words of forum data. As with the emails, the data is rich in metadata for the participants but comes from non-native English speakers (from Sweden).

Yahoo! Answers[47] allows any Internet user to post a question which other Internet users can post answers to. These questions and answers are similar to forums, although in a specific form. Yahoo! have constructed a large corpus of user posted questions and answers by producing a "dump" of the database in October 2007, which contains 4,483,032 questions and their answers. This resource is available for academic use upon request[48]. The corpus is intended for use in training and testing answer-extraction models, see for instance Surdeanu et al. (2008), who use a subset of the corpus. Rafaeli et al. (2005) introduce research using a similar dataset from Google Answers, this however does not seem to be publicly available.

Further studies of forum data are present where researchers have used their own corpora: Collot and Belmore (1996), Ravid and Rafaeli (2004), Waseleski (2006), Yin et al. (2009). Furthermore, Thomson (2006) looked at forum discussion threads on different topics, assessing the effect of gender-stereotypical topics on known gender-based language traits, the study (which was combined with a similar study of chat room language) found that topics produced more gender-based language traits than the actual gender of participants. Kendall (2000) performs an ethnographic study of the online forum BlueSky, looking at masculinity amongst other things.

It is possible to parse text from a range of forums using the same technique as many are structured using software such as vBulletin[49] and phpBB[50]. A web-crawler could be used for this function with boards selected for different subjects, profiles of users are also generally available providing age and gender metadata. One hurdle with this is that vBulletin and phpBB normally require registration to view posts and profile information, this causes ethical issues as requiring a login indicates a private space (Mayer and Till 1996). More discussion on this issue is provided by Eysenbach and Till (2001).

Social Networking (Boyd and Ellison 2008) on sites such as Facebook, MySpace and Bebo has become one of the more popular forms of CMC in recent times. The sites allow users to create a profile which are similar to "home pages", users can share photos, become "friends" with other members and write messages and post status updates (similar to microblogging). No corpora of social networking dialogue or profiles are readily available, however various authors have built their own corpora. Thelwall (2008b) built a corpus of MySpace profiles to analyse the use of swear words by males and females from the UK and US at different ages. He also quantitively analysed various aspects of MySpace usage such as number of friends, date since last visit and various information and answers to questions given by members (Thelwall 2008a). A web crawler, SocSciBot[51] was used to build the corpus. One limitation of studies of gender and especially age differences with retrieved social networking profiles is the potential lack of accuracy in the metadata given; Thelwall (2008a) acknowledges this: "Almost all of the data analysed is self-reported and presumably some of it deliberately or accidentally incorrect. For example, members may lie out their age and probably there are many users under 14 who declare an older age in order to have a profile." Another study of under-18 MySpace profiles by Hinduja and Patchin (2008) found that there was evidence of 8.3% of ages given being inflated. One further study by Shaw (2008) built a corpus of Bebo profiles, looking at non-standard spellings and the difference in spelling trends in different countries. 90 profiles were collected in total, 30 from each of the US, Ireland and England.

2.6. Summary Table

Source	Positive aspects	Potential problems
SMS		
CorTxt	• Recently collected (2004-2007) • Relatively large (nearly 200,000 words) • Large age range (19-68)	• Participants were mainly friends and family of creator, so not necessarily a representative sample
Treasuremytext	• Recently collected (2007) • Contains spelling standardization	• Small (854 messages) • No metadata • Unclear which (if any) texts are British English • Problems with standardisation alignments
NUS SMS Corpus	• Fairly large (10,000+ messages) • Recently collected (2005). • Well structured. • Currently being updated with new material (40,000+ messages) which also contains detailed metadata.	• No metadata (in original collection) • Largely Singaporean non-native speakers of English
HKU SMS Corpus	• Contains some spelling standardisation	• Very small (6,787 words) • No metadata • Mix of Chinese and non-native English
Shortis	• Straightforward to extract data	• Very small (202 messages) • Very little data available about participants • Fairly dated for CMC corpus (2001) • No metadata
IM – No sources available		
Chat		
NPS Corpus	• Fairly recent (2006) • Straightforward to extract text • Split into age-targeted chat rooms	• (Probably) American English • No specific user metadata available • No guarantee that users in chat rooms are of any age group
Email		
Enron Corpus	• Large (500,000+ messages)	• Restricted domain (Business orientated, one company) • Slightly dated (1999-2001) • Little sender metadata
Mini-McCALL	• Rich in metadata • XML based, parsing should be straightforward • Decent sized (500,000+ words) • Recent (2004-2006)	• Non-native English (Swedish students) • Restricted domain (distance learning)
SpamAssassin	• Decent sized (6,000 messages) • Fairly recent (2002-2006)	• Little sender metadata • 31% spam messages
Other CMC		
Usenet Corpus	• Extremely large (25 billion	• No metadata available for

	words) • Recent (2005-2010) • Cleaned up to an extent, with only documents containing 90% English words selected	individuals • Unclear whether Usenet data is significantly different from other CMC varieties
20 Newsgroups	• Quite large (20,000 documents) • Partitioned evenly amongst topics	• Probably fairly dated (1995) compared to Usenet Corpus • No metadata for individuals • Unclear whether Usenet data is significantly different from other CMC varieties
Mini-McCALL (forum data)	• Rich in metadata • XML based, parsing should be straightforward • Decent sized (462,890 words) • Fairly recent (2004-2006)	• Non-native English (Swedish students) • Restricted domain (distance learning)
Yahoo! Answers	• Large (4,483,032 questions and their answers) • Recent (2007) • Category metadata available	• No metadata for individuals • Request required

Table 2 - Summary of relevant CMC corpora.

3. Child/Teenage Language in CMC

As we have highlighted in the introduction, we need to consider child and teenage language on the Internet for the Isis project, therefore CMC corpora with significant contributions from non-adults would be extremely useful. Unfortunately, despite many children and especially teenagers being heavy users of the Internet and CMC, such corpora are few and far between, certainly with reliable metadata confirming the age of participants. Of the CMC corpora in the previous section, few have an indication of the age of participants, fewer still contain child or teenage language - in fact, none of the corpora that we have considered contain texts from under-13s. A few studies previously listed do contain teenage, and a small amount of child language in CMC, but it is unclear whether these are available for external use.

Of the SMS corpora listed, CorTxt is the only source which contains age metadata, however all of the participants in the corpus are 19 and over. Three sets of text messages used in previously listed SMS research studies do contain child/teen language. Plester et al. (2008; 2009) and Plester and Wood (2009) asked thirty-five 10-11 year olds to "translate" a passage of Standard English into "text language", whilst these messages are not naturalistic (as the authors acknowledge (Plester and Wood 2009:158)), they are a rare insight into children's SMS usage. Eldridge and Grinter (2001) and Grinter and Eldridge (2001; 2003) compiled a small corpus of text messages from five boys and five girls aged 15-16, asking the participants to log their text messages over seven days. Finally, Thurlow (2003) had 135 students (older teenagers - mean age was 19) transcribe five recent text messages from their phones, 544 messages in total were recorded.

