Volume 29 — Number 2 — 2014 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics

Herausgegeben von/Edited by
Michael Beißwenger, Nelleke Oostdijk,
Angelika Storrer, Henk van den Heuvel

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 29 – 2014 – Heft 2
Gastherausgeber	Michael Beißwenger, Nelleke Oostdijk, Angelika Storrer, Henk van den Heuvel
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Editorial

Computer-mediated communication (CMC) is an umbrella term used for interpersonal communication mediated through computer networks and accessed via personal computers and/or mobile devices. Examples of CMC genres are written conversations in chats, online forums or instant messaging applications, tweets, comments on weblogs, conversations on Wikipedia talk pages and on "social network" sites (*Facebook* etc.), interactions in multimodal communication environments such as Skype, online role-playing games (MMORPGs) or *SecondLife*, SMS messages, or conversations via smart phone "apps" such as *WhatsApp* or *Threema*.

In the past two and a half decades, the use of CMC genres has become an important part of everyday communication. To support empirical research on these new forms of communication, standard text corpora need to be supplemented by linguistically annotated corpora covering the language use in CMC. Nevertheless, there have been no standards thus far for the representation of the structural peculiarities of CMC genres. In addition, it has become consistently apparent that NLP tools trained on written standard language (e.g. on newspaper corpora) do not perform in a satisfactory manner on CMC data.

This special issue of the JLCL gathers five contributions of scholars and projects who aim to close the "CMC gap" in the corpora landscape for several European languages: the French CoMeRe project (Chanier et al.), the South Tyrolian DiDi project (Glaznieks & Stemle), the German Wikipedia corpus (Margaretha & Lüngen), the Dutch VU Chat corpus (Spooren & van Charldorp), and the research on language variation in Dutch Twitter data (van Halteren & Oostdijk).

Thierry Chanier, *Céline Poudat*, *Benoit Sagot*, *Georges Antoniadis*, *Ciara Wigham*, *Linda Hriba*, *Julien Longhi* and *Djamé Seddah* present a TEI representation schema centered on the model of 'interaction space' that has been applied to four French corpora. The schema extends the scope of annotation issues to environments in which several CMC modalities (e.g. chat, email, and forums) are used simultaneously.

Aivars Glaznieks and *Egon W. Stemle* describe the DiDi project where they work with German CMC corpora of internet users from the Italian province of Bolzano – South Tyrol. One focus of this project is on the question of how L1 German speakers of South Tyrol from different age groups are using variants of German and other languages when communicating on social networking sites. The authors describe an approach for collecting Facebook postings and report on experiments to improve POS tagging results on their data by means of normalization.

Eliza Margaretha and *Harald Lüngen* present an approach developed at the IDS Mannheim for the transformation of Wikipedia articles and talk pages into TEI-based corpora for integration in the *German Reference Corpus* (Deutsches Referenzkorpus, DeReKo). The article's focus lies on issues of representing the conversations on talk pages in TEI. The authors describe a method for automatically segmenting these conversations into user postings and discuss the findings that arise from evaluating the segmentation results.

Wilbert Spooren and *Tessa van Charldorp* describe the design and data collection strategies of a chat corpus which is part of the SoNaR reference corpus for contemporary Dutch. To avoid the problem of collecting chat data "in the wild" with unclear legal status, the

authors created a setting in which data could be collected from a chatroom for secondary school students with the consent of both the participating pupils and their parents. The authors explain the logistical, ethical and technological challenges they encountered during the collection of the data and discuss general considerations regarding CMC data collection that can be derived from their experiences.

Hans van Halteren and *Nelleke Oostdijk* present results from their experiments in automatically estimating the proportions of word tokens in Dutch tweets that are not covered by standard resources and can therefore be expected to cause problems for standard NLP applications. Based on a fully annotated pilot corpus, the authors present a detailed typology of types of non-word tokens, out-of-vocabulary tokens and in-vocabulary tokens whose form deviates from standard Dutch. The annotated corpus was used to calibrate automatic estimation procedures which were then applied to about 2 billion Dutch tweets. The discussion of their results is an important foundation for getting a better picture of challenges that are faced in adapting NLP tools to the peculiarities of (Dutch) CMC data.

The idea for this special issue developed from an international workshop "*Building Corpora of Computer-Mediated Communication: Issues, Challenges, and Perspectives*", which was held at TU Dortmund University in February 2013 in connection with the German DFG network *"Empirical research of Internet-Based Communication"* (*Empirikom*)[1] and with financial support from the *Global Young Faculty* program of the Mercator Research Center Ruhr. The workshop brought together CMC corpus projects from several European countries. The participants discussed issues in collecting, representing, annotating and processing CMC data with the common goal of improving interoperability between CMC resources for different languages on the one hand, and between CMC corpora and standard text corpora on the other hand. The workshop resulted in the formation of a network of CMC corpus projects and a joint application for the installation of a special interest group (SIG) "Computer-mediated communication" in the *Text Encoding Initiative* (*TEI*)[2], which was approved by the TEI Council in the autumn of 2013. The articles by *Eliza Margaretha* and *Harald Lüngen* and *Thierry Chanier et al.* are explicitly related to this initiative; the ongoing work in this SIG is documented on the SIG pages in the TEI wiki[3].

Our gratitude goes to the colleagues who contributed to this special issue as external reviewers. We also thank Lothar Lemnitzer for supporting us while editing and finalizing the issue.

Dortmund, Mannheim and Nijmegen, December 2014

Michael Beißwenger
Nelleke Oostdijk
Angelika Storrer
Henk van den Heuvel

[1] http://www.empirikom.net
[2] http://tei-c.org
[3] http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis,
Ciara R. Wigham, Linda Hriba, Julien Longhi, Djamé Seddah

The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres

Abstract

The CoMeRe project aims to build a kernel corpus of different computer-mediated communication (CMC) genres with interactions in French as the main language, by assembling interactions stemming from networks such as the Internet or telecommunications, as well as mono and multimodal, and synchronous and asynchronous communications. Corpora are assembled using a standard, thanks to the Text Encoding Initiative (TEI) format. This implies extending, through a European endeavor, the TEI model of text, in order to encompass the richest and the more complex CMC genres. This paper presents the Interaction Space model. We explain how this model has been encoded within the TEI corpus header and body. The model is then instantiated through the first four corpora we have processed: three corpora where interactions occurred in single-modality environments (text chat, or SMS systems) and a fourth corpus where text chat, email, and forum modalities were used simultaneously.

The CoMeRe project has two main research perspectives: discourse analysis, only alluded to in this paper, and the linguistic study of idiolects occurring in different CMC genres. As natural language processing (NLP) algorithms are an indispensable prerequisite for such research, we present our motivations for applying an automatic annotation process to the CoMeRe corpora. Our wish to guarantee generic annotations meant we did not consider any processing beyond morphosyntactic labelling, but prioritized the automatic annotation of any freely variant elements within the corpora. We then turn to decisions made concerning which annotations to make for which units and describe the processing pipeline for adding these. All CoMeRe corpora are verified thanks to a multi-stage quality control process that is designed to allow corpora to move from one project phase to the next.

Public release of the CoMeRe corpora is a short-term goal: corpora will be integrated into the forthcoming *French National Reference Corpus*, and disseminated through the national linguistic infrastructure *Open Resources and Tools for Language* (ORTOLANG). We, therefore, highlight issues and decisions made concerning the OpenData perspective.

1 Introduction: the CoMeRe project

Various national reference corpora have been successfully developed and made available over the past few decades, e.g. the *British National Corpus* (Aston and Burnard 1998), the SoNaR *Reference Corpus of Contemporary Written Dutch* (Oostdijk et al. 2008), the DWDS *Corpus for the German Language of the 20th century* (Geyken 2007), the DeReKo *German Reference Corpus* (Kupietz and Keibel 2009) and the *Russian Reference Corpus* (Sharoff 2006). Despite being in strong demand, no French national reference corpus currently exists. Thus, the *Institut de la Langue Française* (ILF) has recently taken the first steps to lay the groundwork for such a project. The aim is for the national project to both collect existing

data and to develop new corpora, in order to ensure the representativeness of the final data set.

The French CoMeRe project (CoMeRe 2014)[1] is an ongoing pilot project whose deliverables will form part of the forthcoming *French National Reference Corpus*. It aims to build a kernel corpus of different computer-mediated communication (CMC) genres with interactions in French as the main language. Three fundamental principles underlie CoMeRe: variety, standards and openness.

"Variety" is one of our key words since we expect to assemble interactions stemming from networks such as the Internet or telecommunications (mobile phones), as well as mono and multimodal, and synchronous and asynchronous communications. Our interest covers genres such as text or oral chats, email, discussion forums, blogs, tweets, audio-graphic conferencing systems (conference systems with text, audio, and iconic signs for communication), even collaborative working/learning environments with verbal and nonverbal communication. A variety of discourse situations is also sought: public or more private conversations, as well as informal, learning, and professional situations. One part of our (sub)corpora is taken from existing corpora, since partners involved in the project had previously collected almost all the genres mentioned earlier. Other parts, such as Wikipedia talk pages, will be extracted from the Web following the recommendations of the *New Collections* workgroup.

"Standards" is our second key word. It refers to two different aspects of corpus linguistics. Firstly, corpora will be structured and referred to in a uniform way. The Text Encoding Initiative (TEI) format (Burnard & Bauman 2013) has been chosen jointly with our European partners, alongside existing metadata formats including the Dublin Core. The TEI is not only a format for corpus structure. First and foremost, it is a model of text. This model needs to be extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse, as we will discuss in Section 2. The European TEI-CMC (2013) special interest group (SIG) aims to propose such extensions to the TEI consortium.

"Standard" also refers to the uniform basic level of automatic annotations, related to segmentation and part of speech (POS) tagging. This will be applied to all of our CMC genres, and is presented hereafter in Section 3.

The third key word is "openness". At the end of the first stage (2013-2014) of the project, a sample of corpora (including those described in this paper, see Section 3.1) that are representative of CMC genres and that have been organized and processed in standard ways, will be released as open data on the French national platform of linguistic resources ORTOLANG (2014). Dissemination will take two different forms: one version of a corpus with the "raw" text without any tokenization and annotation (v1), and a second version of the same corpus with the annotations (v2). This openness is motivated, on the one hand, by the fact that CoMeRe will become part of the larger reference corpus for the French language; the latter is expected to become a reference for studies in French linguistics. On the other hand, the wish to release CoMeRe corpora as open data stems from the fact that, although studies

[1] CoMeRe stands for "Communication Médiée par les Réseaux," an updated equivalent to Computer-Mediated Communication (CMC) or Network-mediated Communication.

on new CMC communication genres draw much attention, there is currently no existing dataset with significant coverage or that encompasses a variety of genres to form the basis for systematic research. This situation is not specific to the French language, as aforementioned languages, which already benefit from reference corpora, also face the same challenge. That being said, a few genre-based corpora are being developed (*e.g.* Rehm et al. 2008). This may explain why a common motivation amongst European partners encouraged, from the outset, the design of a shared framework for the development of models of CMC genres. Indeed, there is a need for open-access corpora that can be cross examined in order to exemplify the way models could be instantiated.

This OpenData perspective paves the way for scientific examination, replication and cumulative research. Of course, this type of openness implies specific considerations of licenses, ethics and rights, as discussed in Section 4.2. In order to achieve this goal, CoMeRe is supported by the research consortium *Corpus-Écrits* (2014), a subsection of the national infrastructure Huma-Num (2014, cf. Digital Humanities), and ORTOLANG, French infrastructures linked to DARIAH (2014), the European infrastructure for humanities.

2 CoMeRe 2013: moving from existing data to models of CMC interaction

The CoMeRe project developed out of collaborations between researchers who had previously collected and structured different types of CMC corpora within their local teams. Once the project was officially underway, it was decided, building upon the SoNaR experience (Oostdijk et al., ibid), to organize workgroups (WG) with distinct tasks in the project: *TEI & Metadata, New Collections, Automatic Processing,* and *Quality*.

The present section firstly describes four of the corpora that individual researchers brought to the CoMeRe project (Section 2.1). Secondly, we discuss how these four corpora helped the *TEI & Metadata* WG to instantiate a model of CMC interaction, working collaboratively with the TEI-CMC SIG (Sections 2.2 and 2.3). Section 2.4 details how the same WG then structured corpora according to this model. The work of other WGs will be the focus of Sections 3 (*Automatic Processing* WG), 4 (*Quality*), and 5 (*New Collections*).

2.1 Gathering existing data

Illustrations in this article will be based on the first four corpora processed by the CoMeRe project in fall 2013. They were collected within the frameworks of national and/or international projects. After their conversion to the new TEI format, they were renamed *cmr-smsalpes*, *cmr-smslareunion*, *cmr-simuligne* and *cmr-getalp_org*.

Our first corpus, *cmr-getalp_org* (Falaise 2014), is a **text chat corpus**, collected from a public Internet Relay Chat (IRC) website. Eighty different discussion channels focusing on a variety of, mainly informal, topics were collected in 2004. The corpus includes more than three million messages. The first version of the corpus (Falaise 2005) had been organized using a simple eXtensible Markup Language (XML) structure.

The data we organized in *cmr-smsalpes* (Antoniadis, 2014) and *cmr-smslareunion* (Ledegen, 2014) emanated from the international project "Faites don de vos SMS à la science" (Fairon et al. 2006) that began in 2004 and was coordinated by the Institute for Computational Linguistics (CENTAL) of the Catholic University of Louvain (Belgium). The project,

named *sms4science*, aims to collect **SMS text messages** worldwide (Panckhurst et al. 2013). It regroups researchers from several countries to collaboratively conduct scientific research on a large number of languages with the objective of contributing to SMS message communication studies.

Data from *cmr-smslareunion* were issued between April and June 2008 within the framework of the first French investigation which led to the collection of 12,622 SMS messages sent by 884 participants. The *Laboratoire de recherche dans les espaces créolophones et francophones* (LCF) of the Université de La Réunion was responsible for the local coordination. As described in the project presentation (LaRéunion4Science, 2008), the uniqueness of the investigation in Réunion is the new scientific dimension that it adds: French-Creole bilingualism, the ludic neographies in SMS messages, and the communication practices of young people which are characterized by multiple alternating languages (French, Creole, English, and Spanish).

Data from *cmr-smsalpes* were collected in 2011 (Antoniadis et al. 2011). The corpus includes 22,117 messages sent by 359 participants mainly living in the French Alps. The project was coordinated by the *Laboratoire de linguistique et didactique des langues étrangères et maternelles* (LIDILEM) of the Université Stendhal in Grenoble.

For both SMS text message corpora, the harvesting of SMS messages required the intervention of technical partners. Indeed, the companies *Orange Informatique* and *Cirrus Private* were responsible for receiving the SMS messages and transferring them to the laboratories concerned. Researchers in charge of compiling data for the two corpora anonymized and structured the messages in different formats: XML for the French Alps corpus and in the form of a spreadsheet for the Réunion corpus. Note that researchers from Réunion also added manual annotations to the messages, providing orthographic transcription and language identification (either pidgin or French), as we will see further on.

Lastly, the *cmr-simuligne* corpus was built out of interaction data resulting from an online language learning course, Simuligne. Data have been extracted from the LETEC (LEarning and TEaching Corpus) Simuligne (Reffay et al. 2009), a corpus deposited in the MULCE repository (2013), which has its own XML schema. Sixty-seven participants (language learners, teachers, native speakers a.k.a. language experts) followed the same pedagogical scenario, but were divided into four groups. All interactions occurred within a Learning Management System (LMS), namely WebCT. They include text chat turns (7,000), emails (2,300), and forum messages (2,700). Since the LMS had no export facilities, data were extracted from its internal database by the LETEC corpus compiler, then structured and anonymized.

Such disparities in corpus compilation choices may have represented a major handicap for the CoMeRe project, particularly when in the linguistics field many individual researchers still pose the question as to whether spending the time to make data shareable and accessible is worthwhile. However, data heterogeneity soon turned into a real asset, favoring exchanges between project participants concerning the data collection contexts and different ways of interpreting the data, as well as increasing our motivation to design a common model and share different areas of expertise.

2.2 Rationales for modelling CMC discourse

Before determining the TEI-compliant structural markup of the corpus, the *TEI & Metadata* WG found it necessary to first settle on a common document model that would fit all of our CMC data as well as new collections of data to be added to the corpus repository in the future. Indeed, annotation is basically an interpretation and the TEI markup naturally encompasses hypotheses concerning what a text is and what it should be. Although the TEI was historically dedicated to the markup of literature texts, various extensions have been developed for the annotation of other genres and discourses, including poetry, dictionaries, language corpora or speech transcriptions.

If one wants to still apply the word "text" to a coherent and circumscribed set of CMC interactions, it is not so much in the sense developed by the TEI. Indeed, it would be closer to the meaning adopted by Baldry and Thibault (2006). These authors consider (ibid: 4) "texts to be meaning-making events whose functions are defined in particular social contexts," following Halliday (1989: 10) who declared that "any instance of living language that is playing a role some part in a context of situation, we shall call a text. It may be either spoken or written, or indeed in any other medium of expression that we like to think of."

Bearing the above in mind, we found it more relevant to start from a general framework, that we will term "Interaction Space" (see next section), encompassing, from the outset, the richest and the more complex CMC genres and situations. Therefore, we did not work genre by genre, nor with scales that would, for instance, oppose simple and complex situations (*e.g.* unimodal versus multimodal environments)—as stated, our goal is to release guidelines for *all* CMC documents and not for each CMC genre. This also explains why we did not limit ourselves solely to written communication. Indeed, written communication can be simultaneously combined with other modalities. For these reasons, the CoMeRe model takes multimodality into account and our approach is akin to the one adopted by the French research consortium IRCOM (2014). This consortium rejected the collection and study of oral corpora as self-contained elements and decided that it was preferable for oral and multimodal corpora to be studied within a common framework, before becoming part of the French reference corpus.

2.3 The notion of 'Interaction Space'

2.3.1 Interaction space: time, location, participants

An Interaction Space (henceforth referred to as IS) is an abstract concept, located in **time** (with a beginning and ending date with absolute time, hence a time frame) where interactions between **a set of participants** occur within an **online location** (see Figure 1 for a general overview). The online location is defined by the properties of the **set of environments** used by the set of participants. *Online* means that interactions have been transmitted through networks; Internet, Intranet, telephone, etc.

The set of participants is composed of **individual** members or **groups**. It can be a predefined learner group or a circumscribed interest group. A mandatory property of a group is the listing of its participants.

The range of types of interactions (and their related locations) is widespread. It is related to the environment(s) participants use and their corresponding modes and modalities.

2.3.2 Environment, mode and modality

An environment may be synchronous or asynchronous, mono or multimodal. **Modes** (text, oral, icon, image, gesture, etc.) are semiotic resources which support the simultaneous genesis of discourse and interaction. Attached to this sense of mode orienteered towards communication, we use the term **modality** as a specific way of realizing communication (this sense refers to the Human Computer Interaction field (Bellik and Teil 1992)). Within an environment, one mode may correspond to one modality, with its own grammar that constraints interactions. For example, the icon modality within an audio-graphic environment is composed of a finite set of icons (*raise hand, clap hand, is talking, momentarily absent*, etc.). In contrast, within an environment, one mode may correspond to several modalities: a text chat has a specific textual modality that is different from the modality of a collective word processor, although both are based on the same textual mode. Consequently, an interaction may be multimodal because several modes are used and/or several modalities (Chanier and Vetter 2006; see also Lamy and Hampel 2007 for another presentation).

Environments may be simple or complex. On one end of the scale, we find simple types with one environment based on one modality (e.g. one text chat system in the *cmr-getalp_org* corpus). On the other end of the scale, stand complex environments, such as the LMS of the aforementioned *cmr-simuligne* corpus, where several types of textual modalities are integrated, either synchronous—text chat— or asynchronous—email and forum), or in 3D environments, where several modes and modalities appear (see hereafter).

An environment offers the participants one or more locations/places in which to interact. For example, a conference system may have several rooms where a set of participants may work separately in sub-groups or gather in one place. In a 3D environment such as the synthetic world *Second Life*, a location may be an island or a plot. A plot may even be divided into small sub-plots where verbal communication (through text chat or audio chat) is impossible from one to another. Hence we say that participants are in the same location/place if they can interact at a given time. Notions of location and interaction are closely related and are defined by the affordances of the environment.

2.3.3 Interaction

As previously described, participants in the same IS can interact (but do not necessarily do it, cf. lurkers). They interact through input devices (microphone, keyboard, mouse, gloves, etc.), which let them use the modalities and output devices, mainly producing visual or oral signals. (These however, will not be described in this article). Hence when participants cannot hear nor see the other participants' actions, they are not in the same IS. Of course, participants may not be participants during the whole time frame of the IS. They can enter late, or leave early. Note that an IS may have a recursive structure: in an online course when the same participants interact over several weeks, different ISs will be created, correspondingly to different occurrences of interaction sessions.

In an IS, actions occur between participants. Let us call the trace of an action within an environment and one particular modality an "**act**". Acts are generated by participants, and sometimes by the system. Some of them may be considered as directly communicative (e.g. verbal acts in synchronous text or oral modalities). Others may not be directly communicative but may represent the cause of communicative reaction/interaction (e.g. when participants write collaboratively in an online word processor and comment on their work). Participants see and hear what others are doing. These actions may represent the rationale for participants to be there and to interact (produce something collectively). Hence the distinction between acts that are directly communicative, or not, is irrelevant.

A verbal act may be realized as an *en bloc* message or as an adaptive one. For example, there are situations where a participant does not plan an utterance as a one-shot process before it is sent as an *en bloc* message to a server, which in turn displays it to the other participants as a non-modifiable piece of language (e.g. as a text chat act which corresponds to what is generally called a chat turn) (Beißwenger et al. 2012). However, a participant's utterance (e.g. in an audio chat act) can also be planned, then modified in the throes of the interaction while taking into account other acts occurring in other modalities of communication (see Wigham and Chanier 2013 as an example).

Even if all the environments, corresponding to the first four corpora that we have processed, form the basis of our current presentation and all these corpora correspond to messages sent *en bloc*, our IS model still needs to take into account other corpora where this does not hold true. Within other multimodal environments from which we have already collected data and which we are currently processing, verbal (speech, text chat) and nonverbal acts occur simultaneously. The main purpose of transcriptions is then to describe interrelations amongst acts and within acts.

2.4 Describing the interaction space within TEI

Since TEI was the format adopted by national research networks (*Corpus-écrits* and IRCOM) and by the European TEI-CMC SIG, the challenge faced by the *TEI & Metadata* WG was to firstly find out how information related to the IS could be described within the TEI header, and secondly, decide how, within the corpus body, verbal acts could be coded in such a way that all information included in the original version of each corpus be kept.

The choice to adopt TEI was also motivated by two different research interests that members of CoMeRe shared: research on NLP models and research on discourse. The focus of these may appear quite different and although analysis work will only start once the CoMeRe corpora have been disseminated, it was important for the *TEI & Metadata* WG to keep both perspectives in mind when making TEI coding decisions.

One interest of CoMeRe members is to study linguistic idiolects occurring in different CMC genres. NLP algorithms are an indispensable prerequisite for this. However, it should be noted that NLP models may be developed solely on the contents of the verbal acts, whilst ignoring the rest of the IS. However, for other CoMeRe members interested in completing studies on discourse, the IS is fundamental. This especially holds true if members later want to study research questions such as: how does discourse organization vary from one situation to another? What type of interaction supports or hinders discourse amongst partici-

pants? What features of participant groups influence online interactions? What are the relationships between discourse organization and language complexity? These are current topics investigated by researchers in fields such as computer-supported collaborative learning (CSCL) and computer-assisted language learning (CALL).

The difference in the importance attributed to the IS when adopting one or other of these research perspectives seems, however, to be dialectical. Indeed research in CSCL and CALL may take advantage of linguistic annotations, which they previously have never considered, possibly because they had not been available to scientists in these fields.

We now move on to illustrate how the *TEI & Metadata* WG encoded the IS in TEI in the four corpora (*cmr-smsalpes*, *cmr-smslareunion*, *cmr-getalp_org* and *cmr-simuligne*) whilst taking the above research perspectives into account. Figure 1 illustrates the different concepts we introduced and which have to be described in TEI. Note that element <u>, used in speech transcriptions and the new (not yet present in TEI) element <prod> used in non-verbal transcriptions will not be presented, because they do not occur in the corpora used here as examples.

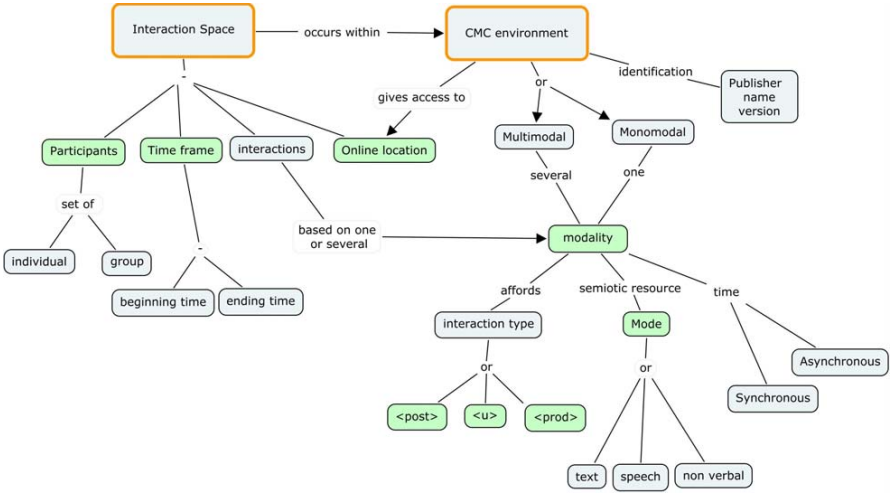

Figure 1: Description of concepts related to the Interaction Space

2.4.1 Environments and affordances

The first step when describing an environment is to define within the <teiHeader> the general features attached to the overall environment type to which it belongs (e.g., IRC text chat systems). However, this needs to be refined in order to elicit specific features of the system. For example, Figure 2, (2a) describes, in TEI, the general text chat modality where inside

one public channel[2] every connected participant may interact with the other participants in a spontaneous way through discussions held in informal settings, in contrast to educational or professional discussions. Example (2b), however, details the affordances related to the specific IRC system used in *cmr-getalp_org*. This simplified extract displays the three main types of chat actions (message, command, and event), and part of the subtype of events. Relationships between this definition of the environment in the <teiHeader> and its actual use by participants in interactions, described in the <body> part of the TEI file, will appear through the attribute @type of the <post> element (see next section).

(2a)

```
<textDesc xml:lang="en-GB">
  <channel mode="w" xml:lang="en-GB">
    <term ref="#texchat-epiknet">text
chat</term></channel>
  <constitution>Messages typed by partici-
pants inside EpikNet IRC Channels and then
collected by Botstats.com </constitution>
  <derivation type="original"/>
  <domain type="public"/>
  <factuality type="fact"/>
  <interaction type="complete" ac-
tive="plural" passive="many"/>
  <preparedness type="spontaneous"/>
  <purpose degree="high">
    <note>Informal discus-
sion</note></purpose>
  </textDesc>
```

(2b)

```
<classDecl>
 <taxonomy>
  <category xml:id="texchat-epiknet" />
  <category xml:id="chat-message"/>
  <category xml:id="chat-command"/>
  <category xml:id="chat-event">
    <category xml:id="connexion" />
    <category xml:id="deconnexion"/>
    <category xml:id="changementpseudo"/>
  </category>
        [...]
```

Figure 2: TEI description of a text chat environment in the <teiHeader>

Figure 2 illustrates a monomodal environment. Distinctively, when the environment is complex, such as the one related to the *cmr-simuligne* corpus where interactions happened in ISs based on text chat, email, or forum modalities (cf. Section 2.1), it is described in the same way in the <teiHeader> thanks to a more complex taxonomy: with one category per modality and each category having its own text description (<textDesc>). Here again, each category corresponds to a type of message appearing in the body of the corpus.

Besides its multimodal environment, the *cmr-simuligne* corpus has another more complex organization. On the LMS platform, there were four distinct interaction spaces where groups of participants completed the same activities. The participants within one group could only communicate with members of that group. These top level ISs have been encoded as distinct TEI texts, and all of them included within a <teiCorpus> file. Every TEI text in *cmr-*

[2] The TEI term "channel", which here corresponds to the environment, should not be confused with channels of the Internet Relay Chat (IRC) environment, where every channel correspond to a particular location of the IS.

simuligne is organized around sets of learning activities that are either simple or complex. A learning activity may include one or several modalities (email, chat or forums). The organization here is strikingly different to that adopted in other corpora. In *cmr-smsalpes*, *cmr-smslareunion*, and *cmr-getalp_org,* all messages are included within one division (element), whereas in *cmr-simuligne* there is one division per modality and a division may be nested several times.

2.4.2 A common post element

As agreed upon in the TEI-CMC SIG, we decided to use a common main new element, called a "post" in order to encode all verbal acts produced by a participant in a textual monomodal environment, prepared in advance by its author and sent *en bloc* to the server. The macro-structure[3] of the post may vary from one modality to another. Every structure is detailed in the header of the TEI files and is accompanied by comments that are of foremost importance because they describe constraints that researchers will have to take into account when conducting future analyses.

Figure 3 provides a simplified extract of the <teiHeader> that describes the structure of a SMS message, specifying how time events and participants' identifications should be interpreted.

```
<tagsDecl>
    [...]
  <post>one post corresponds to one SMS.
     @xml:id  ID of the posting.
     @when corresponds to the date of the message collected by the system. It
depends on the date the participant sent to the system, but not the date of the
conversation. Accordingly, one participant may have sent his/her messages to
his/her correspondent at different times, but may have assembled her messages
and sent them together to the server.
     @who is the anonymized telephone number. Hence one ID identifies one par-
ticipant over the whole corpus. If messages sent by the same participant (send-
er) may be studied, it should be noted that we have no information about the
receiver.
     [...]</post>
</tagsDecl>
```

Figure 3: Simplified structure of the <post> element for an SMS message as described in the <teiHeader>

Figure 4, extracted (and simplified) from *cmr-simuligne*, describes the structure of email and forum messages. In the latter case, each message/<post> in a thread is either the first message of the thread or a response to another message within the thread. The difference is simply made by the XML attribute @ref: A message without an @ref opens a thread whereas a message which has a @ref is an answer to another message and is consequently in-

[3] In this article, we only discuss text (taken in the Halliday (ibid.) sense) macrostructure: IS structure, message/<post> structure (its title, elements which include its contents, relationships with other messages, addressees, etc.). The micro-structure of the text refers to the type of elements found in the actual contents of the message/<post>, for example interaction words, emoticons, hash code, etc. See (Beißwenger et al. 2012) for linguistic consideration on the micro-structure.

cluded in a thread[4]. It has a title (<Title>), may have an attached file (<trailer>) and may also include a list of addressees (<listPerson>). When the message has been read (i.e. opened), this is noted within the structure (@type=Read). The name(s) of the reader(s), as well as the time(s) at which the message was read appear. The latter information is important when studying networks of participants interacting in a group (see, as an example, a CSCL analysis based on *Social Network Analysis* in Reffay and Chanier 2003).

```
<tagsDecl>
    [...]
 <post>one post corresponds to one email message or one forum message or one
text chat act.
   @xml:id ID of the post.
   @when date of the message when created, given by the system
   @who id of the author of the message.
   @type type of the post cf. taxononomy.
   @ref reference to the post ID to which the current post responded to (for
email and forum)
   <head> contains all the rest of the structure of the post, which cannot be
described as TEI elements.
    <title> Title of the forum, or subject of an email.
    <listPerson> list of people who received / read the post.
      @type=SendTo addressee(s) of an email
      @type=Read who opened (read?) an email or a forum message?
      [...]
   <trailer>At the end of a post when there is an attached file
 </tagsDecl>
```

Figure 4: Simplified presentation of the structure of an email or a forum message in the <teiHeader>

2.4.3 Locations and time frame

Locations and time frame are also components of the IS. Different notions of locations need to be distinguished: the server location where data was collected firsthand; locations attached to a modality (e.g., distinct chat rooms or channels); locations of participants (leaving areas, see below). Information on time is given at the level of the IS and also with every post. It is an indispensable component of the data, not only for studying interactions within one IS, but also for the study of group or individual activities within the overall corpus (for example by means of tools for displaying discussion forum time lines see Calico, 2013). For space reasons, we shall not detail here how locations and time frames have been encoded in TEI.

2.4.4 Participants

Since CoMeRe has collected different CMC genres, we have a large variety of participant description types—types which highly constrain further research analysis. On the one hand, in *cmr-smsalpes* and *cmr-smslareunion*, the only information we have about each participant is her/his identification number and information on his/her location given at a regional level (respectively in the French Alps or Réunion). On the other hand, in *cmr-simuligne*, we have

[4] This simple description of the structure of a forum (also used in analysis tools of forums based on XML structures such as Calico 2013) is a sufficient one. Describing the structure of the modality forum should not be confused with the visual description of a forum a participant can adjust when using it: threads of discussions visualized as a sequence of indented messages, or as messages ordered accordingly the date of posting, etc. Our structure includes all the information required for every specific visual display.

access to detailed information about participants (individuals and groups), as shown in Figure 5. An individual female learner, aged 51, who is affiliated to The Open University and who has adopted the alias *Alba* is detailed, as well as information about a learner group.

```
<particDesc>
  <listPerson>
   <person role="learner" xml:id="Gl1">
      <sex>female</sex><age value="51"/>
      <residence>United Kindom</residence>
      <affiliation>The Open University</affiliation >
      <persName><addName type="alias">Alba</addName></persName>
   </person >
         [other participants]
   <personGrp role="learnerGroup" xml:id="Simu-g-Ga">
      <persName><addName type="alias">Gallia</addName></persName>
   </personGrp>
        [other groups]
    <listRelation corresp="#Simu-g-Ga">
       <relation type="social" name="tutor" active="#Gt"/>
       <relation type="social" name="native" active="#Gn1 #Gn2"/>
       <relation type="social" name="learner" active="#Gl1 #Gl2 #Gl3 #Gl4 #Gl5
#Gl6 #Gl8 #Gl9 #Gl10"/>
       <relation type="social" name="researcher" active="#Tm"/>
    </listRelation>
```

Figure 5: Description of one participant, one group and relationships within a group

A common requirement in corpus linguistics is to associate each individual with a single identification code throughout the corpus. In CMC corpora, this is not always easy to achieve. On the one hand, in corpora built from experiments with a limited number of participants, such as *cmr-simuligne*, it was a tedious process to identify each participant every time s/he was named in a post (see example in a message forum in Figure 9). On the other hand, in a public chat channel, it may be difficult to identify participants due to constant changes in their alias names. In one case, analysis of individual contributions, activities, language level, lexical diversity, etc. can become an object of study. In the latter case, it is the variation in alias names which may be interesting to study: see Figure 6 taken from *cmr-getalp_org* where one participant uses suffixes attached to her/his alias in order to reflect different states of mind or activities (e.g. sport, school, busy, away, etc.).

```
    <person xml:id="cmr-get-c027-p4215">
      <persName>
          <addName type="alias">Farlin</addName>
          <addName type="alias">Farlin[AuStade]</addName>
          <addName type="alias">Farlin[AwAy]</addName>
          <addName type="alias">Farlin[IRL]</addName>
          <addName type="alias">Farlin[Lycee]</addName>
          <addName type="alias">Farlin[OqP]</addName>
          <addName type="alias">Farlin[Oral]</addName>
          <addName type="alias">Farlin[PALA]</addName>
           [...]
```

Figure 6: Variety of aliases chosen by one participant in a text chat

2.4.5 Examples of posts

Let us now consider examples of messages sent through different modalities. Whereas affordances of the Interaction Space were described previously in the <teiHeader>, here we discuss corpora bodies (element <body>).

Text chat

One of the interests of assembling heterogeneous corpora is to be able to step back from some forms of oversimplification. One such idea is that on the Internet there is one language, often called Netspeak (Crystal, 2001). Figure 7 shows two messages uttered in the same modality, text chat: (7a) is an extract from *cmr-simuligne* and (7b) from *cmr-getalp_org*. Whereas the author of (7b) types as if s/he were sending an SMS—writing some words such as 'vé' phonetically and not using the plural 's', for example in 'les equation'—the author of (7a), a learner of French, seeks to type full sentences. In the latter message, well-formedness is only endangered by a lack of knowledge in the target language or by the speed of typing which may cause typos e.g. 'hueres' in (7a) rather than 'heures'. (7a) is prototypical of CALL interactions where topics such as lexical or grammatical diversity can be studied in comparison to the target language spoken offline. Whether (7b) is prototypical of text chat or only reflects an idiosyncratic behavior is a research question in itself.