Obviously, the lack of IM corpora found in the previous section means there are no publicly available corpora of child/teen IM use, however, several papers use IM Corpora with teenage participants. Craig (2003) gathered 11,341 lines of text from IM conversations sent from US youths aged 12-17. Tagliamonte and Denis (2008) collected a corpus of IM conversations from 71 Canadians ages 15-20. Jones and Schieffelin (2009) built an IM corpus containing 132 conversations, 66 from 2003 and 66 from 2006, participants were aged 18-22. Varnhagen et al. (2009) collected IM conversations from 40 youths ages 12-17. There are also several papers analysing IM use by teenagers not using corpus analysis techniques (Grinter and Palen 2002; Schiano et al. 2002; Lewis and Fabos 2005; Bryant et al. 2006; Grinter et al. 2006; Boneva et al. 2006).

The only chat corpus found, the NPS corpus, has chat messages from different age-targeted chat rooms, one of which was "teens". Whilst the text in this portion of the corpus should be from teenagers, it cannot be guaranteed that all participants are under any given age. Tynes et al. (2004) and Subrahmanyam et al. (2006) use a corpus of 38 chat sessions from teen orientated chat rooms (monitored and unmonitored), some of the participants reveal their age and gender in the transcripts, but it is debatable whether this could be used as reliable metadata. Further studies of teen chat use not using corpus analysis include Clark (1998), Merchant (2001), Greenfield and Subrahmanyam (2003) and Subrahmanyam et al. (2004).

Unlike other forms of CMC, the general focus of research into email use and language has focused away from adolescent use, possibly due to the widespread use of email in the workplace. One corpus, however, focuses on teenage language use entirely; Harvey et al. (2007; 2008) introduce a growing corpus from the Teenage Health Freak website[52]. Harvey et al. (2008) state that the corpus comprises 62,794 messages collected between January 2004 and December 2005, totalling 1 million words (according to an ESRC awarded research grant to exploit the data (Adolphs and Mullany 2009) the corpus now contains 4 million words). As (Harvey et al. 2008:2) acknowledge, the messages in the corpus are not strictly emails but "one-off postings, which may or not receive replies". The corpus is not publicly available. One issue with the corpus is that it has a restricted domain due to the nature of messages from teenagers all being about their health concerns.

The Mini-McCall corpus provides a small amount of data from teenagers. The corpus contains 12 participants between 15 and 20 (mostly towards 20, rather than 15), as previously mentioned, however, the participants in the corpus are non-native English speakers. The Mini-McCall corpus also provides forum data, although only from a small number of participants as with the email data.

For blogs, Huffaker and Calvert (2005) created a corpus of "teenage blogs" by searching various blog sites for search terms such as "teens", "teen blogs" and "teenager". After reviewing and removing some blogs (such as those from over-18s) 70 blogs remained, of these 67% revealed their age (range: 13-17), however, the authors state "It must be noted that there was no way to validate the physical identities of blog authors. Because actual age or gender could be falsified in the virtual environment, this study could only explore the online personae that were displayed in the blogs." Schler et al. (2006) built a large corpus of blogs, 11,069 of which stated their age as between 13 and 17. Again, it is debatable whether the ages given can be relied upon as being genuine.

Source	Positive aspects	Potential problems
SMS – No sources available		
IM – No sources available		
Chat		
NPS Corpus	• Fairly recent (2006) • Straightforward to extract text • Separate teen targeted chat room	• (Probably) American English • No specific user metadata available • Cannot know for sure whether users of teen chat room are all teens. • Teen chat room portion is small (2,118 messages)
Email		
Teenage Health Freak website	• Large (4 millions words) • UK website (should be mainly British English) • Age and gender may be	• Cannot 100% guarantee age and gender of users • Unclear what format data is in, how much cleanup will be

	available on request	necessary - Restricted domain of health issues
Mini-McCALL	- Rich in metadata - XML based, parsing should be straightforward - Fairly recent (2004-2006)	- Non-native English (Swedish students) - Only contains 15-19 year olds (majority of which are likely to be upper-end of that group) - Only small portion (about 12 students) are under-20 - Restricted domain (distance learning)
Other CMC		
Mini-McCALL (forum data)	- Rich in metadata - XML based, parsing should be straightforward - Fairly recent (2004-2006)	- Non-native English (Swedish students) - Only contains 15-19 year olds (majority of which are likely to be upper-end of that group) - Only small portion (about 12 students) are under-20 - Restricted domain (distance learning)

Table 3 - Summary of CMC corpora containing child/teenage language.

4. Summary and Conclusions

Table 4 shows all of the available corpora discussed in this paper. Child and Teenage language is fairly well represented for spoken language, and to a lesser extent traditional written language. CMC corpora exist for most forms, except Instant Messaging, for adult (or unknown) ages. It is the CMC language used by children and teenagers, and marked-up as such, which is massively under-represented in available corpora. Studies of CMC language by children and particularly teenagers are abundant, however, many of these either do not utilise a corpus of data or have not publicly released the corpus they have created. Even the few corpora found containing teen CMC language have issues: the NPS corpus age range is based on the chat room being named "Teen" - there is no way of identifying the actual age of users, the Mini-McCALL corpus contains non-native English speakers, and the Teenage Health Freak website corpus contains messages from a very restricted domain of health concerns. It may be possible to fill this gap by requesting the use of non-public corpora from research studies discussed in this paper and included in the extended bibliography, alternatively new corpora could be built using similar methods employed elsewhere (such as chatbots and web crawlers). However, obtaining a substantial corpus of language from children and teenagers using different CMC platforms with reliable metadata (even just age and gender) will, in all likelihood, require extra endeavour to collect and compile a new corpus.

	Adult (20+) / Unknown	**Child (5-12)**	**Teen (13-19)**
Spoken		- PoW Corpus - CHILDES - Kids' Speech - SCOTS - BNC	- Kids' Speech - COLT - LIC - IViE Corpus - SACODEYL - BNC

			• ICE
Written		• CLS (Lancaster) • LUCY Corpus • LCCPW	• LUCY Corpus • LOCNESS • Mini-McCALL • ARCHER
SMS	• CorTxt • Treasuremytext • NUS SMS Corpus • HKU SMS Corpus • Shortis		
IM			
Chat	• NPS Corpus		• NPS Corpus
Email	• Enron Corpus • SpamAssassin • Min-McCALL		• Mini-McCALL • Teenage Health Freak website
Other CMC	• Mini-McCALL • Yahoo! Answers • Usenet • 20 Newsgroups		• Mini-McCALL (forum data)

Table 4 - Summary of corpora relevant to our survey.