(7a)

```
<post xml:id="cmr-Simu-
Chat_Lugdunensis_Room1_S47_00528"
when-iso="2001-05-11T12:30:13"
who="#cmr-Simu-Ll8" type="chat-
message">

    <p>Le bateau est ammare a St
Helier dans un marina qui s'ouvre
seulement trois hueres avant la
maree</p></post>
```

(7b)

```
<post xml:id="cmr-get-c043-a21693"
when-iso="2004-03-18T14:09" who="#cmr-
get-c043-p39174" alias="cortex_taff"
type="chat-message">

  <p>Apres je vé faire ma physique c
aussi les equation bilan</p>

</post>
```

Figure 7: Linguistic diversity in text chat acts

SMS

Idiosyncratic ways of communicating within a specific modality have also been identified within our SMS corpora. Messages (8a) and (8b) were sent by the same author, who regularly introduces spaces into her/his message, whereas (8c) and (8d) come from another author following a serious conversation with her/his correspondent. As we will later see in Section 3, both the whitespaces in (8a) and (8b) and the abbreviations and agglutination-abbreviations in (8c) and (8d) will pose issues for the process of automatic annotation of the corpora.

```
(8a)  <post xml:id="cmr-slr-c001-a11644" when-iso="2008-06-16T11:59:00"
      who="#cmr-slr-c001-p868" type="sms">
      <p>à k e l a d r e s e n v o y e r d e s f l e u r s ?</p>
      </post>
      […]
(8b)  <post xml:id="cmr-slr-c001-a11647" when-iso="2008-06-16T12:00:39"
      who="#cmr-slr-c001-p868" type="sms">
      <p>e n f o n t d e n t i s t e</p>
      </post>
      […]
(8c)  <post xml:id="cmr-slr-c001-a00011" when-iso="2008-04-14T10:17:11"
      who="#cmr-slr-c001-p010" type="sms">
      <p>é@??$?Le + triste c ke tu na aucune phraz agréabl et ke tu va encor me
      dir ke c moi ki Merde par mon attitu2! Moi je deman2 pa mieu ke klke mot
      agréabl échangé</p>
      […]
(8d)  <post xml:id="cmr-slr-c001-a00304" when-iso="2008-04-15T20:23:59"
      who="#cmr-slr-c001-p010" type="sms">
      <p>.2 te comporter comme ca avec moi. Je ve bien admettr mes erreur kan
      j'agi vraimen mal comm hier mé fo pa exagérer. Si t pa d'accor c ton droi.
      Si tentain.le rest c à dirreposer dé question sur 1 sujet déjà expliké c
      pa 1 raison valabl pr ke tu te monte contr moi.pr moi ossi ca suffi.</p>
```

Figure 8: Different composition of graphemes and lexical items between two authors of SMS messages

Forum

As shown in Figure 9, the structure of a forum message is more complex. The example in this figure is taken from *cmr-simuligne*. The author of the message is a native speaker of French who is replying to a post made by a learner of French. Each person mentioned has been identified in the message structure (author, list of readers—here shortened) and in its contents (signature of the author). This information may lead to other types of research on discourse and group interactions. For example, who takes the position of a leader, or an animator within a group? Can subgroups of communication be traced within a group, thanks to an analysis of clusters or cliques (Reffay and Chanier ibid.)?

```
<post xml:id="cmr-Simu-Gall_e2a2_hymne-234" when="2001-06-06T08:17:00"
who="#cmr-Simu-Gn2" type="forum-message" ref="#cmr-Simu-Gall_e2a2_hymne-209">
  <head>
     <title>constitution des groupes</title>
     <listPerson>
        <person corresp="#cmr-Simu-Gt">
           <event type="Read" when="2001-06-06T08:17:00">
           <label>Read</label>  </event> </person>
           [….] </head>
  <p><name ref="#cmr-Simu-Gl4" type="person"><forename>Nick</forename></name>,
Est ce que c'est de l'humour anglais? Tu risques de le regretter Amicalement
<name ref="#cmr-Simu-Gn2" ty-
pe="person"><forename>Laurence</forename></name></p>
</post>
```

Figure 9: Message posted in a forum

Encoding manual annotations

Finally, Figure 10 illustrates another challenge faced by the CoMeRe editors when elaborating the TEI schema: the inclusion of manual annotations by researchers within the corpus.

In *cmr-smslareunion,* a large number of the SMS messages mix French from France (French-fra) with French pidgin from Réunion (pidgin-cpf). The content of the post before the <reg> element (which is a standard element belonging to the core TEI) corresponds to the actual message sent. The contents of the <reg> includes the researcher's manual annotations as s/he tries to identify, with various degrees of certainty (cf. @cert), whether part of the message is in French-fra or pidgin-cpf, and who, at the same time, transposes various segments into a more standard orthography.

```
<post xml:id="cmr-slr-c001-a2860" when="2008-05-01T09:49:36" who="#cmr-slr-
c001-p000424" type="sms">
   <p>Oui ver20h mc do st benoit vu ke mi mange la ba. tu mange avan de venir?
Tu me sone kan t la?</p>
   <reg type="transortho"><seg xml:lang="fra" cert="medium">Oui vers 20h Mac Do
Saint Benoît vu que </seg> <seg xml:lang="cpf">mi manj la ba.</seg> <seg
xml:lang="fra">Tu manges avant de venir ? tu me sonnes quand t'es
là ?</seg><add type="F"><seg xml:lang="cpf" cert="low"> Wi vèr 20h Mac Do Sin
Benoi  vu ke</seg> </add> <add type="trad"> <seg xml:lang="fra">je mange là-
bas</seg> </add> </reg>
</post>
```

Figure 10: Annotation of an SMS

The challenge here was to find out how the researcher's annotations, contained within a spreadsheet, could be kept and coded into TEI. The next challenge is to measure the extent to which these manual annotations will correspond to automatic annotations made during the next phase of our project.

3 Automatic corpora annotations

Drawing on previous NLP experience applied to various types of linguistic data issued from social media, the *Automatic Processing* WG is in charge of processing the first layer of annotations on TEI-compliant corpora. This project stage will begin in spring 2014. In this section, we present our motivations for applying an automatic annotation process to the CoMeRe corpora (Section 3.1) before turning to the decisions made concerning which annotations to make for which units (Section 3.2) and to a description of the processing pipeline for adding such annotations to the CoMeRe data (Section 3.2).

3.1 Motivations

If the usefulness of corpora has already been proven in numerous studies and applications, the real value of these corpora relies, most of the time, on the quantity and quality of the information that has been added to them. This information (as annotations) allows content characteristics that are useful (and often essential) for operational use to be highlighted. For example, knowing the grammatical nature of the "words" of a text chat or SMS corpora allows the syntactic structure of each element of the corpora to be identified, as well as the possibility to calculate the vocabulary used and analyze the syntactic or semantic context of a word or class of words, etc.

Depending on the nature of annotations, they can be added automatically, when possible, or manually with the help of appropriate interfaces. The often high cost of manual annotations represents a real handicap for their elaboration. Most of the time, only automatic anno-

tations are used, due to limited budgets that cannot allow for better, more descriptive manual ones. The CoMeRe corpora do not overcome this constraint. Provided by the project partners (corpus compilers), the corpora can contain annotations added by the compilers (as detailed in Section 2.4). One goal of the CoMeRe project is to automatically add additional annotations that will prove useful to improve the operational use of the CMC corpora.

Our starting point for this automated annotation processing is based on anonymized initial corpora that partners brought to the CoMeRe project. Anonymization of the corpora had previously been completed by the compilers. However, the anonymization rules were often different from one corpus to the next, and the CoMeRe team have therefore made them consistent across corpora.

The automated processing of annotations that was performed concerns the textual corpora (or part of them), regardless of the text's form (standard French, text chat, SMS, etc.). Its purpose is to split the interactions into minimal textual units and associate each of them with a label representing their membership to specific morphosyntactic classes as well as additional information, for example, the lemma associated with each unit. This processing is based on automated language processing procedures and techniques.

If the CoMeRe annotated corpora are to be used by any researcher for his/her own personal research questions (see Section 2.2), the set of morphosyntactic labels used (as well as the associated information) must be as "consensual" as possible. Ideally, they must be able to be projected/transformed into the specific model the researcher wants to use, without requiring extensive work and calculation. Even though such a configuration currently seems quite difficult to determine (does it even exist?), our goal is to get as close to this as possible, using a set of labels and "generic" associated information, which are readily understood, and can be used and transformed at a minor cost. We especially think that the association of a lemma to each minimal unit should allow for easier "customization" for researchers to conduct future studies on the contents of the CoMeRe repository.

This need for generic annotations led the *Automatic Processing* WG not to consider any processing beyond morphosyntactic labelling. Therefore, even though syntactic annotations (components, dependencies, etc.) could be considered, the diversity and specificity of existing syntactic analysis models undermines our concern for "genericity" and substantially handicaps any use/adaptation of such annotated corpora.

As part of the CoMeRe corpora consists of text with freely variant spelling (see examples in 2.4), the robustness of the processing tools used is an important factor for their choice. Indeed, they must allow us to automatically process (annotate) any element of these corpora, regardless of the level of variation: misspelling, agglutination (e.g., "*cp*" instead of *je ne sais pas* 'I do not know'), phonetic spelling (e.g. "*2m1*" instead of *demain* 'tomorrow'), shortened elements (e.g. "biz" instead of *bises* 'kisses'), etc. These occurrences are present in several, if not all, of the CoMeRe corpora. Furthermore, they often represent the majority of the interactions within a corpus, e.g. the *cmr-smsalpes* corpus. Tools with such robustness are currently quite rare for morphosyntactic processing of French texts; they are close to non-existent (in the form of complete and autonomous tools) for syntactic processing/annotation. This aspect is an important reason for not processing and annotating the CoMeRe corpora beyond a morphosyntactic level.

3.2 Which annotations for which units?

One of the first stages of processing consists of marking off the "processing units" (most of the time equivalent to a sentence) of the text, in order to apply the same processing to each of them. If splitting "normal" texts into these units does not pose any major problem (except for some specific cases), things are a bit different when it comes to CMC data. These corpora include interactions that only contain partial punctuation, if any. Moreover, it is usually based on punctuation elements that the splitting into units is done. Based on this observation, the processing hypotheses and the processing itself that we apply to each type of corpora are different. For corpora with punctuation that is often missing (SMS, text chat, tweets) our processing unit will be each post; no splitting will be performed. Each SMS, tweet or text chat message will be considered the final unit. For the rest of the corpora (email, forum messages, etc.), content will be split into processing units akin to a sentence and annotated accordingly. We are aware that the absence of clear unit delimitation marks can result in troubles with the processing of further elements of these corpora, for example syntactic analysis.

Apart from the definition of the processing unit, the type of processing/annotations that we apply to the corpora (morphosyntactic annotations) requires the definition of the typographic unit to which annotations can be associated. The targeted annotations being linguistic, they can only be obtained by relying on the linguistic notion of lexical unit (lexeme), which is, however, hard to automate due to the variety of possible ambiguities. For standard texts, these lexical units are often assimilated to units defined purely typographically, units that we will call *tokens.* These tokens are simply defined as a sequence of characters (excluding punctuation and spaces) preceded and followed by a space or a punctuation mark. The morphosyntactic taggers thereby consider the tokens as lexical units based on which language calculations can be performed to select the correct labels. The same goes for the lemmatizers.

This purely typographic approximation of the splitting into lexical units is very simple to obtain automatically. However, this process will not suffice for corpora that contain non-standard text. Indeed, putting aside the partial or complete absence of punctuation, other phenomena, for example abbreviations ("*bis*" or "*biz*" for *bises* 'kisses') or agglutination-abbreviations ("*chépa*" for *je ne sais pas* 'I do not know,' "*mdr*" for *mort de rire* 'LOL', "*ct*" or "*c t*" for *c'était* 'it was'), prevent any identification of lexical units and tokens, even in an approximative way. Following (partially or totally) the approach used in similar work (Fairon and Paumier 2006; Cook and Stevenson 2009; Chabert et al. 2012), the *Automatic Processing* WG decided upon the following: the tokens will receive the annotations but these annotations will provide as much information about the underlying lexical units as possible. As a consequence, "*chépa*" or "*ct*" will be considered as tokens, but will need to be annotated, through linguistic information describing the complexity of their correspondence, to the lexical units to which they are linked.

In order to obtain such annotations, some kind of mapping between tokens and (an approximation of) lexical units is required, as only the sequence of lexical units could be successfully tagged by existing POS taggers. This raises a new question: what kind of lexical units should we try and associate with observable tokens? Today, the answer to this question

results from the following fact: virtually all POS taggers are trained on edited corpora (often journalistic data). This means that for now, the easiest way to get an acceptable POS tagging and lemmatization accuracy on CMC data is to *temporarily* transform the data so that it appears as "edited" (as journalistic) as possible—in order for the POS tagger and the lemmatizer to be applied, and then to project the resulting information onto the original text.

3.3 Processing pipeline

The processing pipeline used in CoMeRe implements the ideas presented in Section 3.2. It has previously been applied to CMC data in two different ways: as a pre-annotation tool on French (Seddah et al. 2012a) and as a pre-parsing processing tool on English (Seddah et al. 2012b). It can be summarized in the following steps, which we criticize and illustrate below:[5]

- **Pre-processing** step: We first apply several regular-expression-based grammars taken from the SxPipe shallow analysis pipeline (Sagot and Boullier 2008) to detect smileys, URLs, e-mail addresses, Twitter hashtags, and similar entities, in order to consider them as one token even if they contain whitespaces.
- **Tokenization** step: The raw text is tokenized (i.e., split into typographic units) and segmented into processing units which play the role usually devoted to sentences (see above), using the tools included in SxPipe.
- **Normalization** step: We apply a set of 1,807 rewriting rules,[6] together with a few heuristics that rely on a list of highly frequent spelling variations (errors or on-purpose simplifications) and on the Le*fff* lexicon (Sagot 2010). The number of "corrected tokens" obtained by applying these rules might be different to the number of original tokens. In such cases, we use 1-to-*n* or *n*-to-1 mappings. For example, the rule *ni a pa* → *n'_y a pas '[there] isn't'* explicitly states that *ni* is an amalgam for *n'* and *y* (negative clitic and locative clitic, which will be POS tagged and lemmatized as two distinct lexical units), whereas *a* should be left unchanged in this context (the lexical unit matches the typographic unit), and finally *pas* is the correction of *pa* (negative adverb, approx. 'not').
- **Annotation** step: Lexical units are POS tagged and lemmatized using standard tools—in our case, the standard French model from the MElt tagger (Denis and Sagot 2012) and the associated lemmatizer. This POS tagging model was trained on the French TreeBank (FTB; Abeillé et al. 2003), "UC" version (FTB-UC), and on the Le*fff* lexicon (see Denis and Sagot 2012 for details).
- **Post-annotation** step: We apply a set of 15 generic and almost language-independent manually-crafted rewriting rules that aim to assign the correct POS to tokens that belong to categories not found in MElt's training corpus, i.e., in FTB; for example, all URLs and e-mail addresses are post-tagged as proper nouns whatever the tag provided by MElt; likewise, all smileys get the POS for interjections.

[5] During the whole process, XML annotations in the corpus are protected and ignored (but preserved).

[6] These rules were forged as follows: first, we extracted *n*-gram sequences involving unknown tokens or occurring at an unexpectedly high frequency from various development corpora (the development part of the *French Social Media Bank*, parts of the CoMeRe data); then we manually selected the relevant ones and provided them manually with a corresponding "correction".

- **Denormalization** step: We assign POS tags and lemmas to the original tokens based on the mappings between "normalized" lexical units and original token. If a unique lexical unit is associated with more than one original token, all tokens except the last one are assigned the tag *Y* and an empty lemma. The last token receives the tag of the lexical unit and its lemma. If more than one corrected tokens are mapped to one original token (non-standard contraction), the original token is assigned a tag obtained by concatenating the tags of all the lexical units, separated by the '+' sign. The same holds for lemmas. This convention is consistent with the existing P+D and P+PRO tags, which correspond to standard French contractions (e.g., *aux* 'to the(plur)', contraction of *à* 'to' and *les* 'the(plur)'). If the mapping is one-to-one, the POS tag provided by MElt for the lexical unit is assigned to the corresponding token.

We shall now illustrate this process by way of three examples: first, a single (contracted) token, then a simple non-standard compound and, finally, a whole sentence. Let us first consider the token *chépa* 'dunno'. Steps one and two (pre-processing, tokenization) have no particular effect on it. Step three normalizes this token by associating it with four lexical units, namely *je ne sais pas* 'I do not know.' Steps four and five POS tag and lemmatize these lexical units, thus producing, for example, the output *je*/CLS/*je ne*/ADV/*ne sais*/V/*savoir pas*/ADV/*pas*.[7] Then step six denormalizes this output by associating these POS tags and lemmas with the single input token, thus producing the following output: *chépa*/CLS+ADV+V+ADV/*je+ne+savoir+pas*.

Let us now consider the sequence *l'après midi*. It contains three tokens, *l'*, *après* and *midi*. The underlying lexical units are *l'* 'the' and *après-midi* 'afternoon'. In other words, the two last tokens are a non-standard compound. The result of step three is *l' après-midi*, thanks to an adequate normalization pattern, and step five produces *l'*/DET/*le après-midi*/NC/*après-midi*. Then step six applies the convention mentioned above for compounds while denormalizing: *l'*/DET/*le* is unchanged, the token *après* receives the special tag *Y* and no lemma, and the last token of the compound, *midi*, gets the tag of the corresponding lexical unit, NC, and the full lemma *après-midi*. Hence the final output: *l'*/DET/*le après*/Y/ *midi*/NC/*après-midi*.