In this paper, we have carried out a survey of two types of corpora; child language and CMC language plus the overlap between the two. Previous surveys of language corpora have focussed on well-known corpora or corpora for specific languages, rather than language produced by children online. Our original motivation for this was a project specific one, driven by the needs of our online child protection research and the work with law enforcement agencies to detect adults masquerading as children online. However, we have seen through this survey the paucity of corpora containing child language and online varieties of language with verifiable metadata such as age and gender. This is a drawback not only to our own research but to others wishing to carry out replicable studies in these areas using the kinds of reference corpora that corpus linguists have become accustomed to using in their research.

Numerous studies and projects have been carried out where a new dataset is collected, analysed and results published without release of the underlying data. This is most likely due to ethical restrictions on the secondary analysis and release of project data. Unfortunately, this duplication of work and lack of comparability of results has negative consequences for the linguistic research community as a whole. Where previous studies have made use of age and gender information in online corpora, in some cases they use self-reported information which has been found to be inaccurate in other studies.

We focussed in the introduction on two main uses for online and child-language corpora. For the first use, building age and gender-based profiles of language, we required verifiable age and gender metadata. Here, our survey has found very few available corpus resources. For the second use or application, that of retraining our NLP tools to deal with child and online language text types, the metadata requirements are less strict and we have seen that very large corpora of email and Usenet data are available, for example. In addition, for online CMC data, it is certainly possible to utilise web-crawlers that will collect vast quantities of such data. However, even in this scenario, legal and ethical restrictions may apply, although this is still a grey area for corpus linguists.

Whilst the specific needs of the Isis project have been largely unfulfilled by the available corpora, various corpora have been highlighted which may be of use for future research in the fields of child language and CMC. It is hoped that the highlighted scarcity of suitable corpus data representing children online will encourage future projects to establish new datasets for research use, and that researchers will consider making their own datasets available for wider research wherever possible.

Acknowledgements

This research is co-funded by the EPSRC and ESRC (EP/F035438/1). Project title: “Isis: Protecting children in online social networks”.

Notes

[1] i.e. communication through computer networks
[2] http://www.comp.lancs.ac.uk/isis/
[3] http://tiny.cc/corpora
[4] http://www.helsinki.fi/varieng/CoRD/
[5] http://www.clarin.eu/view_resources
[6] The current focus of our project is to work with UK law-enforcement agencies who will be investigating primarily British targets.
[7] http://khnt.hit.uib.no/icame/manuals/pow.htm
[8] http://childes.psy.cmu.edu/
[9] Described in more detail at http://childes.psy.cmu.edu/manuals/
[10] An online manual exists at http://childes.psy.cmu.edu/manuals/chat.pdf
[11] An online manual exists at http://childes.psy.cmu.edu/manuals/clan.pdf
[12] See http://talkbank.org/share/
[13] http://cslu.cse.ogi.edu/corpora/kids/
[14] see http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2007S18
[15] http://torvald.aksis.uib.no/colt/
[16] http://www.ling.lancs.ac.uk/activities/278/
[17] http://www.phon.ox.ac.uk/files/apps/IViE/
[18] http://www.um.es/sacodeyl/
[19] http://www.um.es/sacodeyl/en/pages/license.htm
[20] TEI encodes spelling ‘corrections’ in the form <orig>gonna</orig><corr>going to</corr>.
[21] http://www.grsampson.net/RLucy.html
[22] http://www.lancs.ac.uk/fass/faculty/activities/540/
[23] http://www.lancs.ac.uk/fass/projects/lever/index.htm
[24] http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Icle/LOCNESS1.htm
[25] http://www.scottishcorpus.ac.uk/
[26] http://www.natcorp.ox.ac.uk/
[27] Note: The COLT corpus forms part of the BNC, so some overlap will occur.
[28] http://ice-corpora.net/ice
[29] see, for example, http://www.netlingo.com
[30] http://www.txt2nite.com/forum/viewtopic.php?t=247
[31] http://www.treasuremytext.com/
[32] http://www.cel.iitkgp.ernet.in/~monojit/sms.html
[33] http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/
[34] http://wing.comp.nus.edu.sg/SMSCorpus/
[35] http://www.hku.hk/linguist/research/bodomo/MPC/
[36] http://www.netting-it.com/
[37] http://www.le.ac.uk/pc/aa/ked6/index.html
[38] http://www.sms4science.org/
[39] http://faculty.nps.edu/cmartell/NPSChat.htm
[40] http://www.cs.cmu.edu/~enron/
[41] http://spamassassin.apache.org/publiccorpus/
[42] http://people.csail.mit.edu/jrennie/20Newsgroups/
[43] Although the authors state that the corpus is publicly available, our search failed to find the data
[44] http://wordpress.com/
[45] http://www.blogger.com/
[46] http://twitter.com/
[47] http://answers.yahoo.com/

[48] `http://www.stanford.edu/class/cs345a/YahooData.pdf`

[49] `http://www.vbulletin.com/`

[50] `http://www.phpbb.com/`

[51] `http://socscibot.wlv.ac.uk`

[52] `http://www.teenagehealthfreak.org/`

References

Acharyya, S., S. Negi, L. V. Subramaniam, and S. Roy. 2009. "Language independent unsupervised learning of short message service dialect". *International Journal on Document Analysis and Recognition*, 12 (3), 175–184.

Adams, P. H. and C. H. Martell. 2008. "Topic detection and extraction in chat". In *ICSC '08: Proceedings of the 2008 IEEE International Conference on Semantic Computing*, 581–588, Washington, DC, USA. IEEE Computer Society.

Adolphs, S. and L. Mullany. 2009. "Health communication and the Internet: An analysis of adolescent language use on the teenage health freak website". ESRC Research Grant Award.

af Segerstad, Y. H. 2002. *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. PhD thesis, Department of Linguistics, Gö teborg University, Sweden.

Agarwal, S., S. Godbole, D. Punjani, and S. Roy. 2007. "How much noise is too much: A study in automatic text classification". In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, 3–12, Washington, DC, USA. IEEE Computer Society.

Al-Sa'di, R. A. and J. M. Hamdan. 2005. ""synchronous online chat" English: Computer-mediated communication". *World Englishes*, 24 (4), 409–424.

Anderson, J., D. Beavan, and C. Kay. 2007. "SCOTS: Scottish Corpus of Texts and Speech". In Beal, J., K. Corrigan, and H. Moisl (Eds.), *Synchronic Databases*, 1 of *Creating and Digitizing Language Corpora*. Basingstoke: Palgrave Macmillan.

Anderson, W. J. 2007. "The SCOTS Corpus: a resource for language contact study". In Ureland, P., A. Lodge, and S. Pugh (Eds.), *Language contact and minority languages in Europe*, 5 of *Studies in Eurolinguistics*. Berlin: Logos Verlag.

Anderwald, L. and B. Szmrecsanyi. 2009. "Corpus linguistics and dialectology". In Lüdeling, A. and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, 2 of *Handbooks of Linguistics and Communication Science*, chapter 53, 1126–1139. Berlin and New York: Mouton de Gruyter.

Androutsopoulos, J. and M. Beißwenger. 2008. "Introduction: Data and methods in computer-mediated discourse analysis". *Language@Internet*, 5.

Argamon, S., M. Šarić, and S. S. Stein. 2003. "Style mining of electronic messages for multiple authorship discrimination: first results". In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 475–480, New York, NY, USA. ACM.

Aw, A., M. Zhang, J. Xiao, and J. Su. 2006. "A phrase-based statistical model for SMS text normalization". In *Proceedings of the COLING/ACL on Main conference poster sessions*, 33–40, Morristown, NJ, USA. Association for Computational Linguistics.

Baker, P. 2009. "The BE06 corpus of British English and recent language change". *International Journal of Corpus Linguistics*, 14 (3), 312–337.

Baron, A. and P. Rayson. 2009. "Automatic standardisation of texts containing spelling variation: How much training data do you need?". In *Online Proceedings of Corpus Linguistics 2009*. University of Liverpool.

Baron, N. S. 1998. "Letters by phone or speech by other means: the linguistics of email". *Language and Communication*, 18 (2), 133–170.

Baron, N. S. 2003. "Why email looks like speech: Proofreading, pedagogy, and public face". In Aitchison, J. and D. M. Lewis (Eds.), *New Media Language*, chapter 9. London, UK: Routledge.

Baron, N. S. 2004. “See you online: Gender issues in college student use of instant messaging”. *Journal of Language and Social Psychology*, 23 (4), 397–423.

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. “The WaCky wide web: a collection of very large linguistically processed web-crawled corpora”. *Language Resources and Evaluation*, 43 (3), 209–226.

Beaufort, R., S. Roekhaut, L.-A. Cougnon, and C. Fairon. 2010. “A hybrid rule/model-based finite-state framework for normalizing SMS messages”. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 770–779, Uppsala, Sweden. Association for Computational Linguistics.

Beißwenger, M. and A. Storrer. 2009. “Corpora of computer-mediated communication”. In Lüdeling, A. and M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, 1 of *Handbooks of Linguistics and Communication Science*, chapter 17, 292–308. Berlin and New York: Mouton de Gruyter.

Bengel, J., S. Gauch, E. Mittur, and R. Vijayaraghavan. 2004. “Chattrack: Chat room topic detection using classification”. *Intelligence and Security Informatics*, 266–277.

Berman, R. A. and B. Nir-Sagiv. 2007. “Comparing narrative and expository text construction across adolescence : A developmental paradox”. *Discourse Processes*, 43 (2), 79–120.

Berman, R. A. and L. Verhoeven. 2002. “Cross-linguistic perspectives on the development of text-production abilities: Speech and writing”. *Written Language and Literacy*, 5 (1-2), 1–43.

Biber D., S. Conrad, and R. Reppen. 1998. “Appendix: commercially available corpora and analytical tools”. In Biber, D., S. Conrad, and R. Reppen (Eds.), *Corpus Linguistics: investigating language structure and use*, 281-287. Cambridge: Cambridge University Press.

Bieswanger, M. 2007. “2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different shortening strategies in English and German text messages”. In Hallett, T., S. Floyd, S. Oshima, and A. Shield (Eds.), *Texas Linguistic Forum*, 50, Austin, Texas. University of Texas, Austin.

Boneva, B., A. Quinn, R. Kraut, S. Kiesler, and I. Shklovski. 2006. “Teenage communication in the instant messaging era”. In Kraut, R., M. Brynin, and S. Kiesler (Eds.), *Computers, Phones, and the Internet: Domesticating Information Technology*, Series in Human-Technology Interaction, chapter 14, 201–218. Oxford and New York: Oxford University Press.

Boyd, D. M. and N. B. Ellison. 2008. “Social network sites: Definition, history, and scholarship”. Journal of Computer-Mediated Communication, 13 (1), 210–230.

Bryant, J. A., A. Sanders-Jackson, and A. M. K. Smallwood. 2006. “IMing, text messaging, and adolescent social networks”. *Journal of Computer-Mediated Communication*, 11 (2), 577–592.

Çamtepe, A., M. S. Krishnamoorthy, and B. Yener. 2004. “A tool for Internet chatroom surveillance”. *Intelligence and Security Informatics, 252*–265.

Carterette, E. C. and M. H. Jones. 1974. *Informal Speech: Alphabetic and Phonemic texts with statistical analyses and tables*. Berkley, California: University of California Press.

Chafe, W. and D. Tannen. 1987. “The relation between written and spoken language”. *Annual Review of Anthropology*, 16, 383–407.

Cherny, L. 1999. *Conversation and Community: Chat in a Virtual World*. Stanford, CA, USA: CSLI Publications.

Chipere, N., D. Malvern, B. Richards, and P. Duran. 2001. “Using a corpus of school children’s writing to investigate the development of vocabulary diversity”. In Rayson, P., A. Wilson, T. McEnery, A. Hardie, and S. Khoja (Eds.), *Proceedings of Corpus Linguistics 2001*, 13 of *UCREL Technical Papers*, 126–133. UCREL, Lancaster University.

Chipere, N., D. Malvern, and B. Richards. 2004. “Using a corpus of children’s writing to test a solution to the sample size problem affecting type-token ratios”. In Ashton, G., S. Bernardini, and D. Stewart (Eds.), *Corpora and Language Learners*, 17 of *Studies in Corpus Linguistics*, 137–147. Amsterdam and Philadelphia: John Benjamins.

Choudhury, M., R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. 2007. “Investigation and modeling of the structure of texting language”. *International Journal on Document Analysis and Recognition*, 10 (3), 157–174.

Clark, A. 2003. “Pre-processing very noisy text”. In *Proceedings of the Workshop on Shallow*

Processing of Large Corpora at Corpus Lingusitics 2003.

Clark, L. S. 1998. “Dating on the net: Teens and the rise of “pure” relationships”. In Jones, S. G. (Ed.), *Cybersociety 2.0: revisiting computer-mediated communication and community*, chapter 6, 159–183. Thousand Oaks, CA, USA: Sage Publications.

Collot, M. and N. Belmore. 1996. “Electronic language: A new variety of English”. In Herring, S. C. (Ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, 39 of Pragmatics and Beyond: New Series, 13–28. Amsterdam: John Benjamins.

Cook, P. and S. Stevenson. 2009. “An unsupervised model for text message normalization”. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity,* 71–78, Boulder, Colorado. Assocation for Computational Linguistics.

Corney, M., O. de Vel, A. Anderson, and G. Mohay. 2002. “Gender-preferential text mining of email discourse”. In *18th Annual Proceedings of Computer Security Applications Conference*, 282–289. Los Alamitos, CA, USA: IEEE Computer Society.