Before moving on to the last example, it is important to be aware of the following three points concerning this approach. First, there is no clear-cut way of deciding what should be normalized and what should not. Second, normalization can sometimes be achieved in different ways. For example, *chépa* could be normalized as *je sais pas* (informal) or *je ne sais pas* (standard, formal, would be used in journalistic data). For these two points, the answer is the same: as the normalization is only temporary (just for the POS tagger and lemmatizer to work), the general guideline is to "normalize" everything that departs from standard (journalistic) French in such a way that it matches as closely as possible to standard (jour-

[7] This tagged and lemmatized example is given in the MElt format, an extension of the Brown Corpus format, in which the "word", its POS tag and its lemma are separated by slashes. A whitespace is a word-separator, and each sentence (i.e., each unit of treatment) is in one line. The tagset used here is the tagset used in the *French Social Media Bank*, which extends the so-called FTB-UC tagset (see Seddah et al. 2012a and references therein); CLS is the POS tag for subject clitics, V for finite non-imperative verbs and ADV for adverbs, including for negative adverbs such as *pas* and (maybe surprisingly) for the negative clitic *ne*.

nalistic) French. The third point worth mentioning is that the mapping between tokens and lexical units can be very strange. For example, let us consider the sequence *c t*. This sequence can be interpreted by actually pronouncing the name of both letters, which produces /setɛ/, the valid pronunciation of *c'était* 'it was,' which is composed of two lexical units, *c'* 'it' and *était* 'was'. Note that this mapping means that the token *c* corresponds to *c'é-* whereas the token *t* corresponds to *-tait*. There is therefore no direct correspondence between the original tokens and the underlying lexical units that are to be POS tagged and lemmatized. In such a situation, we consider that there is no other way but to consider both tokens as forming a *de facto* compound *c_t* that is itself the (nonstandard) contraction of *c'* and *était*. As a result, we tag and lemmatize it as *c*/Y/ *t*/CLS+V/*ce+être*.

Keeping this in mind, we can move on to our last example, a (simplified) sentence from the *French Social Media Bank*, found on a forum from the website Doctissimo (2013) that provides health-related information: *sa fé o moin 6 mois qe les preliminaires sont sauté c a dire qil yen a presk pa* 'Foreplay has disappeared for at least 6 months, that is there is almost none.' Table 1 illustrates the whole process by providing the output of steps three, five, and six together with the tokenized input (output of step two).

Within the CoMeRe project, this processing pipeline has already been tested and improved. For instance, the pre-annotation pipeline (used for developing the *French Social Media Bank*) used 327 instead of 1,804 normalization rewriting rules. There is still room for improvement however, and applying it systematically to the various CoMeRe corpora will certainly lead to further modifications and improvements. It is worth noting that CoMeRe will use this processing pipeline in a way that is similar to its use for developing the *French Social Media Bank*, i.e., as a pre-annotation tool. In other words, because the goal will be to have the best possible annotations on a well-defined set of data, we intend to improve our normalization rules and maybe MElt's (Denis and Sagot 2012) training data by exploiting the very data we want to annotate. This is very different from a standard setting where one expects to process new data as well as possible, thus preventing target data from being used in any way during the development or training of the tools.

The way the processing pipeline described above shall be used in CoMeRe is twofold:

- A fully automatic setting: the whole pipeline will be applied. The resulting annotations might be kept as such or might be manually corrected afterwards.
- A semi-automatic setting: for some corpora, such as *cmr-smsalpes* (Antoniadis 2013), manual normalization was performed, in a way that is approximately compatible with the objectives of step three. In such a setting, the manually normalized data is provided as an input to steps one and two, step three is skipped, steps four and five (tagging and lemmatization) are applied, and step six is replaced by an *a posteriori* alignment step, in order to dispatch the resulting annotations in the original data (before manual normalization). This alignment step has not yet been developed. However, we believe we can achieve it based among others on the set of normalization rewriting rules used by step three.

CoMeRe's automatic annotation process raises several issues, especially important on noisier corpora (SMS, text chat, etc.), which will be mentioned in the conclusion.

Table 1: Automatic correction and annotation (POS tags, lemmas) for a very noisy sentence extracted from the *French Social Media Bank* (Seddah et al. 2012a). Errors produced by the pipeline are mentioned.

Tokenized text (output of step two)	**Normalized text (output of step three)** ***whitespaces denote additional lexical unit boundaries***	**POS tagged and lemmatized normalized text (output of step five)**	**Final output (after step six)**
sa	*ça*	*ça*/PRO/*ça*	*sa*/PRO/*ça*
fé	*fait*	*fait*/V/*faire*	*fé*/V/*faire*
o	*au*	*au*/P+D/*à+le*	*o*/P+D/*à+le*
moins	*moins*	*moins*/ADV/*moins*	*moins*/ADV/*moins*
6	*6*	*6*/DET/*6*	*6*/DET/*6*
mois	*mois*	*mois*/NC/*mois*	*mois*/NC/*mois*
qe	*que*	*que*/PROREL/*que* (erroneous POS tag, should be CS)	*qe*/PROREL/*que*
les	*les*	*les*/DET/les	*les*/DET/les
preliminaires	*preliminaires* (the missing acute accent on the first *e* has not been restored)	*prelimi-naires*/NC/*preliminaire* (despite the missing acute accent, the POS tag is correct, but not the lemma)	*preliminai-res*/NC/*preliminaire*
sont	*sont*	*sont*/V/*être*	*sont*/V/*être*
sauté	*sautés*	*sautés*/VPP/*sauter*	*sauté*/VPP/*sauter*
c	*c'est-à-dire*	*c'est-à-dire*/CC/*c'est-à-dire*	*c*/Y/
a			*a*/Y/
dire			*dire*/CC/*c'est-à-dire*
qil	*qu' il*	*qu'*/CS/*que il*/CLS/*il*	*qil*/CS+CLS/*que+il*
yen	*y en*	*y*/CLO/*y en*/CLO/*en*	*yen*/CLO+CLO/*y+en*
a	*a*	*a*/V/*avoir*	*a*/V/*avoir*
presk	*presque*	*presque*/ADV/*presque*	*presk*/ADV/*presque*
pa	*pas*	*pas*/ADV/*pas*	*pa*/ADV/*pas*

4 Quality control and dissemination

All the data collected in the CoMeRe data bank (CoMeRe Repository 2014), as well as annotations added to the CMC corpora detailed in Section 3, are verified by the *Quality* WG before the public release of the corpora and their dissemination at the end of 2014. In this current section we detail these two processes. Firstly, in Section 4.1 we detail CoMeRe's staged process of quality control that allows a corpus to move from one project phase to the next. Secondly, in Section 4.2, we describe the planned dissemination of CoMeRe which is scheduled for the end of 2014. We also highlight questions this raised for members of the *TEI & Metadata* WG concerning the acknowledgement of individual researchers' work in both the metadata and corpus reference, as well as the need for appropriate licenses for our corpora.

4.1 Corpus quality control process

For the production of any corpus, quality control is an essential aspect, particularly when a corpus undergoes format conversions. As Reynaert et al. state, quality control should "take place all along the production timeline of the resource, rather than being put as a final check at the very end of corpus completion" (2010: 2697). Within the CoMeRe project, quality control is a multi-stage process that allows a corpus to move from one phase of the project to the next.

A first validation step occurs when the corpus compiler deposits the original corpus in the CoMeRe repository. The nomenclature for this version is *corpusname-v0.* At this stage, a member of the *Quality* workgroup checks that the information concerning the corpus license, the corpus size, the context in which data was collected, and descriptions of any previously performed anonymization processes has all been supplied, as well as the legibility of corpus files. Requests for additional information from the compiler are handled. Once these criteria have been met, the corpus moves on to the TEI conversion phase.

Once the corpus has been converted into TEI, it is deposited in the *corpusname-v0* server space under the nomenclature *corpusname-tei-v1.* The corpus then undergoes a second quality control process during which the metadata in the TEI header is firstly validated in relation to the information provided by the corpus compiler. At this stage, the corpus description in both English and French is checked alongside the bibliographic reference for the corpus and the encoding of different participant roles and the description of the corpus license. Secondly, the description of the anonymization process is then compared to the information supplied by the corpus compiler and the identification of the corpus' interaction participants is verified. In a third step, the *Quality* workgroup then proceed by randomly selecting a certain number of <post> elements with the original contents in *corpusname-v0* in order to check that no information has been lost in the TEI conversion process. After any back and forth exchange between the corpus compilers and those inputting the data, the corpus is then validated. The validated version moves into the *corpusname-v1* server space and the automatic annotations phase is set in motion.

Once automatic annotations have been completed, a final quality control occurs during which the version *corpusname-tei-v1* and the post annotation corpus version are compared to ensure that no information has been lost. It is verified that the person who performed the

annotations has been correctly cited in the metadata and that the annotation process has been included in the corpus description. Again, a selection of <post> elements are chosen and compared between the two versions in order to ensure that no interaction information has been lost. This validation is also directed towards the correctness of the annotations. Once this final quality control has validated the corpus, it moves into the *corpusname-tei-v2* server space and both the versions *corpusname-v1* and *corpusname-v2* are then deemed ready for dissemination and are deposited on the national server, ORTOLANG.

At the time of writing, the first stage is achieved where the four corpora previously mentioned are concerned. The *Quality* group has started its work in order to assess the version *corpusname-tei-v1,* before the automatic annotations, which are scheduled for the upcoming months.

4.2 The dissemination of CoMeRe

As mentioned, the CoMeRe corpora will be released at the end of 2014. Meanwhile, new corpora from the *New Acquisitions* WG are under process (see the next section for details) and will be integrated into the CoMeRe repository hosted by ORTOLANG.

ORTOLANG (2014) is a new national infrastructure network for which the objective is, firstly, to allow linguistic data in French (lexicons, corpora, dictionaries) and NLP tools to be disseminated amongst the international community of researchers in linguistics. Secondly, a selection of these data will be saved permanently by another national infrastructure (CINES) which has been mandated to save top-priority French research data in all scientific fields. This data storage is expensive: notably because files need to be converted into different formats regularly, as certain current formats may soon become obsolete.

The dissemination of CoMeRe corpora in open-access formats imposes some specific constraints because our corpora will join other corpora deposited in ORTOLANG that have been prepared within other national projects. All corpora deposited in ORTOLANG will be structured in TEI and made accessible through an interface that is still under development. The latter will allow users to perform linguistic queries using concordancers, lexicometric, and morphosyntactic tools, similar to the one found on the query interface of the German DWDS (2013) corpus. Variations in TEI formats within the range of corpora deposited in ORTOLANG are foreseen. This requires every project to document, in detail, the specific TEI structures used to format their corpora, particularly if any further conversions need to be made to facilitate incorporation of the corpora into the query interface. Releasing corpora in open-access formats also requires the provision of specific information for each corpus concerning the protection of author rights and that future users circumscribe to ethical reuse of the corpora.

Where the CoMeRe project is concerned, we have made some progress towards meeting ORTOLANG's requirements. Firstly, our IS model has been carefully documented in the header of every TEI file, as previously explained. Other metadata were added, detailing how data was collected as well as how ethics and rights were respected. Secondly, in order to encourage data reuse, following the philosophy of OpenData (2013), we have decided to release our corpora under Creative Common licenses or others that are closely related. This includes possibly accepting terms for commercial use (i.e., discarding the Creative Com-

mons' non-commercial option) and the creators waiving their intellectual property rights (CC0 license). We therefore had to ensure that all members' work was given scientific acknowledgement; both within corpus metadata and by way of a specific bibliographic reference attributed to the corpus.

The need to acknowledge the time spent by researchers in compiling and structuring corpora is a well-known, if not always respected, issue in corpus linguistics. In order to acknowledge the contributions made by different members of the CoMeRe project, the *TEI & Metadata* WG chose to use standard and precise terminology to encode participants' roles in each corpus. The OLAC format was adopted for this. This format is an overlayer of the Dublin Core, an ISO standard that is made up of 15 generic tags that, if need be, can be refined. Figure 11 is an extract of the *cmr-smslareunion* corpus' OLAC metadata card. It illustrates the encoding roles (Johnson 2006). These roles can also easily be encoded, as metadata, in the TEI header.

```
<dc:creator>LEDEGEN Gudrun</dc:creator>
<dc:creator>CHANIER Thierry </dc:creator>
<dc:contributor xsi:type="olac:role" olac:code="compiler">LEDEGEN Gudrun</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="editor">CHANIER Thierry</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="depositor">CHANIER Thierry</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="data_inputter">JIN Kun</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="data_inputter">HRIBA Linda</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="developer">LOTIN Paul</dc:contributor>
<dc:contributor xsi:type="olac:role" olac:code="sponsor"> [...]
```

Figure 11: Examples of OLAC encoding roles

While acknowledging different participants' contributions to a corpus is one issue, referring to a corpus as a global entity and to its creator is another. A specific way of referencing corpora must be adopted when citing and referencing the work; much the same way as bibliographic references are constructed and used within scientific publications. Bearing in mind the CODATA/ITSCI (2013) recommendations, CoMeRe decided to encode bibliographic reference to corpora as shown in Figure 12.

```
<dcterms:bibliographicCitation>Ledegen, G.(2014). Grand corpus de sms SMSLa
Réunion [corpus]. In Chanier T. (ed.) Banque de corpus CoMeRe.  Ortolang.fr :
Nancy. [cmr-smsalpes-tei-v1 ; http://handle.net/xxx/cmr-smslareunion-tei-v1]
</dcterms:bibliographicCitation>
```

Figure 12: Corpus citation

In the "Dublin Core – OLAC" metadata set, the bibliographic reference is integrated into the tag <BibliographicCitation>. The contents of this element will be displayed on the internet interface developed by ORTOLANG for corpus consultation and access. Following the CoMeRe example of how to form a bibliographic reference for a corpus, ORTOLANG has taken the decision to ask every corpus depositor to elicit this reference. This is a step in the right direction where standardized citation procedures are concerned.

Within a corpus citation, the permalink is an essential part of the reference[8]. In the same way that a Digital Object Identifier (DOI) allows a user to obtain direct access to the abstract

[8] Note that in Figure 12, the corresponding URL of the Handle type will be obtained when the corpus is deposited.

of a scientific publication, the permalink will be a permanent link to the corpus metadata. The latter allows users to search the ORTOLANG corpus access interface while being in compliance with harvesting protocols including the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). The advantage of this is that every corpus will be easily searchable on the Web. Moreover, each CoMeRe corpus will have an OLAC form (also converted inside the corpus' TEI header), allowing automatic harvesting by European servers since ORTOLANG is a representative of the Common Language Resources and Technology Infrastructure (CLARIN).

5 Conclusion and perspectives

The present article presented a general overview of the ongoing French CoMeRe project. Our ultimate goal is to build a kernel corpus of different CMC genres that is structured in TEI. At the time of writing, the CoMeRe repository comprises eight corpora (out of which four served as examples in this paper), representing different CMC genres: text chats (more than 3 million), SMS (44,000), emails (2,300), forum messages (2,700), and tweets (34,000).

Standardization is one of the key principles of the project and all CoMeRe corpora will be TEI-compliant. With this in mind, the CoMeRe project is involved in the European TEI-CMC SIG to design and write TEI guidelines for the markup of CMC data. The four corpora were marked up in TEI under a format that is now part of the draft proposal of the TEI-CMC SIG. As explained above, we found it more adequate to first design a more general framework, termed "Interaction Space", that would fit the richest and the more complex CMC genres and situations. In doing so, the developed model encompasses multimodality. This is particularly important as new data will soon be added to the repository, including, for example, MULCE corpora which comprise data coming from audio-graphic conferencing systems. Each CMC genre was then described through its interaction space and the TEI markup was determined regarding the IS.

Several of our TEI-compliant corpora are currently being tagged. The *Automatic Processing* WG has presented its motivations for applying an automatic annotation process to the CoMeRe corpora before turning to the decisions made concerning which annotations to make for which units and to a description of the processing pipeline for adding these to the CoMeRe data.

CoMeRe's automatic annotation process raises several issues, which are especially important where noisier corpora are concerned (SMS, text chat, etc.). Ongoing work[9] aims to better understand the phenomena that cause such data to depart from standard language corpora, in order to improve their automatic processing. As a first step, the *Automatic Processing* WG will focus on improving its tokenization and normalization scheme. This will require an explicit definition of the scope of the normalization process and a definition of the notion of *noisy token*.

The "genericity" of CoMeRe's POS and lemma annotation is a baseline that makes sense only if it can serve as input for various transformations, in order to be used in various types

[9] Among other, these issues are the main topic of a PhD funded by the Région Rhône-Alpes about the study and exploitation of SMS French.

of linguistic and NLP uses of the CoMeRe corpora. Further work is now required to study the balance between our annotations and the requirements of the various uses of CoMeRe corpora. This might lead the WG to develop tools for converting annotations from its generic (FTB-UC) tagset—widespread in the French NLP community—into various other tagsets, more adequate for downstream uses.

Finally, the ideas, methods, and tools described above have been designed and deployed on a few types of CMC corpora in two languages (French, English), including for the development of the *French Social Media Bank* (Seddah et al. 2012a), which will soon become part of CoMeRe.[10] Including new types of French CMC corpora within CoMeRe may require improvements and modifications of the approach and pipeline of the group, and even new strategies and tools. We are aware that a small set of gold standard annotations have to be produced and a formal evaluation of the tagging process be conducted. This may not be possible before the end of 2014, the concluding date of the first phase of the CoMeRe project.

Additional corpora are currently being processed by the *New Collections* WG. The **Twitter team** has developed a corpus of political tweets, *cmr-polititweets,* which reflects new political genres (Longhi 2013: 31) in the framework of a more general research project on lexicon. The corpus aims to gather the most influential French political statements. More than 34,000 tweets coming from 206 accounts have been collected and organized in our TEI format (Longhi et al. 2014). The **Wikipedia team** is focusing on controversial talk pages in the fields of science and technology. The corpus of talk pages, *cmr-wikiconflits*, will ultimately reflect different oppositions, such as controversial vs. consensual, people vs. objects. The team endeavors to examine four types of talk pages: (i) pages signaled on the Wikipedia mediation page; (ii) pages listed in the category *Neutral point of view: dispute*,[11] (iii) talk pages of articles having a pertinence controversy; and (iv) protected and semi-protected pages, i.e. pages subject to individual restrictions, thus temporarily or permanently limiting their editing. Data have already been collected and their transformation into our TEI format is in its final stages. Let us add that the Wikipedia team plan to conduct two types of analysis on the data and will concentrate both on the linguistic characteristics and the structure of the discussion pages.

These corpora, besides a selection of MULCE multimodal ones, will increase the representativeness and the variety of the CoMeRe repository, which will be released by the end of 2014. It will be the first milestone in the forthcoming *French National Reference Corpus* and we assume that the efforts we undertook will meet the strong demand for open and standard data within our community.

[10] The other use case is the 2012 SANCL shared task organized by Google on "non-canonical" English parsing, a task based on the English Google WebBank (see Seddah et al. 2012b and references therein).

[11] Signaling articles for which the neutral point of view is controversial, *i.e.* articles deemed to be non-neutral. This is one of the major subjects of dispute on Wikipedia.

References

Abeillé, A., Clément, L. & Toussenel, F. (2003). Building a Treebank for French. Kluwer: Dordrecht.