Corney, M. W. 2003. *Analysing E-mail Text Authorship for Forensic Purposes*. MA thesis, Queensland University of Technology.

Craig, D. 2003. “Instant messaging: The language of youth literacy”. Technical report, Stanford, CA, USA.

Crystal, D. 2001. *Language and the Internet.* Cambridge: Cambridge University Press.

Crystal, D. 2004. *A Glossary of Netspeak and Textspeak.* Edinburgh University Press.

Crystal, D. 2008. *Txting: The Gr8 Db8.* Oxford University Press.

Danet, B. 1998. “Text as mask: gender, play, and performance on the Internet”. In Jones, S. G. (Ed.), *Cybersociety 2.0: revisiting computer-mediated communication and community*, chapter 5, 129–158. Thousand Oaks, CA, USA: Sage Publications.

Daumé III, H., T. Deoskar, D. McClosky, B. Plank, and J. Tiedemann (Eds.). 2010. *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Uppsala, Sweden: Association for Computational Linguistics.

Davies, M. 2009. “The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights”. *International Journal of Corpus Linguistics*, 14 (2), 159–190.

de Vel, O., A. Anderson, M. Corney, and G. Mohay. 2001. “Multi-topic e-mail authorship attribution forensics”. In *Proceedings of the Workshop on Data Mining for Security Applications,* 8th ACM Conference on Computer Security (CCS’2001).

de Vel, O., M. Corney, A. Anderson, and G. Mohay. 2002. “Language and gender author cohort analysis of e-mail for computer forensics”. In *Proceedings of Digital Forensic Research Workshop*.

December, J. 1996. “Units of analysis for Internet communication”. *Journal of Communication*, 46 (1), 14–38.

Deutschmann, M., A. Ädel, G. Garretson, and T. Walker. 2009. “Introducing Mini-McCALL: A pilot version of the Mid-Sweden corpus of computer-assisted language learning”. *ICAME Journal*, 33, 21–44.

Donath, J. 1999. “Identity and deception in the virtual community”. In Kollock, P. and M. Smith (Eds.), *Communities in Cyberspace*, 29–59. London: Routledge.

Driscoll, D. 2002. “The ubercool morphology of Internet gamers: A linguistic analysis”. *Undergraduate Research Journal for the Human Sciences*, 1.

Edwards, J. A. 1993. “Survey of electronic corpora and related resources for language researchers”. In Edwards, J. A. and M. D. Lampert (Eds.), *Talking data: transcription and coding in discourse research*, 263–307. New Jersey: Lawrence Erlbaum Associates.

Eldridge, M. and R. Grinter. 2001. “Studying text messaging in teenagers”. In *CHI 2001 Workshop 1: Mobile Communications: Understanding User, Adoption and Design*.

Eysenbach, G. and J. E. Till. 2001. “Ethical issues in qualitative research on Internet communities”. *BMJ*, 323 (7321), 1103–1105.

Ferrara, K., H. Brunner, and G. Whittemore. 1991. “Interactive written discourse as an emergent register”. *Written Communication*, 8 (1), 8–34.

Fletcher, P. and M. Garman. 1988. “Normal language development and language impairment:

Syntax and beyond". *Clinical Linguistics and Phonetics*, 2 (2), 97–113.

Forsyth, E. N. 2007. *Improving Automated Lexical And Discourse Analysis of Online Chat Dialog*. PhD thesis, Naval Postgraduate School, Monerey, CA.

Forsyth, E. N. and C. H. Martell. 2007. "Lexical and discourse analysis of online chat dialog". In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, 19–26, Washington, DC, USA. IEEE Computer Society.

Foster-Cohen, S. H. 1999. *An Introduction to Child Language Development*. London and New York: Longman.

Gabrielatos, C., E. N. Torgersen, S. Hoffmann, and S. Fox. 2010. "A corpus-based sociolinguistic study of indefinite article forms in London English". *Journal of English Linguistics*, 38 (4), 297–334.

Gathercole, V. C. 1979. *Birdies like birdseed the bester than buns: a study of relational comparatives and their acquisition*. PhD thesis, University of Kansas.

Gathercole, V. C. 1986. "The acquisition of the present perfect: explaining differences in the speech of Scottish and American children". *Journal of Child Language*, 13 (3), 537–560.

Grabe, E. 2004. "Intonational variation in urban dialects of English spoken in the British Isles". In Gilles, P. and J. Peters (Eds.), *Regional Variation in Intonation*, 9–31. Tuebingen: Niemeyer.

Grabe, E. and B. Post. 2002. "Intonational variation in the British Isles". In *Speech Prosody 2002*, 343–346.

Granger, S. (Ed.). 1998. *Learner English on Computer*. Studies in Language and Linguistics. London and New York: Longman.

Greenfield, P. M. and K. Subrahmanyam. 2003. "Online discourse in a teen chatroom: New codes and new modes of coherence in a visual medium". *Journal of Applied Developmental Psychology*, 24 (6), 713–738.

Grinter, R. E. and M. A. Eldridge. 2001. "y do tngrs luv 2 txt msg?". In *ECSCW'01: Proceedings of the seventh conference on European Conference on Computer Supported Cooperative Work*, 219–238, Norwell, MA, USA. Kluwer Academic Publishers.

Grinter, R. E. and M. A. Eldridge. 2003. "Wan2tlk?: everyday text messaging". In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, 441–448, New York, NY, USA. ACM.

Grinter, R. E. and L. Palen. 2002. "Instant messaging in teen life". In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, CSCW '02, 21–30, New York, NY, USA. ACM.

Grinter, R. E., L. Palen, and M. Eldridge. 2006. "Chatting with teenagers: Considering the place of chat technologies in teen life". *ACM Transactions Computer-Human Interaction*, 13 (4), 423–447.

Harvey, K., D. Churchill, P. Crawford, B. Brown, L. Mullany, A. Macfarlane, and A. McPherson. 2008. "Health communication and adolescents: what do their emails tell us?". *Family Practice*, 25.

Harvey, K. J., B. Brown, P. Crawford, A. Macfarlane, and A. McPherson. 2007. "'am i normal?' teenagers, sexual health and the internet". *Social Science & Medicine*, 65 (4), 771–781.

Haslerud, V. and A.-B. Stenström. 1995. "The Bergen Corpus of London Teenager Language (COLT)". In Leech, G., G. Myers, and J. Thomas (Eds.), *Spoken English on Computer: Transcription, mark-up and application*, chapter 20, 235–242. New York: Longman.

Herring, S. C. 2003. "Computer-mediated discourse". In Schiffrin, D., D. Tannen, and H. E. Hamilton (Eds.), *Handbook of Discourse Analysis*, chapter 31. Oxford: Wiley-Blackwell.

Herring, S. C. 2007. "A faceted classification scheme for computer-mediated discourse". *Language@Internet*, 4.

Hinduja, S. and J. W. Patchin. 2008. "Personal information of adolescents on the Internet: A quantitative content analysis of MySpace". *Journal of Adolescence*, 31 (1), 125–146.

Hoffmann, S. 2007. "Processing Internet-derived text – creating a corpus of Usenet messages". *Literacy and Linguistic Computing*, 22 (2), 151–165.

Hoffstaedter, P. and K. Kohn. 2009. "Real language and relevant language learning activities: Insights from the SACODEYL project". In Kirchhofer, A. and J. Schwarzkopf (Eds.), *The Workings of the Anglosphere - Contributions to the Study of British and US-American*

Cultures, presented to Richard Stinshoff. Trier: Wissenschaftlicher Verlag Trier.

Hofland, K. and S. Johansson. 1982. *Word frequencies in British and American English*. Bergen, Norway: The Norwegian Computing Centre for the Humanities.

Holmer, T. 2008. "Discourse structure analysis of chat communication". *Language@Internet*, 5.

Honeycutt, L. 2001. "Comparing e-mail and synchronous conferencing in online peer response". *Written Communication*, 18 (1), 26–60.

How, Y. and M.-Y. Kan. 2005. "Optimizing predictive text entry for short message service on mobile phones". In Smith, M. J. and G. Salvendy (Eds.), *Proceedings of Human Computer Interfaces International 2005 (HCII 05)*, Las Vegas. Lawrence Erlbaum Associates.

Hudson, J. M. and A. Bruckman. 2004. ""go away": Participant objections to being studied and the ethics of chatroom research". *The Information Society: An International Journal*, 20 (2), 127–139.

Huffaker, D. A. and S. L. Calvert. 2005. "Gender, identity and language use in teenage blogs". *Journal of Computer-Mediated Communication*, 10 (2), article 1.

Hyland, K. and J. Milton. 1997. "Qualification and certainty in l1 and l2 students' writing". *Journal of Second Language Writing*, 6 (2), 183–205.

Jacobson, D. 1999. "Doing research in cyberspace". *Field Methods*, 11 (2), 127–145.

Java, A., X. Song, T. Finin, and B. Tseng. 2007. "Why we Twitter: understanding microblogging usage and communities". In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56–65, New York, NY, USA. ACM.

Jones, G. M. and B. B. Schieffelin. 2009. "Enquoting voices, accomplishing talk: Uses of be + like in instant messaging". *Language and Communication*, 29 (1), 77–113.

Kalman, Y. M., G. Ravid, D. R. Raban, and S. Rafaeli. 2006. "Pauses and response latencies: A chronemic analysis of asynchronous CMC". *Journal of Computer-Mediated Communication*, 12 (1), 1–23.

Kang, H.-S. and H.-D. Yang. 2006. "The visual characteristics of avatars in computer-mediated communication: Comparison of internet relay chat and instant messenger as of 2003". *International Journal of Human-Computer Studies*, 64 (12), 1173–1183.

Karlgren, J. 2006. "Preface to the proceedings of the workshop on NEW TEXT: Wikis and blogs and other dynamic text sources". In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Kendall, L. 2000. ""OH NO! I'M A NERD!": Hegemonic masculinity on an online forum". *Gender & Society*, 14 (2), 256–274.

Klimt, B. and Y. Yang. 2004. "Introducing the Enron corpus". In *First Conference on Email and Anti-Spam (CEAS) 2004 Proceedings*, Mountain View, California, USA.

Kobus, C., F. Yvon, and G. Damnati. 2008. "Normalizing SMS: are two metaphors better than one?". In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, 441–448. Association for Computational Linguistics.

Kochanski, G., E. Grabe, J. Coleman, and B. Rosner. 2005. "Loudness predicts prominence: Fundamental frequency lends little". *The Journal of the Acoustical Society of America*, 118 (2), 1038–1054.

Lee, D. Y. W. 2010. "What corpora are available?". In McCarthy, M. and A. O'Keeffe (Eds.), *The Routledge Handbook of Corpus Linguistics*, 107–121. Abingdon: Routledge.

Leech, G. 1992. "Corpus linguistics and theories of linguistic performance". In Svartvik, J. (Ed.), *Directions in Corpus Linguistics: proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 105–122. Berlin: Mouton de Gruyter.

Lewis, C. and B. Fabos. 2005. "Instant messaging, literacies, and social identities". *Reading Research Quarterly*, 40 (4), 470–501.

Lin, J. 2007. *Automatic Author Profiling of Online Chat Logs*. MA thesis, Naval Postgraduate School, Monerey, CA.

Ling, R. and N. S. Baron. 2007. "Text message and IM: Linguistic comparison of American college data". *Journal of Language and Social Psychology*, 26 (3), 291–298.

Lopresti, D., S. Roy, K. Schulz, and L. V. Subramaniam. 2008. "Foreword". In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, 303 of

ACM International Conference Proceeding Series, 1–1, New York, USA. ACM.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates, 3rd edition.

Mayer, M. and J. E. Till. 1996. “The Internet: A modern pandora’s box?”. *Quality of Life Research*, 5 (6), 568–571.

McEnery, T. and A. Wilson. 1996. “Appendix A: corpora mentioned in the text”. In *Corpus Linguistics*, 181–187. Edinburgh: Edinburgh University Press.

Mehler, A., S. Sharoff, and M. Santini (Eds.). 2011. *Genres on the web: Computational Models and Empirical Studies*, 42 of Text, Speech and Language Technology. Springer.

Merchant, G. 2001. “Teenagers in cyberspace: an investigation of language use and language change in Internet chatrooms”. *Journal of Research in Reading*, 24 (3), 293–306.

Milton, J. 1998. “Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment”. In Granger, S. (Ed.), *Learner English on Computer*, Studies in Language and Linguistics, chapter 14, 186–198. London and New York: Longman.

Milton, J. and K. Hyland. 1999. “Assertions in students’ academic essays: A comparison of English NS and NNS student writers”. In Berry, R., B. Asker, K. Hyland, and M. Lam (Eds.), *Language Analysis, Description and Pedagogy*, 147–161, Hong Kong. HKUST.

Miranda, E., L. Camp, L. Hemphill, and D. P. Wolf. 1992. “Development changes in children’s use of tense in narrative”. In *Boston University Conference on Language Development*, Boston.

Myslin, M. and S. T. Gries. 2010. “k dixez? A corpus study of Spanish Internet orthography”. *Literacy and Linguistic Computing*, 25 (1), 85–104.

Nikkhou, M. and K. Choukri. 2005. “Survey on Arabic language resources and tools in the Mediterranean countries. Nemlar project report.”. Technical report.

Nowak, K. L. and C. Rauh. 2006. “The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction”. *Journal of Computer-Mediated Communication*, 11 (1), 153–178.

Ooi, V. B. Y. 2001. “Aspects of computer-mediated communication for research in corpus linguistics”. *Language and Computers*, 36, 91–104.