Antoniadis, G (2014). "Corpus de SMS réels dans les Alpes, smsalpes" [corpus]. In Chanier T. (ed.) Banque de corpus CoMeRe. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-smsalpes

Antoniadis, G., Chabert, G. & Zampa V. (2011). "Alpes4science: Constitution d'un corpus de SMS réels en France métropolitaine". TEXTOS conference: dimensions culturelles, linguistiques et pragmatiques. Annual conference of ACFAS, 9-10 May 2011, Sherbrooke, Canada.

Aston G. & Burnard L. (1998). The BNC handbook: exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.

Baldry, A. & Thibault, P-J. (2006). Multimodal Transcription and Text Analysis. Equinox: London.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2012). "A TEI Schema for the Representation of Computer-mediated Communication". In Journal of the Text Encoding Initiative (jTEI), 3, http://jtei.revues.org/476 ; DOI : 10.4000/jtei.476.

Beißwenger, M., Chanier, T. ,Chiari, I., Ermakova, M., van Gompel, M., Hendrickx, I, Herold, A., van den Heuvel, H., Lemnitzer, L. & Storrer, A. (2013). "Computer-Mediated Communication in TEI: What Lies Ahead". Special Topic Panel, TEI Conference and Members Meeting 2013, 2-5 October 2013, Rome, Italy.

Bellik Y. & Teil D. (1992). "Définitions terminologiques pour la communication multimodale". Conference Interaction Humain-Machine IHM'92, Paris. http://perso.limsi.fr/bellik/publications/1992_IHM_1.pdf

Burnard, L. & Bauman, S. (2013). TEI P5: Guidelines for electronic text encoding and interchange [document]. TEI consortium, tei-c.org. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Calico (2013). Online Tools for analysing and visualising discussion forums [website]. http://www.stef.ens-cachan.fr/calico/outils/outils.htm

Chabert, G., Zampa, V., Antoniadis, G. & Mallen, M. (2012). Des SMS Alpins. Éditions de la Bibliothèque départementale des Hautes-Alpes: Gap.

Chanier, T. & Vetter, A. (2006). "Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". Apprentissage des Langues et Systèmes d'Information et de Communication (ALSIC), 9. DOI: 10.4000/alsic.270, http://alsic.revues.org/270

CODATA/ITSCI Task Force on Data Citation (2013). "Out of cite, out of mind: The Current State of Practice, Policy and Technology for Data Citation". Data Science Journal 12, pp. 1-75, DOI: 10.2481/dsj.OSOM13-043

CoMeRe (2014). Communication Médiée par les Réseaux, project documentation [website]. http://comere.org

CoMeRe Repository (2014). Repository fo the CoMeRe corpora [website]. Ortolang.fr: Nancy, http://hdl.handle.net/11403/comere

Cook, P. & Stevenson, S. (2009). "An Unsupervised Model for Text Message Normalization". In Feldman, A. & Lönneker-Rodman, B. (Ed.). Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, pp. 71–78. http://aclweb.org/anthology/W/W09/W09-2000.pdf

Corpus-écrits (2014). Consortium Corpus-écrits [website]. http://corpusecrits. huma-num.fr

DARIAH (2014). Digital Research Infrastructure for Arts and Humanities [website]. http://www.dariah.eu/

Denis, P. & Sagot, B. (2012). "Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging". In Language Resources and Evaluation, 46(4), pp. 721–736.

Doctissimo (2013). Discussion forum linked to the website Doctissimo, general public welfare and health care [webservice]. Lagardère Active : doctissimo.fr. http://forum.doctissimo.fr/

DWDS (2013). Das Digitale Wörterbuch der deutschen Sprache [website]. http://www.dwds.de/

Fairon, C. & Paumier, S. (2006). "A translated corpus of 30,000 French SMS". In Proceedings of LREC 2006, 22-28 May 2006, Genova, Italy. http://www.lrec-conf.org/proceedings/lrec2006/

Falaise, A. (2014). "Corpus de français tchaté getalp_org" [corpus] . In Chanier T. (ed) Banque de corpus CoMeRe Banque de corpus CoMeRe. Ortolang.fr: Nancy. http://hdl.handle.net/11403/comere/cmr-getalp_org

Falaise, A. (2005). "Constitution d'un corpus de français tchaté". In Actes de RECITAL 2005, 6-10 June, Dourdan, France. http://hal.archives-ouvertes.fr/hal-00909667

Geyken, A. (2007). "The DWDS-Corpus: A reference corpus for the German language of the 20th century". In C. Fellbaum (Ed.). Collocations and idioms: linguistic, lexicographic, and computational aspects. London: Continuum Press.

Huma-Num (2014). French Infrastructure for Digital Humanities [website]. http://www.huma-num.fr/

IRCOM (2014). Consortium Corpus Oraux et Multimodaux [website]. http://ircom.huma-num.fr

Kupietz, M. & H. Keibel (2009): "The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research".Working Papers in Corpus-based Linguistics and Language Education, No. 3, pp. 53–59. Tokyo: Tokyo University of Foreign Studies (TUFS).

Lamy, M-N. & Hampel, R. (2007). Online Communication in Language Learning and Teaching. Basingstoke: Palgrave Macmillan.

LaRéunion4Science (2008). Site of the project sms4science located in La Réunion [website]. http://www.lareunion4science.org/

Ledegen, G. (2014). "Grand corpus de sms SMSLa Réunion" [corpus]. In Chanier T. (ed.) Banque de corpus CoMeRe. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-smslareunion

Ledegen, G. (2010). "Contact de langues à La Réunion: «On ne débouche pas des cadeaux. Ben i fé qoué al?». Langues et Cité, 'Langues en contact', vol.16, pp. 9-10. http://www.dglf.culture.gouv.fr/Langues_et_cite/LC16.pdf

Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014). "Polititweets, corpus de tweets provenant de comptes politiques influents" [corpus]. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr: Nancy. http://hdl.handle.net/11403/comere/cmr-polititweets

Longhi, J. (2013). "Essai de caractérisation du tweet politique". L'Information Grammaticale. vol.136, pp.25-32.

MULCE repository (2013). Repository of learning and teaching (LETEC) corpora [webservice] . Clermont Université : MULCE.org. http://repository.mulce.org

Oostdijk, N., Reynaert, M., Monachesi, P., Van Noord, G., Ordelman, R., Schuurman I., & Vandeghinste, V. (2008). "From D-Coi to SoNaR: A reference corpus for Dutch". In Proceedings of LREC, 28-30 May, Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/index.html

Johnson, H. (2006). OLAC Role Vocabulary [document]. Open Language Archive Community (OLAC). http://www.language-archives.org/REC/role.html

OpenData (2013) Principles for "openness" in relation to data and content [document]. Open Kowledge Foundation: http://okfn.org/. http://opendefinition.org/od/

ORTOLANG (2014). Open Resources and TOols for LANGuage [website]. ATILF / CNRS – Université de Lorraine : Nancy, http://www.ortolang.fr

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., & Verine B. (2013). "Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS". Épistémè— revue internationale de sciencessociales appliquées, 9: Des usages numériques aux pratiques scripturales électroniques, 107-138. http://hal.archives-ouvertes.fr/hal-00923618

Reffay, C. Chanier, T. Lamy, M.-N. & Betbeder, M.-L. (2009). (editors). LETEC corpus Simuligne [corpus]. MULCE.org: Clermont Université. [oai:mulce.org:mce.simu.all.all; http://repository.mulce.org]

Reffay, C. & Chanier, T. (2003). "How social network analysis can help to measure cohesion in collaborative distance-learning". In Proceedings of Computer Supported Collaborative Learning Conference (CSCL'2003). June 2003, Bergen, Norway. Kluwer Academic Publishers : Dordrecht, pp. 343-352. http://edutice.archives-ouvertes.fr/edutice-00000422

Rehm, G. et al. (2008). "Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems". In Proceedings of LREC, 28-30 May, Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/index.html

Reynaert, M., Oostdijk, N., De Clercq, O., van den Heuvel, H., & de Jong, F. (2010). "Balancing SoNaR: IPR versus Processing Issues in a 500-million-Word Written Dutch Reference Corpus". In, Seventh conference on International Language Resources and Evaluation, LREC '10, 19-21 May 2010, Malta. http://doc.utwente.nl/72111/1/LREC2010_549_Paper_SoNaR.pdf

Sagot, B. & Boullier, P. (2008). "SxPipe 2: architecture pour le traitement présyntaxique de corpus bruts". Traitement Automatique des Langues, 49(2), pp. 1-35.

Sagot, B. (2010). "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French". In Calzolari, N., et al. (Ed.). Proceedings of LREC'10, 17-23 May, Valetta, Malta. http://lrec-conf.org/proceedings/lrec2010/index.html

Seddah, D., Sagot, B., Candito, M., Mouilleron, V. & Combet, V. (2012a). "The French Social Media Bank: a Treebank of Noisy User Generated Content". In Kay, M. & Boitet, C. (Ed.). Proceedings of CoLing 2012: Technical Papers, 8-15 Decembre 2012, Mumbai, India, pp. 2441-2458. http://aclweb.org/anthology/C/C12

Seddah, D., Sagot, B. & Candito, M. (2012b). "The Alpage Architecture at the SANCL 2012 Shared Task: Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing". In Notes of the first workshop of Syntactic Analysis of Non Canonical Languages (SANCL'2012), in conjunction with NAACL'2012, 3-8 June 2012, Montreal, Canada.

Sharoff, S.(2006). "Methods and tools for development of the Russian Reference Corpus". In Wilson, A., Archer, D. & Rayson, P. (Ed.). Language and Computers, Corpus Linguistics Around the World. Rodopi: Amsterdam, pp.167-180. http://npu.edu.ua/!e-book/book/djvu/A/iif_kgpm_Corpus%20Linguistics.pdf

TEI-CMC (2013). TEI Special Interest Group on Computer-Mediated Communication [website]. http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

Wigham, C.R. & Chanier, T. (2013). “Interactions between text chat and audio modalities for L2 communication and feedback in the synthetic world Second Life”. Computer Assisted Language Learning, DOI: 10.1080/09588221.2013.851702.

Aivars Glaznieks, Egon W. Stemle

Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project

Abstract

This paper introduces the project DiDi in which we collect and analyze German data of computer-mediated communication (CMC) written by internet users from the Italian province of Bolzano – South Tyrol. The project focuses on quasi-public and private messages posted on Facebook, and analyses how L1 German speakers in South Tyrol use different varieties of German (e.g. South Tyrolean Dialect vs Standard German) and other languages (esp. Italian) to communicate on social network sites. A particular interest of the study is the writers' age. We assume that users of different age groups can be distinguished by their linguistic behavior. Our comprehension of *age* is based on two conceptions: a person's regular *numerical age* and her/his *digital age*, i.e. the number of years a person is actively involved in using new media. The paper describes the project as well as its diverse challenges and problems of data collection and corpus building. Finally, we will also discuss possible ways of how these challenges can be met.

1 Language in computer-mediated communication

There is a wealth of studies in the corpus linguistic literature on the particularities of language used in computer-mediated communication (CMC) (e.g. for German Bader 2002, Demuth and Schulz 2010, Dürscheid et al. 2010, Günthner and Schmidt 2002, Härvelid 2007, Kessler 2008, Kleinberger Günther and Spiegel 2006, Siebenhaar 2006, Siever 2005, Salomonsson 2011). Especially, the use of "netspeak" phenomena (Crystal 2001) such as emoticons, acronyms and abbreviations, interaction words, iteration of letters, etc. have attracted attention. The studies describe different functions of such phenomena within CMC. Features transferred from spoken language, such as discourse particles, vernacular and dialectal expressions are frequently mentioned characteristics of CMC. They serve to transmit informality of a given message, comment, or status post. Writers often use emoticons, interaction words (e.g. **grin**), abbreviations (e.g. *lol*), and spelling changes such as the iteration of letters (e.g. *cooooolll*) to compensate for the absence of facial expressions, gestures and other kinesic features, and prosody. Many emoticons, interaction words, and abbreviations are "verbal glosses" for performed actions and aspects of specific situations. In addition, there are also particularities in spelling that people use without the aim of representing features of spoken language and that deviate from the standard variety. To cover such phenomena (e.g. *n8* for 'night'), we will follow Androutsopoulos (2007; 2011) and use the term "graphostylistics". Finally, all forms of shortening (e.g. *lol*, *n8*, and *thx* for *thanks*) are often used for economic reasons to perform speedy conversations in chats and instant messages. The use of shortenings can also be motivated due to character restrictions of the used services.

Differences between the use of language in CMC and in traditional written genres were often described with respect to the model of Koch/Oesterreicher (1985; 2008). The model

differentiates written and spoken genres by separating the characteristic style of a text or a discourse from the medium (graphic vs phonic) in which it appears. Modelled on a continuum of proximity vs distance between the interlocutors, style is determined by the conditions of a specific communication (dialogue vs monologue, familiarity of interlocutors, presence in time and space, etc.) and the strategies of verbalization (permanence, density of information, complexity, etc.). Spoken language prototypically displays several characteristics with respect to morpho-syntax (anacoluthon, paratactic sequences, holophrastic utterances, etc.), lexis (e.g. low variation of lexical items), and pragmatics (discourse particles, self-corrections, etc.) that vary from prototypical written genres such as fictional and journalistic prose. In many cases the characteristic style corresponds with the medium, i.e. informal conversations between friends are carried out orally while an administrative regulation is published in written form. However, the relation between style and medium is not immutable, and the importance of distinguishing between the two becomes obvious when they are on opposite ends. A sermon, for example, features characteristics of written texts, and, in fact, it will be produced in written form. Nevertheless, it is usually orally presented. Thus, a sermon is transmitted orally (hence the medium is phonic), but, with respect to the conception of the text, it is based on a written tradition and reflects the characteristic style of written language. Changes in the relation between style and medium can also show the other way around. Particularly nowadays, with the rise of the new media, people can chat with each other using keyboards and touch-screens – a form of communication that was not possible some twenty years ago. New media thus facilitate proximity communication such as informal chatting using a graphic representation of language resembling habits of spoken language use.

However, CMC is characterized by more than a simple transfer of features of oral conversation into written form. Graphostylistic elements, for example, cannot be explained as adoptions from oral practices. They have originated from a writing system. Within the possibilities of written communication, people often use graphostylistic elements to signal affiliation to social groups. Therefore, occurrences of the emerging "new literacy" including all the before mentioned features of CMC must be considered as self-contained social practices that allow for the use of writing in realms of interactions that were formerly reserved to oral communication (Androutsopoulos 2007; 2011).

In opposition to public concerns, linguists found that new practices of literacy supported by the new media do not substitute traditional forms of writing but rather complement the individual repertoire. Dürscheid et al. (2010) for example compared texts of different genres produced by the same writers. They collected texts written at school and texts written for out-of-school activities. The corpus of extracurricular texts consisted of e-mails, instant messages, newsgroup communication, and postings on social network sites (SNS). The writers' age ranged from 14 to 24 years, and all participants attended schools in German-speaking Switzerland. Dürscheid et al. found that the style of writing varies in all kind of texts. Out-of-school texts share certain features such as the usage of special characters and images as well as graphostylistic elements, which the authors rarely found in texts written at school. Therefore, the authors claim that there is no direct influence of out-of-school style

on texts written in class, but pupils use their repertoire of different styles appropriately to the expectations of the reader and the conditions of text production.[1]

Storrer (2012) confirms the basic findings of Dürscheid et al. (2010). People are aware of the different occasions of writing, and they vary their style according to the motive of writing. In her study, she compares Wikipedia talk pages with article pages. She finds different styles of writing for talk and article pages indicating that authors of Wikipedia differentiate between the two types of pages. Talk pages are pages for interaction and discussion between Wikipedia authors whereas article pages aim for an explanatory dictionary entry. Hence, talk pages display features of spoken language and graphostylistic elements, but article pages do not. Assuming that the same people have produced both kinds of pages, it becomes obvious that people usually do not conflate *interaction-oriented* and the *text-oriented writing* (cf. Storrer 2012; 2013; 2014) but maintain a division between these two types of writing.

The present article introduces an ongoing project that investigates the use of German on SNS with a specific focus on the writers' age. Since interaction-oriented writing is a relatively recent but thanks to CMC ubiquitous phenomenon, we will investigate the question whether language use in CMC is similar across generations. The project concentrates on a selected regional user group, namely on users from the Italian province of South Tyrol. Linguistically, South Tyrol is characterized as a multilingual area where 70% of its inhabitants declare to be L1 German speakers, 26% L1 Italian speakers and 4% L1 Ladin speakers (Autonome Provinz Bozen 2012). With respect to German, two varieties are used. The standard variety of German in South Tyrol (STG) is used for text-oriented writing, e.g. for documents and newspaper texts, and as a spoken language in formal public contexts (e.g. at regional broadcast stations, in political speeches), in educational and academic settings at schools and at the university, and in conversations with people from other German speaking areas. South Tyrolean Dialect is used as a spoken language between German-speaking inhabitants of South Tyrol on almost all occasions. In recent years, the South Tyrolean Dialect has become more and more important for interaction-oriented writing in CMC.

We have structured the article in the following way: In Section 2, we will outline why "age" is a relevant category for linguistic analysis of CMC. Section 3 presents the research questions and the method of the study. In Section 4, we will describe challenges and problems of data collection and corpus building within the project and discuss possible solutions. The paper concludes with an outlook on our future work.

1 Extracurricular texts can be written in any variety, and were often entirely written in Swiss Dialect in Dürscheid et al.'s study, whereas curricular texts must be written in Swiss Standard German in German-speaking Switzerland. However, the authors of the study found the use of dialect in both extracurricular and curricular writing. The study shows that pupils sometimes resort to dialect words even in curricular texts. For this phenomenon, the authors assume an indirect interference from the new media. According to them, the increase of keyboard-based CMC supports the use of the dialect as an everyday written language. Thereby, dialect becomes an alternative for diglossic Swiss German writers, and thus dialect words may interfere with the standard variant even in situations in which the standard variety should be preferred.

2 Computer-mediated communication and the writers' age

Age has been considered in linguistics from various perspectives. Fiehler and Thimm (2003) list four different uses of *age* in social sciences, and there might be more: *age* can refer to (1) a numerical value, (2) a biological phenomenon with respect to maturation and aging, (3) a social phenomenon, and (4) a communicative construct.

(1) As a numerical value, *age* is easily countable and suitable for many methods of metadata collection. It can easily be measured by considering a person's date of birth. Yet, counting ages does not intrinsically refer to biological maturation or one's life experience, but people usually link the numerical age to a biological view of *age*. (2) The biological aspects of an individual are related to her/his numerical age, and that plays a major role in studies on language acquisition (for an overview see e.g. Tomasello and Bates 2001) and aging (Lindorfer 2012). (3) Among other social categories such as *gender*, *ethnicity*, and *class*, *age* becomes relevant when a person's status and behavior within the society gets affected (Mattheier 1987); for example, a numerically old person can be progressive with respect to political ideas, technological development, and other aspects of social life that are usually associated with young people. Therefore, such a numerically old person can be held to be "young" with respect to both his/her attitudes towards socially relevant topics and his/her behavior as a social actor. Linguistic behavior in general has been shown to vary with age, and, thus, linguistic features are often correlated with the membership to an age group. Therefore, the interrelation between linguistic behavior and *age* is a topic in many sociolinguistic studies (for an overview see e.g. Chambers 2003: 163-225). (4) Related to the social aspect, *age* is a relevant category in conversations and thus influences communication in general and the communicative behavior of the participants in particular. One example could be the attribution of "being old" or "being young" during conversation by oneself or by others (e.g. Coupland et al. 1991, Linke 2003).

On the other hand, there is no reason to assume an "age-specific language", and to consider people of a certain age (say older than 65) as one sociolinguistic group that uses language in a different way compared to people between 40 and 60 years, for example. Fiehler (2003: 39) lists features of age-specific language, but at the same time, he states that the group of elderly people is too heterogenic (cf. Digmayer and Jakobs 2013) and thus he cannot identify an age-specific variety or style. He suggests using prototypical features to describe the linguistic behavior of different age groups. Using the framework of prototype theory allows for features that do not necessarily occur in every "old" person's language, but may influence the perception and attribution of "oldness" to an interlocutor anyway ("doing age").

With respect to language use in CMC, little is known about communication of the elderly, neither between generations nor within an age group. Older people still do not use the internet to the same extent as young people do. For example, while more than 90% of adolescent people in Germany use the internet daily or several times a week (JIM-Studie 2012), only 25% of the elderly (more than 65 years) can be considered as internet users at all (Generali Altersstudie 2013), although more than 40% of the elderly live in houses with private internet access (infas 2011). In recent years, the number of users at the age of 60 and older has grown, and some researchers expect that this increase will continue in future years (Initiative

D21 2013). However, older people are reluctant to use new information and communication technologies that in turn generally use programs and tools online for data exchange; the older the people are, the less they use new media. The reasons for this *digital divide* are manifold. For many elderly even the possibilities of the internet seem to be difficult to discover since the use of the internet in general is often considered complex and complicated. In addition, the effort necessary to become familiar with the technical aspects of the new media (cf. Siever 2013) is too high compared to the amenities that are associated with the internet (cf. Schelling and Seifert 2010, Janßen and Thimm 2011). Mostly, people older than 70 years do not see any personal advantages of accessing the internet (infas 2011).

However, those of the elderly that use the internet frequently (so-called *silver surfers*) benefit from it as a source of information (Janßen and Thimm 2011: 386). In the realm of communication, writing e-mails is the dominant activity (infas 2011). Yet, only 3% of people in Germany over 65 years of age are members of a social network (Generali Altersstudie 2013). According to data from the US, however, SNS become more and more attractive to older people with the consequence that they will be used more frequently by older people in the future (cf. Janßen and Thimm 2011: 380). In recent years, some SNS emerged that have specialized in the demands of elderly people that serve as an SNS as well as an online dating service for the elderly (e.g. *www.platinnetz.de*). Compared to the use of e-mail and postings on SNS, silver surfers rarely use instant messaging services or take part in chat rooms. According to Janßen and Thimm (2011: 391), the synchronous nature of chat may frighten elderly people, since chatting requires rapid interaction and a good command of typing which may overstrain people who are not used to near-synchronous written communication.

In studies on CMC, there seems to be a consensus that age has to be considered as a variable (e.g. Androutsopoulos 2013). Sometimes, it remains disregarded due to the chosen methodology, for example in corpus linguistic approaches aiming at large and freely accessible data, for which no personal data is available (e.g. Beißwenger 2013). However, even if age is collected as a variable, the value of this information for the particular study may be questionable. Androutsopoulos criticizes that scholars in the field mostly focus on those variables that are easy to obtain no matter if they are relevant to the research interest: "The preference for clear-cut social variables such as gender and age may reflect scholarly convention rather than the categories that are relevant to participants in online communication" (Androutsopoulos 2011: 280). For research in the field of CMC, an age concept may be relevant that takes into account that parts of the population have been using new media devices such as personal and laptop computers only for the last 20 years; others started even more recently. Smartphones and tablet computers did not exist until 2007 and 2010, respectively. Though all kinds of web-enabled devices are widely spread (cf. Initiative D21 2013), the practice of using new media to communicate with friends and family members has neither affected the whole population, nor those people who are now using the new media every day but might not have done so when the services became first available. In addition, most people in western societies who were born after 1985 were more or less socialized with the new media. They are used to the digital world and they have practiced the handling of electronic devices from early on. Thus, they are often described as *digital natives*, and it is sometimes assumed that they differ from adults who were not socialized with the new media

(so-called *digital immigrants*, cf. Prensky 2001). These categories reflect the respective ease of handling new media tools. For this reason, it might be helpful to consider the peoples' experience with the internet in general and computer-mediated communication in particular, when studying CMC data. To refer to that experience, we use the term *digital age* which covers the span of time a person is actively using the facilities of the internet with the help of electronic devices, as well as the amount of time that a person is occupied with new media-related activities within a certain time span (per day, week or month). The numerical value of a person's age does by no means cover his/her experience with the digital world, and cannot be equated with the digital age. Since using the internet for communicative purposes is a social practice, and age as a numerical value may generate groups of people that are too heterogeneous with respect to this practice, this conception of age may not properly describe linguistic behavior in CMC. In this regard, the digital age must be conceived as a functional age because it is based on a person's competence and behavior rather than on his/her mere numerical age (cf. Kohrt and Kucharczik 2003: 32-33).

Summing up what has been said so far, communication in the new media demands a great deal of its users. The genre of CMC is characterized by interaction-oriented text production which displays several aspects of spoken language (inter alia the use of the dialect). In addition, genre-specific elements such as graphostylistic writing are widely used. Thus, new strategies for using language are observable in contemporary CMC. Despite being rendered in written form, the language of CMC is tailored to the requirements of interaction and thus displays features that characterize the proximity of the interlocutors such as embedding in the actual situation, little planning of speech acts, as well as emotional and dialogic use of language. Those people who were socialized with CMC (the *digital natives*) are used to interaction-oriented writing. Little is known about people who came into contact with CMC after their primary socialization phase at an advanced age or during adulthood (*digital immigrants*). So, do both groups have the same competences in producing CMC-specific interaction-oriented texts, and how do they apply interaction-oriented writing in CMC? For David Crystal, the rising number of older writers will have an impact on the width of writing styles that are observable in CMC: "[…] many emailers, for example, are now senior citizens – 'silver surfers', as they are sometimes called. The consequence is that the original colloquial and radical style of emails (with their deviant spelling, punctuation, and capitalization) has been supplemented by more conservative and formal styles, as older people introduce their norms derived from the standard language" (Crystal 2011: 11). Crystal's statement rests on a subjective evaluation rather than on verified analysis of data, and we, too, are not aware of any empirical studies that analyze older peoples' styles in CMC. In light of this current lack of research, the DiDi project aims to fill this gap. The following sections will introduce the project and discuss its challenges and potential solutions for investigating linguistic habits among users of CMC.

3 The DiDi project

In the *DiDi* project we analyze the linguistic strategies employed by users of SNS. The data analysis will focus on South Tyrolean users, and we will investigate how they communicate with each other. Another focus of the project is on the users' age and on the question

whether a person's age influences language use on SNS. As outlined above, we understand *age* in two ways: as a numerical value that reflects the life span of an individual and as *digital age* that reflects a person's experience with the new media.

We address three research questions:

1. How do South Tyrolean users of SNS (with L1 German) use German in
   a. quasi-public communication?
   b. private communication?
2. How do South Tyrolean users of SNS (with L1 German) use other languages in
   a. quasi-public communication?
   b. private communication?
3. Are there differences in language use that can be explained by *age*
   a. with respect to the numerical age (younger vs older users)?
   b. with respect to the digital age (experienced vs less experienced internet users)?

3.1 Design

Participants: The project aims to collect linguistic data from at least 120 different volunteers participating in the study. It is important that the participants have German as L1, and their center of life in South Tyrol. With respect to research question number 3, we envisage six age cohorts on the basis of their numerical age:

(1) between 15 and 24 years,
(2) between 25 and 34 years,
(3) between 35 and 44 years,
(4) between 45 and 54 years,
(5) between 55 and 64 years, and
(6) from 65 years on.

We will form age cohorts for the digital age on the basis of the relevant specifications made in the questionnaire (see below).

Period of recording: We will collect data within the time span of one year (2013).

3.2 Method

Data collection: We will collect data from the social networking platform Facebook. The corpus will consist of two kinds of data: (1) quasi-public communication on the participants' wall (i.e. status updates and comments)[2], and (2) non-public private messages[3] that

2 The privacy settings of Facebook distinguish between a public and a customizable non-public setting. In the default non-public setting, all communication on one's own wall is readable by one's Facebook friends. Since the number of friends can reach several hundreds of people, we do not consider conversations published in this setting as being private. By using the term quasi-public communication we refer to those communicative events that are potentially readable and joinable by one's friends, i.e. status updates and comments. We avoid the term public communication to refer to such communicative events because the non-public privacy setting guarantees a limited access to the walls and thus status updates and comments on these walls are not public. However, we do not yet distinguish between status updates that are published in a non-public setting from those published in a public setting. Since privacy settings are modifiable from one post to another, we will check for users that do so, and consider the privacy settings

we will harvest from the social network site Facebook. We will use past data, for both (1) and (2). A declaration of consent for the scientific utilization of one's own posts can be legally given for both past quasi-public posts and past private messages, and we aim at a minimum of 20 quasi-public posts and additional 20 private messages of each participant. Altogether, we expect at least 2,400 wall posts and 2,400 instant messaging conversations, 400 in each age cohort.

We use an online questionnaire for the collection of the participants' metadata which comprise personal data (sex, year of birth, center of life, L1), data regarding the use of the internet (year of first internet access, motive, activities), data regarding CMC (preferences of communication services, frequency of use, devices, languages), and socio-economic data (education, occupation). We will consider the metadata when we will interpret the result of the data analysis (see below).

Corpus creation: One aim of the project is to build a linguistic corpus of South Tyrolean CMC data. Therefore, we will enrich the data with additional relevant linguistic information such as information about lemma and part of speech (POS). We will use standard tools for the POS tagging annotation such as the TreeTagger (Schmid 1994), which embeds other automatic processing of linguistic data such as tokenization, sentence splitting and lemmatization. The features of the language used in CMC (Beißwenger et al. 2013) and of non-standard varieties (cf. Ruef and Ueberwasser 2013) will make manual corrections of the automatic annotations necessary. After that, the data will be prepared for a corpus query system (e.g. CQP, cf. Christ 1994), which is necessary for the quantitative and qualitative analysis of the data. Finally, we will make the corpus publicly available via the existing interface of the Korpus Südtirol initiative (cf. Anstein et al. 2011).

Data analysis: We will analyze the data quantitatively and qualitatively. Descriptive statistics including reports on the number of postings and messages as well as the length of words, sentences, and postings and messages constitute the main part of the quantitative analysis. Regarding the qualitative analysis, the focus will be on linguistic features of interaction-oriented writing in South Tyrol, including "netspeak" phenomena as well as speech characteristics due to proximity vs distance (dialect vs non-dialect vs standard variety).

Interpretation: All results of the data analysis will be interpreted considering the metadata coming from the online questionnaire. The calculation of correlations will reveal systematic interrelations between extra-linguistic factors and linguistic behavior. Calculations will focus on the variable *age* in its numerical as well as digital conception.

4 Challenges for corpus building

This Section points out some challenges for the corpus building process within the project DiDi. We first start with ethical and legal questions connected to our method of language data collection on SNS. After that, we present some challenges of finding participants of all

in our analyses. For the time being, we refer to all communicative events on the wall by using the term quasi-public communication, even though they may be published in the public setting. In contrast, we use the term private communication to refer to messages that are directly sent to a limited, preselected audience (e.g. instant messages).

3 From 2014 on, there is only one kind of messages on Facebook. The differentiation between instant messaging (chat) and sending messages (mail) was disestablished.

sighted age groups, especially those older than 65. Finally, we switch to the technical aspects of processing South Tyrolean Dialect. To estimate the performance of a customary tagger for Standard German on South Tyrolean Dialect, we performed a pretest in which we evaluated the performance of the TreeTagger on both original South Tyrolean Dialect and the same South Tyrolean Dialect data with adaptations to Standard German.

4.1 Ethical and legal aspects of the data collection in DiDi

Ethical and legal questions of the scientific use of language data collected from the internet have been raised in several publications (e.g. Beißwenger and Storrer 2008: 300-301, Crystal 2011: 14). The main question is who owns the text messages and who has the right to use it. Answers to this question may vary depending on which perspective you choose: (1) the legal or (2) the ethical.

(1) According to the legal terms of Facebook (Version 11, December 2012, cf. https://www.facebook.com/legal/terms), the legal status varies with the user's privacy setting. For all postings in the public setting, the legal status is well defined: "When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you (i.e., your name and profile picture)" (§2.4). On the contrary, all posted messages using the private setting are owned by the writer: "You own all of the content and information you post on Facebook, and you can control how it is shared through your privacy and application settings" (§2). Thus, messages sent in the public setting are legally free to use, while for messages posted in the Private setting the researcher needs a consent to use the data. Facebook even defines what a declaration of consent has to look like: "If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it." (§5.7)

(2) Beißwenger and Storrer (2008: 300-301) emphasize the importance of ethical considerations that have to be taken into account before creating a CMC corpus. Ethical considerations concern all personal information of the writer and other people mentioned in the texts. There is no doubt that scholars are obliged to handle any collected data responsibly. Therefore, they have to de-personalize all data that will be accessible in the corpus or published elsewhere. The de-personalization of the data concerns names and nicknames of people and places, and all contact information. Moreover, from the ethical point of view, it is questionable if a researcher interested in CMC data should collect messages without asking for permission, even though they are freely and legally available anyway. Storrer (2013) assumes that authors of such texts may not agree on the use of their texts in a searchable corpus regardless of the fact that the texts are available. Therefore, a consent for the use of the data would always be preferable. With respect to our method to collect past conversations instead of recording new data in a certain period, there may also be some ethical considerations. For example, before you record language data – even without collecting personal information – researchers are obliged to inform the interlocutors about the fact that they will be recorded and their utterances will be used for scientific purposes (*obligation to inform*). Interlocutors can then decide whether they want to take part in the recorded conversation.

When using past conversations, the researcher ignores his obligation to inform, which could be both an ethical and a legal issue.

As mentioned in Section 3, the project DiDi collects two types of data: quasi-public posts and private messages. For the collection of the language data, we will ask for an informed consent to use the language data and the personal data collected by the online questionnaire for a publicly available CMC corpus. We will use past wall posts and messages that are stored by Facebook for each user. Legally, the author of every post or message can provide the consent for the use of the data, even for past ones. We do not use status updates, comments on the participants' walls, and messages by users who do not participate in the project. Therefore, our procedure is in line with the ethical demands of using personal data from the internet.

4.2 Data collection

For the project to be successful, we need two kinds of data: first, language data in terms of written CMC data, and second personal data of the writers (metadata) (cf. section 3.2). All data should come from participants of different age groups (cf. section 3.1). The challenges for the recruitment of participants for the project are diverse. (1) Volunteers are not easy to find, especially those older than 65 years of age. Therefore, we need a recruitment procedure to address potential participants that considers habits of Facebook users of different numerical and digital ages within the new media as well as in the "real" world. The difficulty is how to attract attention for the study in members of all age groups. (2) Even if we cannot avoid any expenditure of time for the participants, we want to keep the effort for participating in the study as low as possible. Low expenditure of time may increase the probability for potential participants to take part in the study. While we can automate to a large extent the collection of language data, the metadata collection always involves the participants' cooperation and time. However, we need a procedure for the entire process of data collection that restricts the effort for each participant to an acceptable degree. (3) Since we collect two types of data, language data and metadata, we have to ensure that these data are related to each other. Therefore, there is also a technical challenge to integrate a mechanism into the data collection that guarantees that language data and metadata are connected in a secure and biunique way.

We face all three challenges by recruiting participants for our study using an app that we share with the South Tyrolean Facebook community. We will start with spreading an announcement for participating in the project. There will also be a request to share the announcement with other members of the South Tyrolean Facebook community. Everyone who is interested in participating has to install the app. Before people can start the app, they have to read the terms and conditions of participating, and provide a consent. After starting the app, we will be able to read out the language data (status updates, comments, and messages) of the last six month from the personal site of each participant. In addition, the app provides a secure and biunique link to an online questionnaire.

With respect to challenge (1), our procedure of recruiting participants follows the principles of how information is spread on the internet. People who want to support the study can both share the app and participate in the study by installing the app and fill out the question-

naire. Thereby, we will find active SNS users that guarantee the necessary amount of posts for each participant. At present (end of January 2014), we cannot estimate if we will find participants of all ages with this procedure.

A big advantage of this procedure is that it keeps the effort for each participant to a minimum (2). When they have decided to take part in the study, they just have to agree with the terms and conditions of the study at the beginning of the login process of the app. At the same time, the participants have to decide which language data they want to provide, their quasi-public posts and comments as well as their private messages or one of the two kinds only. Some more effort is necessary to fill out the questionnaire. We restricted the questionnaire to relevant questions with respect to our research interest. Therefore, participants will need a maximum of ten minutes to complete it. We are aware of the fact that not all Facebook users will use the inherent messaging service. Older users usually do not use instant messaging services (cf. Janßen and Thimm 2011: 391). Therefore, the collection of private messages is independent from the collection of the quasi-public posts, and participants can decide deliberately which kind of data they want to provide.

Finally, the app will ensure the connection of the language data with the metadata (3). For the metadata collection, we will resort to the online survey system *opinio* (http://www.objectplanet.com/opinio/). After starting the app, it will redirect each participant to the online survey and will use a secure and biunique identifier for each questionnaire.