Palomares, N. A. and E.-J. Lee. 2009. “Virtual gender identity: The linguistic assimilation to gendered avatars in computer-mediated communication”. *Journal of Language and Social Psychology*.

Perera, K. 1986. “Language acquisition and writing”. In Fletcher, P. and M. Garman (Eds.), *Language Acquisition: Studies in first language development*, chapter 23, 494–518. Cambridge University Press, second edition.

Petrović, S., M. Osborne, and V. Lavrenko. 2010. “The Edinburgh Twitter corpus”. In Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, 25–26, Los Angeles, California, USA. Association for Computational Linguistics.

Plester, B. and C. Wood. 2009. “Exploring relationships between traditional and new media literacies: British preteen texters at school”. *Journal of Computer-Mediated Communication*, 14 (4), 1108–1129.

Plester, B., C. Wood, and V. Bell. 2008. “Txt msg n school literacy: does texting and knowledge of text abbreviations adversely affect children’s literacy attainment?”. *Literacy*, 42 (3), 137–144.

Plester, B., C. Wood, and P. Joshi. 2009. “Exploring the relationship between children’s knowledge of text message abbreviations and school literacy outcomes”. *British Journal of Developmental Psychology*, 27 (1), 145–161.

Pooley, N., K. Alcock, K. Cain, A. Hardie, S. Hoffman, and P. Rayson. 2008. “Variability in child language”. In *Posters at ICAME 2008 Conference*, Ascona, Switzerland.

Pravec, N. A. 2002. “Survey of learner corpora”. *ICAME Journal*, 26, 81–114.

Pusch, C. 2002. “A survey of spoken language corpora in Romance”. In Pusch, C. and W. Raible (Eds.), *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache / Romance Corpus Linguistics: Corpora and spoken language*, 245–264. Tübingen: Narr.

Rafaeli, S., D. R. Raban, and G. Ravid. 2005. “Social and economic incentives in Google Answers”. In *ACM Group 2005 Conference*, Sanibel Island, Florida. ACM.

Ravid, G. and S. Rafaeli. 2004. “Asynchronous discussion groups as Small World and Scale Free

Networks". *First Monday*, 9 (9).

Rettie, R. 2003. "A comparison of four new communication technologies". In Jacko, J. A., C. Stephanidis, and D. Harris (Eds.), *Human-Computer Interaction: Theory and Practice*, 1, 686–690. Mahwah, New Jersey, USA and London, UK: Lawrence Erlbaum Associates.

Ringlstetter, C., K. U. Schulz, and S. Mihov. 2006. "Orthographic errors in web pages: Toward cleaner web corpora". *Computational Linguistics*, 32 (3), 295–340.

Sallis, P. and D. Kassabova. 2000. "Computer-mediated communication: experiments with e-mail readability". *Information Sciences*, 123 (1-2), 43–53.

Sampson, G. 2003. "The structure of children's writing: moving from spoken to adult written norms". In Granger, S. and S. Petch-Tyson (Eds.), *Extending the scope of corpus-based research: New applications, new challenges*. Amsterdam and New York: Rodopi.

Sampson, G. 2005. "The LUCY Corpus: Documentation".
http://www.grsampson.net/LucyDoc.html.

Schiano, D. J., C. P. Chen, J. Ginsberg, U. Gretarsdottir, M. Huddleston, and E. Isaacs. 2002. "Teen use of messaging media". In *Proceedings of ACM Conference on Human Factors in Computing Systems CHI '02*, 594–595. ACM Press.

Schleppegrell, M. J. 1989. *Functions of because in spoken discourse*. PhD thesis, Georgetown University, Washington University, USA.

Schleppegrell, M. J. 1991. "Paratactic because". *Journal of Pragmatics*, 16 (4), 323–337.

Schleppegrell, M. J. 2001. "Linguistic features of the language of schooling". *Linguistics and Education*, 12 (4), 431–459.

Schleppegrell, M. J. 2004. *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah, New Jersey, USA: Lawrence Erlbaum Associates.

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker. 2006. "Effects of age and gender on blogging". In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Schulze, M. 1999. "Substitution of paraverbal and nonverbal cues in the written medium of IRC". In Naumann, B. (Ed.), *Dialogue analysis and the mass media*, 65–82. Tübingen: Niemeyer.

Seale, C., S. Ziebland, and J. Charteris-Black. 2006. "Gender, cancer experience and Internet use: A comparative keyword analysis of interviews and online cancer support groups". *Social Science & Medicine*, 62 (10), 2577–2590.

Shaoul, C. and C. Westbury. 2011. *A USENET corpus (2005-2010).* Edmonton, Alberta: University of Alberta. Available from
http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html

Sharoff, S. 2006. "Open-source corpora: using the net to fish for linguistic data". *International Journal of Corpus Linguistics*, 11 (4), 435–462.

Shaw, P. 2008. "Spelling, accent and identity in computer-mediated communication". *English Today*, 24 (2), 42–49.

Shortis, T. 2001. *The Language of ICT: Information and Communication Technology*. London and New York: Routledge.

Shortis, T. 2007. "Gr8 Txtpectations: The Creativity of Text Spelling". *English, Drama, Media*, 8, 21–26.

Singh, S. 2001. "A pilot study on gender differences in conversational speech on lexical richness measures". *Literacy and Linguistic Computing*, 16 (3), 251–264.

Smith, N., T. McEnery, and R. Ivanic. 1998. "Issues in Transcribing a Corpus of Children's Handwritten Projects". *Literacy and Linguistic Computing*, 13 (4), 217–225.

Sofkova Hashemi, S. 2003. *Automatic Detection of Grammar Errors in Primary School Children's Texts: A Finite State Approach*. PhD thesis, Department of Linguistics, Göteborg University.

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

Sproat, R., A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. "Normalization of non-standard words". *Computer Speech and Language*, 15 (3), 287–333.

Stenström, A.-B., G. Anderson, and I. K. Hasund. 2002. *Trends in Teenage Talk: Corpus compilation, analysis and findings*. Studies in Corpus Linguistics. Amsterdam and Philadelphia: John Benjamins.

Stuart, K. 2006. “Towards an analysis of academic weblogs”. *Revisita Alicantina de Estudios Ingleses*, 19, 387–404.

Subrahmanyam, K., P. M. Greenfield, and B. Tynes. 2004. “Constructing sexuality and identity in an online teen chat room”. *Journal of Applied Developmental Psychology*, 25 (6), 651–666.

Subrahmanyam, K., P. M. Greenfield, and D. Smahel. 2006. “Connecting developmental constructions to the Internet: Identity presentation and sexual exploration in online teen chat rooms”. *Developmental Psychology*, 42 (3), 395–406.

Surdeanu, M., M. Ciaramita, and H. Zaragoza. 2008. “Learning to rank answers on large online QA collections”. In *Proceedings of ACL-08: HLT*, 719–727, Columbus, Ohio, USA. Association for Computational Linguistics.