4.3 Automatic processing of CMC data in South Tyrolean Dialect

Several studies on automatic processing of CMC data have shown that the so-called "noisy" language of CMC causes a remarkable decrease in the accuracy rate of the automatic processing (cf. the overview of studies in Eisenstein 2013: 359; see also Baldwin et al. 2013, Giesbrecht and Evert 2009). With respect to POS tagging, spelling variants that differ from the canonical spelling lead to problems for the processing. Spelling variants appear for instance in the form of expressive lengthening (*cooooll*) or abbreviations of words (*u* for *you*), and in all spellings that represent regional or social varieties (Gadde et al. 2011). Giesbrecht and Evert (2009: 32) use web data and report on well-known errors due to shortcomings of German taggers (cf. Schmid 1995 for the TreeTagger). Furthermore, they find "'new' error types due to the confusion of punctuation signs, foreign words and cardinals with common nouns, proper nouns and adjectives", especially in non-edited text genres such as *TV episode guides* or *postings from online forums*. In addition, CMC-specific writings (e.g. tokens starting with a hashtag) lower the accuracy rate of the tagger (cf. Baldwin et al. 2013: 359). For the successful processing of CMC data, there are mainly two solutions to this problem: (1) domain adaptation and (2) normalization (e.g. Eisenstein 2013). In (1), the tools will be adapted to the language data; in (2) on the other hand, the data will be adapted to the processing tools.

(1) CMC data features linguistic and structural particularities that are rather rarely found in traditional writings. Because it deviates more or less from the standardized variety, it presents challenges to the automatic processing tools that are trained on the standard variety found in newspaper texts. Traditional taggers cannot tag most of the specific features of CMC data adequately because the tag sets they use lack CMC-specific tags. Furthermore,

due to deviations from the standard orthography errors occur already during the tokenization process. In addition, deviations from newspaper texts generally cause a higher number of tagging errors on familiar tokens compared to newspaper texts. Some researchers hence are concerned with a specification of a CMC schema including an improved tag set for the POS tagging that covers typical CMC phenomena (e.g. Bartz et al. 2013, Beißwenger et al. 2013, Gimpel et al. 2011). Traditional tag sets, for example, such as the STTS (Schiller et al. 1999) do not provide special tags for CMC phenomena. When using the STTS, emoticons, for instance, must be additionally defined as one unit that should be tagged in a certain way, for example, either as a non-word (*XY*) or as an interjection (*ITJ*), depending on the theoretical considerations. Otherwise it takes each component of a given emoticon and uses one of the punctuation signs (*$.*, *$,*, *$(*). However, there are some suggestions for an extension of the existing STTS tag set that could be a solution to the problem. Beißwenger et al. (2012, 2013) for example introduce a new tag for *interaction signs* with six subcategories comprising interjections, responsives, emoticons, interaction words, interaction templates, and addressing terms. Without any doubt, an adjusted tag set will facilitate the automatic processing of CMC data. However, an adapted NLP tool chain for CMC is not yet available, and the adaption of tools for CMC data is still work in progress.[4]

(2) A recent example for normalization of CMC data is the *sms4science* project (cf. Dürscheid and Stark 2011). The project aimed at building a corpus of Swiss SMS messages including all Swiss national languages. Since many of the German messages were written in Swiss German Dialect, the researchers decided to translate word by word the dialect data into Standard German in order to automatically process the texts (Ueberwasser 2013). The generation of an interlinear gloss is extremely labor-intensive and time consuming, and realizable only with support of specifically designed computer programs (Ruef and Ueberwasser 2013).

To estimate the quality of South Tyrolean SNS data and to understand what kind of adaptation on South Tyrolean SNS data would be necessary, we decided to run a pretest. Recent collections of data coming from South Tyrol indicate that with respect to dialect use, South Tyrolean SMS data is similar and comparable to the Swiss data (e.g. Huber 2013). For other genres of CMC, no such data is available. The pretest should allow for an evaluation of POS tagging results on CMC data containing South Tyrolean Dialect. Another objective of the pretest was to determine if normalization of the data is inevitable to obtain acceptable tagging results, or if selective adaptations of the tools and specific corrections of the original data can be sufficient to achieve the same results for accuracy. A third possibility would be to use some adaptions of the tool as a means to "clean" the CMC data before continuing with further (manual) normalization tasks (cf. Baldwin et al. 2013). Our assumption was that some adaptations would have more effect on the POS tagging accuracy than others. That

4 For example, for further suggestions on a revised version of the STTS, so called STTS 2.0, cf. the STTS workshops 2012 and 2013 organized by CLARIN-D at IMS Stuttgart (http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/GermanTagsets.html).

would mean that domain adaptation is worthwhile for some but not for all aspects. If this assumption turns out to be valid, we could then establish a combined process using normalization and domain adaptation for the POS tagging of CMC data.

More precisely, our hypothesis was that the domain adaptation for words coming from closed classes (adpositions, pronouns, articles, conjunctions, modal and auxiliary verbs, particles) is more efficient with respect to the POS tagging accuracy of the entire corpus than an adaptation for open class words (nouns, main verbs, adjectives, adverbs). There are several reasons for our hypotheses: (1) the number of closed class words are by definition restricted whereas the number of open class words is much larger and increasing. That means that it is easier to provide a list of closed class words with their spelling variants for the domain adaptation process than one of open class words. (2) In addition, closed class words occur more frequently, so that many tokens per type are found in a corpus (cf. the type-token ratio (TTR) for closed class words of 0.0525 vs 0.2760 for open class words in our pretest corpus). Therefore, we assume that we can cover many tokens by enriching the inherent lexicon of the tagger with a relatively low number of closed class words. (3) The tagger can use this information provided by the closed class words to recognize nouns, adjectives, main verbs, and adverbs. This means that the initialization of closed class word may have a positive impact on the tagging accuracy of open class words. We will evaluate the efficiency of the procedure using the POS tagging accuracy rate.

We will use the findings of the pretest to determine how to process the data of the main study. An adaptation of the processing tool to the dialectal closed class words for example would be less labor intensive than to create and provide an entire lexicon for South Tyrolean Dialect. Furthermore, the normalization of the open class words would be less time-consuming than the normalization of the whole corpus, even though a gloss tool is able to support the normalization process and suggests candidates for the word-by-word translation (cf. Ruef and Ueberwasser 2013). The aim of the pretest therefore was to find out if a combined approach of domain adaptation and normalization is worthwhile.

For the pretest corpus, we collected 72 messages and 231 corresponding comments from the Facebook web page *Spotted: Südtirol*. The web page describes itself as a fan page that helps to find unknown people who a person saw somewhere recently in the real world and wants to know who they are. The person searching for another one has to describe the person and the occasion where she/he saw her/him. The description is published as a depersonalized post on the *Spotted: Südtirol* wall. The community has to comment on the post and help to disclose the identity of the person someone is looking for. We decided to use data from the *Spotted: Südtirol* page for the following reasons: First, it is an open-access page where all content is publicly available, and second, the page addresses a locally restricted community. Therefore, we were sure to use authentic language data coming from South Tyrolean users. From the original pretest corpus, we excluded seven comments that were entirely written in a language other than German (2 in English, 5 in Italian), and one comment that was not understandable because of non-transparent writings. All remaining messages and comments were written in German; however, they were not written in the standard variety but in South Tyrolean Dialect.

The pretest corpus consists now of 72 messages and 223 comments. First we tokenized, then checked the corpus for tokenization errors, and finally corrected them. In the original file, 348 tokens were corrected, mostly merged, resulting in 266 tokens. For the corrected version consisting of 5,138 tokens (3,506 tokens in 72 messages and 1,632 tokens in 223 comments), we created a gold standard for POS tagging to evaluate the TreeTagger performance. The gold standard was created in four phases: One annotator tagged ~5% of the corpus (250 tokens) from scratch. In addition, an ensemble of three taggers tagged the same part of the corpus. For the ensemble, we used the TreeTagger, the Berkeley Parser (Klein and Manning 2001), and the Stanford POS Tagger (Toutanova et al. 2003). Differences between the annotator and the ensemble were discussed by four members of the project team until a consensus was reached. We took the consensus as the gold standard for the 250 tokens (phase 1). After that, the ensemble tagged the remaining part of the corpus and all cases of deviance were tagged from scratch partly by the first annotator, partly by a second annotator, and partly by both the first and the second annotator (phase 2). As phase 1 confirmed some well-known deficiencies of POS taggers for German (cf. Schmid 1995, Giesbrecht and Evert 2009, Glaznieks et al. forthcoming), the annotators checked the annotations of the ensemble even if the taggers agreed on the same POS tag. Since both annotators tagged an overlapping part of the corpus, the project team compared the annotations and found a consensus whenever the annotation did not agree (phase 3). Systematic discordant cases between the two annotators were finally checked for the entire corpus to finalize the gold standard for the pretest corpus (phase 4).

To test our hypothesis, we compared the automatic tagging result of the TreeTagger on the original corpus with those on various normalized corpus versions. The versions differ in tokens that have been included into the normalization procedure. The baseline of the comparison is the original version (ORG) of the corpus corrected for errors with respect to automatic tokenization. In addition, in the first version of normalization (CLSD), we replaced all tokens coming from closed class words which differed from the standard German version with the corresponding standard German expression (1,321 replacements). In the second version of normalization (OPEN), we did the same for all tokens coming from open class words which differed from the standard German version (1,498 replacements). The normalization procedure for CLSD and OPEN included corrections of all kind of misspellings as well as a normalization of abbreviations. For the combined version (C&O), we normalized all tokens, coming from closed class and open class words as well as those words that have not been considered in CLSD and OPEN, i.e. interjections (ITJ) and non-words (XY) (altogether 2,857 replacements). Finally, in the last version of normalization (FULL), we also corrected punctuation errors (3,401 replacements). From CLSD, OPEN, and C&D we can estimate the success of a possible domain adaptation for closed class words, open class words, and the whole lexicon, respectively. Table 1 shows the accuracy rate for the baseline and the three versions of normalization compared to our gold standard. The results are split for messages and comments.

Table 1: Evaluation of the POS tagging results for different versions of normalization of the pretest corpus split for messages and comments.

	accuracy of POS tagging results				
	ORG	CLSD	OPEN	C&O	FULL
messages	42.98%	67.60%	63.35%	87.36%	90.34%
comments	37.93%	57.66%	52.82%	72.73%	76.60%
total corpus	41.38%	64.44%	60.00%	82.72%	85.95%

Table 1 shows that each normalization step leads to improvements of the POS tagging performance. McNemar's Chi-squared tests with Yates's continuity correction (R Development Core Team 2011) demonstrated that all pairwise comparisons of the corpus versions (ORG, CLSD, OPEN, C&O) are significant ($df=1$, $p < 0.001$). POS tagging results on the original data (ORG) are very low (41.4%). The normalizations performed for CLSD improved the accuracy rate in the total corpus to 64.4%; those performed for OPEN lead to a slightly lower rate of 60%. The accuracy rates reveal that our hypothesis stated before is correct: The normalization of closed class words led to a better accuracy rate on the whole corpus than the normalization of the open class words. The mere number of tokens that we have normalized in CLSD (1,321) is lower than for OPEN (1,498) but the accuracy rate for CLSD is significantly higher than for OPEN. However, the accuracy rates in both versions CLSD and OPEN are still low indicating that even if we adapt our tools to the language data, further manual normalization tasks would be necessary. The C&D column shows the accuracy rates that we reach when we normalize all tokens in the corpus. The tagging accuracy of about 80% in C&D is far from the one that can be obtained for newspaper texts (95-97%). We obtained the best result in FULL (~86%) in which we performed additional corrections of punctuation errors.

Note that the results for accuracy between messages and comments generally differ. The reason for the difference between the accuracy rate in messages and comments is most likely related to the fact that messages contain predominantly complete sentences whereas comments often consist of incomplete clauses and phrases.

As a further step, we were interested in the accuracy rate of the POS tagger on unknown words (UW) vs known words (KW). Table 2 shows the distribution of the accuracy rate for all versions of the pretest corpus for UWs and KWs. For all the corpus versions, the UWs decrease the accuracy rate of total corpus version. The analysis of the most frequent UWs revealed that most of them were emoticons and uncommon signaling of ellipsis (e.g. two dots instead of three) for which the TreeTagger did not have a proper tag and tagged these mostly as nouns (NN) or adjectives (ADJA, ADJD). Another group of UWs are proper names that were often tagged as common nouns (NN instead of NE).

Table 2: Evaluation of the POS tagging results for different versions of normalization of the pretest corpus split for unknown words (UW) and known words (KW)

	accuracy of POS tagging results				
total corpus	ORG	CLSD	OPEN	C&O	FULL
UW	11.60%	21.70%	10.48%	24.50%	27.97%
KW	71.16%	85.93%	80.75%	91.13%	93.02%
total	41.38%	64.44%	60.00%	82.72%	85.95%

The most frequent tagging errors in ORG (cf. Figure 1) are wrongly tagged adjectives and adverbs that were predicted to be either nouns or verbs, verbs that were predicted to be either nouns or adjectives, and nouns that were predicted to be verbs. In addition, many pronouns and articles were tagged as adjectives, and CMC-specific tokens, especially emoticons were often tagged as adjectives. Finally, a considerable number of wrongly tagged tokens occurred within the categories of nouns (NN instead of NE), adjectives (ADJA instead of ADJD), and verbs (VVPP instead of VVFIN). Tagging errors could be reduced in CLSD as well as in OPEN. In CLSD (cf. Figure 2), we provided the standard German version of all closed class words to the tagger. Consequently, most of the tagging errors within the closed class words could be avoided. However, some tagging errors of demonstrative and relative pronouns remained due to syntactical differences between South Tyrolean Dialect and Standard German. As we have assumed, providing closed class words had also an impact on wrongly tagged open class words but the number of tagging errors is still high. In OPEN in contrast, we could reduce tagging errors for nouns, adjectives, adverbs, and verbs, but at the same time, many closed class words were erroneously tagged (cf. Figure 3). The full normalization of the corpus in C&D led to a decline of tagging errors for all tokens compared with ORG but the error rate remains relatively high in general (cf. Figure 4). Errors persisted for many unknown nouns that were wrongly identified as adjectives, adverbs, or verbs, for CMC-specific tokens, and for well-known shortcomings of the German TreeTagger regarding for example homographic finite and infinite verb forms (cf. Schmid 1995: 7-8). The corrections of punctuation errors in FULL (cf. Figure 5) eliminated a couple of tagging errors in all POS tags and thus increased the accuracy rate to more than 85% in total. If we exclude all CMC-specific tokens such as emoticons and links to webpages from the corpus, the accuracy rate can be raised to almost 88.91% for the total corpus, and to 83.30% and 91.35% for comments and messages respectively.

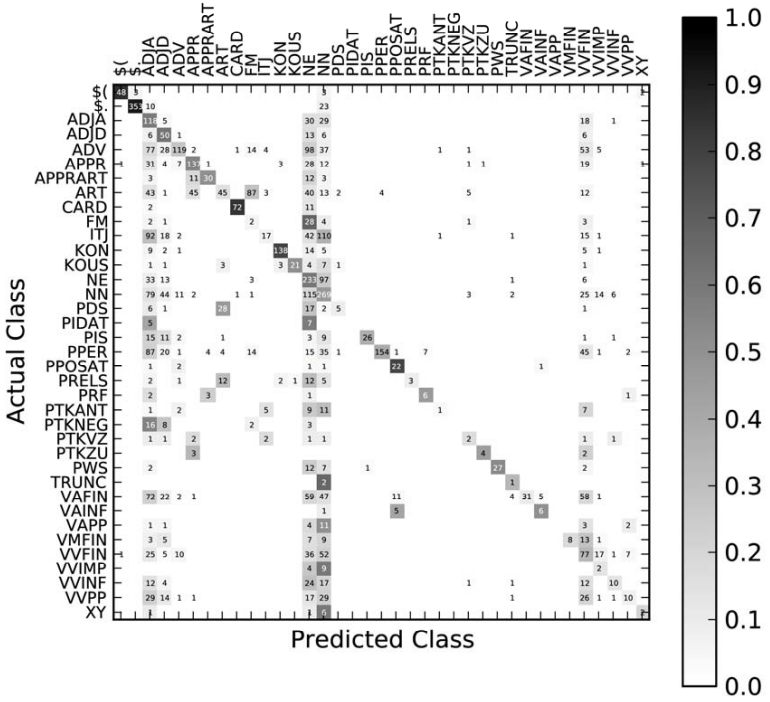

Figure 1: Confusion matrix for all wrongly tagged tokens (>5 instances) in ORG (see Schiller et al. (1999) for a glossary of all POS tags)