Sussman, N. M. and D. H. Tyson. 2000. “Sex and power: gender differences in computer-mediated interactions”. *Computers in Human Behavior*, 16 (4), 381–394.

Swan, M. 2005. *Practical English Usage*. Oxford University Press, third edition.

Tagg, C. 2009. *A corpus linguistics study of SMS text messaging*. PhD thesis, School of English, Drama and American and Canadian Studies, University of Birmingham, Birmingham, UK.

Tagliamonte, S. A. and D. Denis. 2008. “Linguistic ruin? LOL! Instant messaging and teen language”. *American Speech*, 83 (1), 3–34.

Tavosanis, M. 2007. “A causal classification of orthography errors in web texts”. In *Proceedings of AND 2007*, 99–106.

Thelwall, M. 2008a. “Social networks, gender, and friending: An analysis of MySpace member profiles”. *Journal of the American Society for Information Science and Technology*, 59 (8), 1312–1330.

Thelwall, M. 2008b. “Fk yea I swear: cursing and gender in MySpace”. *Corpora*, 3 (1), 83–107.

Thelwall, M. and D. Stuart. 2007. “RUOK? blogging communication technologies during crises”. *Journal of Computer-Mediated Communication*, 12 (2), 523–548.

Thomson, R. 2006. “The effect of topic of discussion on gendered language in computer-mediated communication discussion”. *Journal of Language and Social Psychology*, 25 (2), 167–178.

Thomson, R. and T. Murachver. 2001. “Predicting gender from electronic discourse”. *British Journal of Social Psychology*, 40 (2), 193–208.

Thurlow, C. 2003. “Generation txt? the sociolinguistics of young people’s text-messaging”. *Discourse Analysis Online*, 1 (1).

Thurlow, C. and M. Poff. Forthcoming. “Text messaging”. In Herring, S. C., D. Stein, and T. Virtanen (Eds.), *Handbook of the Pragmatics of CMC*. Berlin and New York: Mouton de Gruyter.

Trudgill, P. 1999. *The Dialects of England*. Blackwell Publishing, second edition.

Trudgill, P. and J. Chambers (Eds.). 1991. Dialects of English: Studies in grammatical variation. *Longman Linguistics Library*. London and New York: Longman.

Tsuboi, Y. and Y. Matsumoto. 2002. “Authorship identification for heterogeneous documents”. *Joho Shori Gakkai Kenkyu Hokoku*, 17–24.

Tuulos, V. H. and H. Tirri. 2004. “Combining topic models and social networks for chat data mining”. In *WI ’04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, 206–213, Washington, DC, USA. IEEE Computer Society.

Tynes, B., L. Reynolds, and P. M. Greenfield. 2004. “Adolescence, race, and ethnicity on the Internet: A comparison of discourse in monitored vs. unmonitored chat rooms”. *Journal of Applied Developmental Psychology*, 25 (6), 667–684.

University of Leicester. 2006. “Texts to reveal ’whodunnit”’. Press Release.

Van Dyke, N. W., H. Lieberman, and P. Maes. 1999. “Butterfly: a conversation-finding agent for Internet Relay Chat”. In *IUI ’99: Proceedings of the 4th international conference on Intelligent user interfaces*, 39–41, New York, NY, USA. ACM.

Varnhagen, C., G. Mcfall, N. Pugh, L. Routledge, H. Sumida-Macdonald, and T. Kwong. 2009. “’lol’: new language and spelling in instant messaging”. *Reading and Writing*, Online First.

Voida, A., W. C. Newstetter, and E. D. Mynatt. 2002. “When conventions collide: the tensions of instant messaging attributed”. In *CHI ’02: Proceedings of the SIGCHI conference on Human factors in computing systems*, 187–194, New York, NY, USA. ACM.

Waldvogel, J. 2007. “Greetings and closings in workplace email”. *Journal of Computer-Mediated*

Communication, 12 (2), 456–477.

Wales, K. 2000. “North and South: An English linguistic divide”. *English Today*, 16 (1), 4–15.

Wang, J. 2001. “Recent progress in corpus linguistics in China”. *International Journal of Corpus Linguistics*, 6 (2), 281–304.

Waseleski, C. 2006. “Gender and the use of exclamation points in computer-mediated communication: An analysis of exclamations posted to two electronic discussion lists”. *Journal of Computer-Mediated Communication*, 11 (4), 1012–1024.

Werry, C. C. 1996. “Linguistic and interactional features of Internet Relay Chat”. In Herring, S. C. (Ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, 39 of *Pragmatics and Beyond: New Series*, 47–63. Amsterdam: John Benjamins.

Witmer, D. F. and S. L. Katzman. 2006. “On-line smiles: Does gender make a difference in the use of graphic accents?”. *Journal of Computer-Mediated Communication*, 2 (4).

Wong, W., W. Liu, and M. Bennamoun. 2006. “Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text”. In Peter, C., P. J. Kennedy, J. Li, S. J. Simoff, and G. J. Williams (Eds.), *Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006)*, CRPIT, 83–89, Sydney, Australia. ACS.

Wong, W., W. Liu, and M. Bennamoun. 2008. “Enhanced integrated scoring for cleaning dirty texts”. *IJCAI Workshop on Analytics for Noisy Unstructured Text Data (AND), 2007*, 55–62.

Xiao, R. 2008. “Well-known and influential corpora”. In Lüdeling, A. and M. Kyto (Eds.), *Corpus Linguistics: An International Handbook*, 1, 383–457. Berlin: Mouton de Gruyter.

Xiao-jun, Y. 2006. “Survey and prospect of China’s corpus-based research”. In Wilson, A., D. Archer, and P. Rayson (Eds.), *Corpus linguistics around the world*, 219–233. Amsterdam: Rodopi.

Yates, S. J. 1996. “Oral and written linguistic aspects of computer conferencing: A corpus based study”. In Herring, S. C. (Ed.), *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, 39 of *Pragmatics and Beyond: New Series*, 29–46. Amsterdam: John Benjamins.

Yates, S. J. 2001. “Gender, language and CMC for education”. *Learning and Instruction*, 11 (1), 21-34.

Yin, D., Z. Xue, L. Hong, B. D. Davidson, A. Kontostathis, and L. Edwards. 2009. “Detection of harassment on web 2.0”. In *CAW2.0 2009: Proceedings of Content Analysis in the Web 2.0 Workshop at WWW2009*, Madrid Spain.

Yvon, F. 2010. “Rewriting the orthography of SMS messages”. *Natural Language Engineering*, 16 (2), 133–159.

Zhu, C., J. Tang, H. Li, H. T. Ng, and T. Zhao. 2007. “A unified tagging approach to text normalization”. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 688–695, Prague, Czech Republic. Association for Computational Linguistics.

Zitzen, M. and D. Stein. 2004. “Chat and conversation: a case of transmedial stability?”. *Linguistics*, 42 (5), 983–1021.