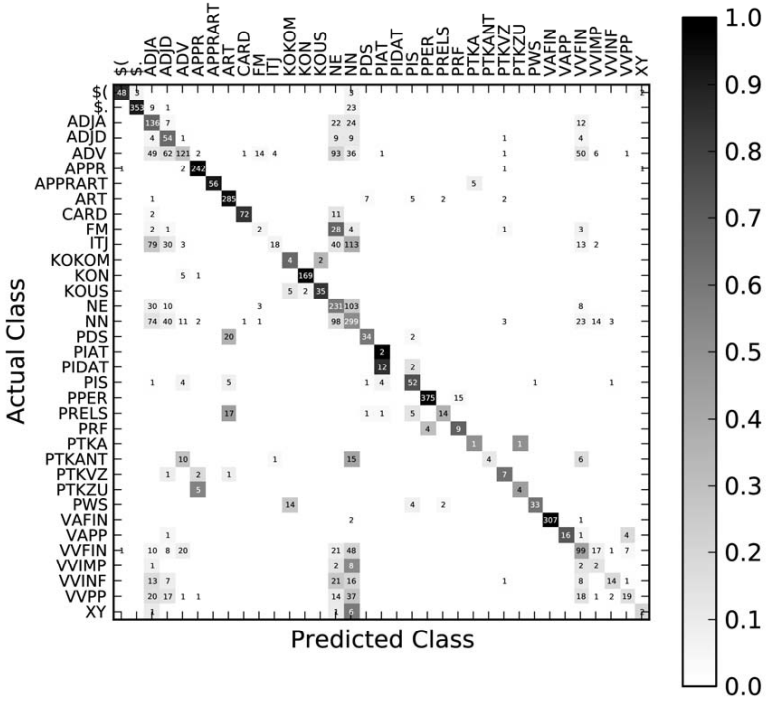

Figure 2: Confusion matrix for all wrongly tagged tokens (>5 instances) in CLSD

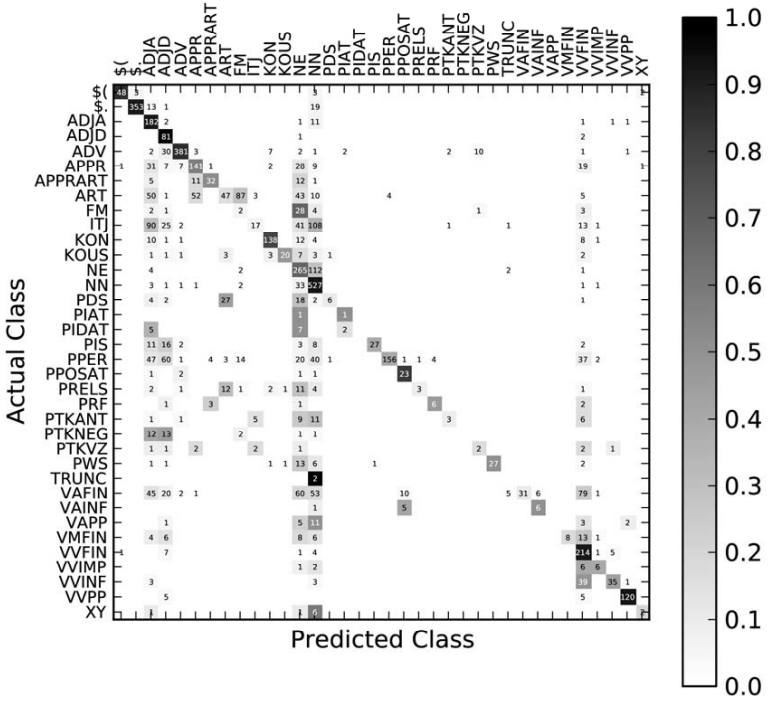

Figure 3: Confusion matrix for all wrongly tagged tokens (>5 instances) in OPEN

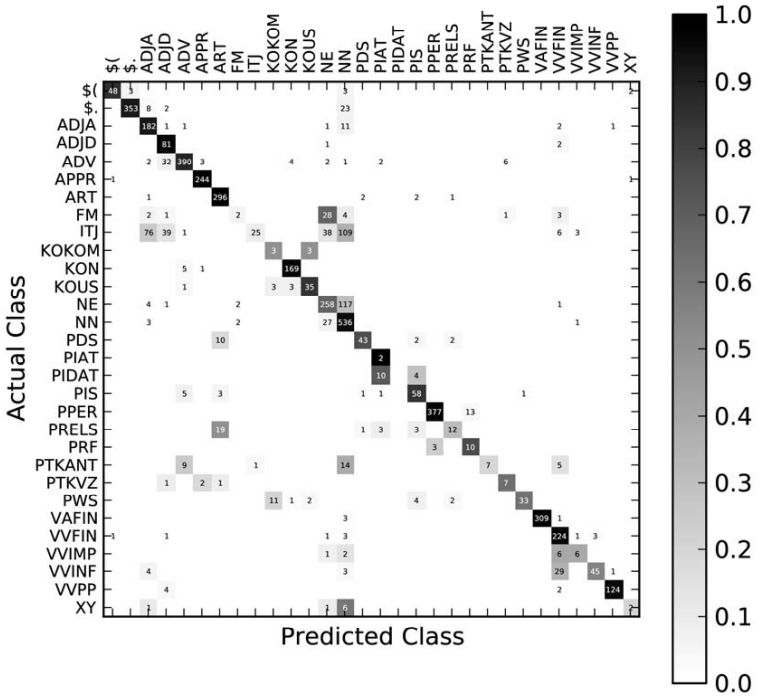

Figure 4: Confusion matrix for all wrongly tagged tokens (>5 instances) in C&O

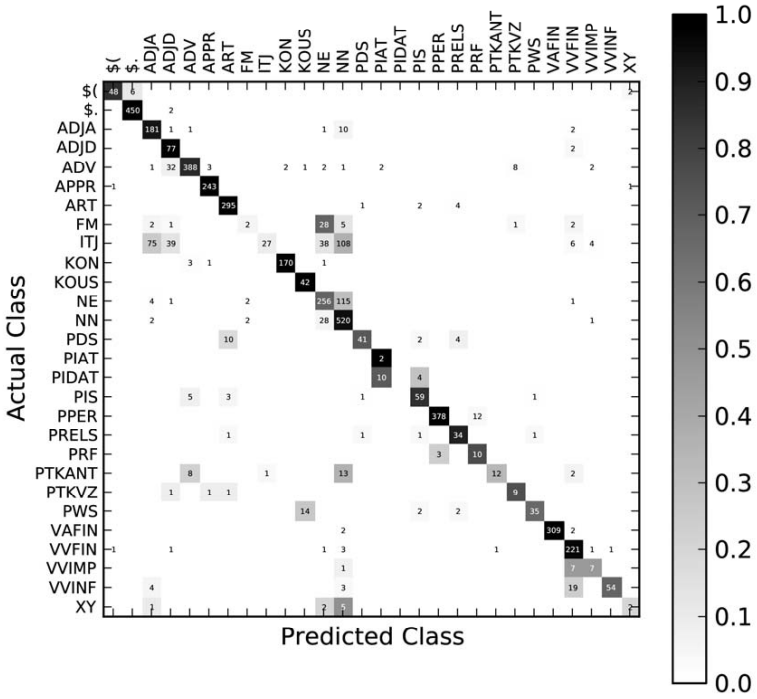

Figure 5: Confusion matrix for all wrongly tagged tokens (>5 instances) in ALL

5 Conclusion and future perspectives

In this article, we provided an overview of the project DiDi that is concerned with writing on social network sites and focuses on a regionally selected group of writers: South Tyrolean users. The main research question is whether people of different age (*numerical age*) and time of exposure to the internet (*digital age*) behave linguistically differently on social network sites. To answer this question, we will collect authentic data from the social networking platform Facebook and compile a CMC corpus.

There are several challenges regarding the corpus building process: We outlined general considerations about collecting personal data from the internet, and described a solution to ethical and legal problems of collecting data from Facebook as well as for recruiting participants that provide the language data. Finally, we considered well-known challenges in the automatic processing of CMC data and possible solutions. For our corpus, we expect a large

part to be written in South Tyrolean Dialect; this will pose problems for NLP tools usually devised for German standard language. A pretest on the POS tagging performance of South Tyrolean Dialect shows that corrections are necessary to obtain ample tokenization and POS tagging results. We tested the hypothesis whether these corrections can be facilitated for a large part of the corpus by furnishing the tagger with a lexicon for closed class words of the South Tyrolean Dialect. We assumed this intervention would also support the processing of open class words and thus diminish the need for further corrections. With our fully normalized pretest corpus we were able to estimate the impact on POS tagger performance for different word classes, i.e. if a POS tagger's lexicon were to be extended with certain entries, how would this extension improve its performance. In a first experiment, we substituted all dialect expressions coming from closed class words with standard German translations. This procedure improved the accuracy rate of the POS tagger from less than 50% to 64%. To obtain a reliable POS tagged corpus, further interventions appear necessary. The data also reveals that a completely normalized version of data (open- and closed words) coming from South Tyrolean Dialect still contains a high number of tagging errors that are only partly traceable to CMC-specific language (emoticons, ellipses, links, etc.). Many errors occur due to grammatical differences between Standard German and South Tyrolean Dialect and become obvious in the normalized version. An example is the use of the relative pronoun *was* to refer to a person, sometime even in the combination of two relative pronouns *der was*. This type of reference cannot be made in Standard German, neither alone nor in combination, and thus leads to tagging errors (e.g. ART instead of PRELS).

It seems that normalizations and corrections are inevitable to produce satisfactory tagging results, but to cover structural differences between South Tyrolean Dialect and Standard German, an improved language model would be necessary. With our pretest corpus being entirely written in South Tyrolean Dialect, with few (easier to handle) phenomena, such as hash-tags and URLs, we believe this corpus to be a worst-case scenario for what to expect. Although South Tyrolean users often use the dialect in CMC, we do not expect the corpus for the DiDi project to consist exclusively of dialect data but to also contain non-dialect data that can be processed by automatic tools. Therefore, the results of the pretest must be considered to represent a special case that may only partly affect the DiDi corpus depending on the distribution of the used German varieties.

Normalization and manual corrections are time consuming and labor-intensive, and therefore, we will have to balance the manual work with the expected outcome and the required quality of the processed data. Good criteria for the decision will be the percentage of South Tyrolean Dialect and Standard German in the main corpus, and the quality of the non-dialect data, i.e. how deviating from the Standard the CMC data in the DiDi corpus will be. We will also try to pool more CMC data and improve language models for the processing tools.

Bibliography

Androutsopoulos, J. (2007). “Neue Medien – neue Schriftlichkeit?” In: Mitteilungen des Deutschen Germanistenverbandes 54, 72-97.

Androutsopoulos, J. (2011). “Language change and digital media: a review of conceptions and evidence.” In: Kristiansen, T. and Coupland, N. (eds.) (2011): Standard languages and language standards in changing Europe. Oslo: Novus, 145-161.

Androutsopoulos, J. (2013). “Networked multilingualism: Some language practices on Facebook and their implications.” In: International Journal of Bilingualism, published online 11 June 2013. *http://ijb.sagepub.com/content/early/2013/06/07/1367006913489198* (accessed 13 June 2013)

Anstein, S., Oberhammer, M., Petrakis, S. (2011). “Korpus Südtirol - Aufbau und Abfrage.” In Abel, A. and Zanin, R. (eds.) (2011): Korpora in Lehre und Forschung. Bozen-Bolzano: University Press, 15-28.

Autonome Provinz Bozen (2012). “Volkszählung 2011/Censimento della populazione 2011.” In: astat info 38/2012.

Bader, J. (2002). “Schriftlichkeit und Mündlichkeit in der Chat-Kommunikation.” In: Networx Nr. 29 *http://www.mediensprache.net/networx/networx-29.pdf* (accessed 14 July 2014)

Baldwin, T., Cook P., Lui, M., MacKinlay, A. and Wang, L. (2013). “How Noisy Social Media Text, How Diffrnt Social Media Sources?” In: Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan, 14-18 October, 2013, 356-364.

Bartz, T., Beißwenger, M. and Storrer, A. (2013). “Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge.” In: JLCL 28, 157-198.

Beißwenger, M. (2013). “Das Dortmunder Chat-Korpus: ein annotiertes Korpus zur Sprachverwendung und sprachlichen Variation in der deutschsprachigen Chat-Kommunikation.“ In: LINSE, Linguistik-Server Essen. *http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf* (accessed 11 October 2013)

Beißwenger, M. and Storrer, A. (2008). “Corpora of Computer-Mediated Communication.” In: Lüdeling, A. and Kytö, M. (eds.) (2008). Corpus Linguistics. An International Handbook. Volume 1. Berlin. New York, 292-308.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. and Storrer, A. (2012). “A TEI Schema for the Representaion of Computer-mediated Communication.” In: Journal of the Text Encoding Initiative (3) November 2012.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. and Storrer, A. (2013). “*DeRiK:* A German reference corpus of computer-mediated communication.” In: Literary and Linguistic Computing 2013.

Chambers, J.K. (2003). Sociolinguistic theory. Linguistic vaiation and its social significance. Oxford: Blackwell.

Christ, O. (1994). “A Modular and Flexible Architecture for an Integrated Corpus Query System.” In: Proceedings of COMPLEX 1994, Budapest, 7-10 July, 1994, 23-32.

Coupland, N., Coupland, J. and Giles, H. (1991). Language, society and the elderly. Discourse, identity and aging. Oxford: Blackwell.

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

Crystal, D. (2011). Internet Linguistics. A Student Guide. London, New York: Routledge.

Demuth, G. & Schulz, E. K. (2010). "Wie wird auf Twitter kommuniziert?" In: Networx Nr. 56 *http://www.mediensprache.net/networx/networx-56.pdf* (accessed 25 October 2013)

Dürscheid, C. and Stark, E. (2011). "sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In: Thurlow, C. and Mroczek, K. (eds.) (2011). Digital Discourse: Language in the New Media. New York, London: Oxford University Press, 299–320.

Dürscheid, C., Wagner F. and Brommer, S. (2010). Wie Jugendliche schreiben. Schreibkompetenz und Neue Medien. Berlin: de Gruyter.

Digmeyer, C. and Jakobs, E.-M. (2013). "Innovationsplattformen für Ältere." In: Marx, K. and Schwarz-Friesel, M. (eds.) (2013). Sprache und Kommunikation im technischen Zeitalter. Wieviel Internet (v)erträgt unsere Gesellschaft? Berlin: de Gruyter, 143-165.

Eisenstein, J. (2013). "What to do about bad language on the internet." In: Proceeding of NAACL-HLT 2013, Atlanta, Georgia, 9–14 June, 2013, 359–369.

Fiehler, R. (2003). "Modelle zur Beschreibung und Erklärung altersspezifischer Sprache und Kommunikation." In: Fiehler, R. and Thimm, C. (eds.) (2013). Sprache und Kommunikation im Alter. Radolfszell: Verlag für Gesprächsforschung, 38-56. *http://www.verlag-gespraechsforschung.de/2004/alter/038-056.pdf* (accessed 4 September 2013)

Fiehler, R. and Thimm, C. (2003). "Das Alter als Gegenstand linguistischer Forschung – eine Einführung in the Thematik." In: Fiehler, R. and Thimm, C. (eds.) (2013). Sprache und Kommunikation im Alter. Radolfszell: Verlag für Gesprächsforschung, 7-16. *http://www.verlag-gespraechsforschung.de/2004/alter/007-016.pdf* (accessed 4 September 2013)

Gadde, P., Subramaniam, L.V. and Faruquie, T.A. (2011). "Adapting a WSJ trained Part-of-Speech tagger to Noisy Text: Preliminary Results." In: Proceedings of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data (J-MOCR-AND 2011), Beijing, China, September, 2011. *http://researchweb.iiit.ac.in/~phani.gadde/pubs/wsjTaggerSMS.pdf* (accessed 13 December 2013)

Generali Altersstudie (2013). Wie ältere Menschen leben, denken und sich engagieren. Frankfurt/Main: Fischer.

Giesbrecht, E. and Evert, S. (2009). Part-of-speech tagging - a solved task? An evaluation of POS taggers for the Web as corpus. In: Alegria, I., Leturia, I. and Sharoff, S. (eds.) (2009). Proceedings of the Fifth Web as Corpus Workshop (WAC5), San Sebastián, Spain, 7 September, 2009, 27-35.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith N.A. (2011). "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments." In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 19-24 June, 2011, 42–47.

Glaznieks, A., Nicolas, L., Stemle, E., Abel, A. and Lyding V. (forthcoming). "Establishing a Standardised Procedure for Building Learner Corpora." In: APPLES – Journal of Applied Language Studies. Special Issue: Proceedings of LLLC 2012, Oulu, Finland, 5-6 October, 2012.

Günthner, S. and Schmidt, G. (2002). "Stilistische Verfahren in der Welt der Chat-Groups." In: Keim, I. and Schütte, W. (eds.) (2002): Soziale Welten und kommunikative Stile. Festschrift für Werner Kallmeyer zum 60. Geburtstag. Tübingen: Narr, 315-337.

Härvelid, F. (2007). "'Wusste gar nicht, dass man schriftlich labern kann.' Die Sprache in Deutschschweizer Newsboards zwischen Mündlichkeit und Schriftlichkeit." In: Networx Nr. 51 *http://www.mediensprache.net/networx/networx-51.pdf* (accessed 25 October 2013)

Huber, J. (2013). Sprachliche Variation in der SMS-Kommunikation. Eine empirische Untersuchung von deutschsprachigen Schreiberinnen und Schreibern in Südtirol. Unpublished BA thesis at the Free University of Bozen - Bolzano, Faculty of Education.

infas (2011). infas-Telekommunikationsmonitor. Größte regionalisierte Studie zur Telekommunikation in Deutschland. Bonn: Institut für angewandte Sozialwissenschaften. *http://www.infas.de/fileadmin/images/themenfelder/kommunikation/infas_Telekommunikations-Monitor.pdf* (accessed 29 August 2013)

Initiative D21 (2013). D21-Digital-Index. Auf dem Weg in ein digitales Deutschland? TNS Infratest. *http://www.initiatived21.de/wp-content/uploads/2013/04/digitalindex.pdf* (accessed 29 August 2013)

Janßen, J. and Thimm, C. (2011). "Senioren im Social Web – entgrenztes Alter?" In: Anastasiadis, M. and Thimm, C. (eds.) (2001). Social Media. Theorie und Praxis digitaler Sozialität. Frankfurt/Main: Peter Lang, 375-395.

JIM-Studie (2012). Jugend, Information, (Multi-)Media. Basisstudie zum Medienumgang 12- bis 19-Jähriger in Deutschland. Stuttgart: Medienpädagogischer Forschungsverbund Südwest. *http://www.mpfs.de/fileadmin/JIM-pdf12/JIM2012_Endversion.pdf* (accessed 29 August 2013)

Kessler, F. (2008). "Instant Messaging. Eine neue interpersonale Kommunikationsform." In: Networx Nr. 52 *http://www.mediensprache.net/networx/networx-52.pdf* (accessed 25 October 2013)

Klein, D. and Manning, C. (2001). An *O* (n^3) Agenda-Based Chart Parser for Arbitrary Probabilistic Context-Free Grammars. Technical Report. Stanford. *http://ilpubs.stanford.edu:8090/491/1/2001-16.pdf* (accessed 29 November 2013).

Kleinberger Günther, U. and Spiegel, C. (2006). "Jugendliche schreiben im Internet: Grammatische und orthographische Phänomene in normungebundenen Kontexten." In: Dürscheid, C. and Neuland, E. (eds.) (2006): Perspektiven der Jugendsprachforschung. Frankfurt: Peter Lang, 101-116.

Koch, P. and Oesterreicher, W. (1985). "Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte." In: Romanistisches Jahrbuch 36, 15-43.

Koch, P. and Oesterreicher, W. (2008). "Mündlichkeit und Schriftlichkeit von Texten." In: Janich, N. (ed.) (2008). Textlinguistik. 15 Einführungen. Tübingen: Narr, 199-215.

Kohrt, M. and Kucharczik, K. (2003). "'Sprache' – unter besonderer Berücksichtigung von 'Jugend' und 'Alter'." In: Fiehler, R. and Thimm, C. (eds.) (2013). Sprache und Kommunikation im Alter. Radolfszell: Verlag für Gesprächsforschung, 17-37. *http://www.verlag-gespraechsforschung.de/2004/alter/007-016.pdf* (accessed 4 September 2013)

Lindorfer, B. (2012). "Psycholinguistische Erkenntnisse zur Sprache im Alter." In: Neuland, E. (ed.) (2013). Sprache der Generationen. Mannheim/Zürich: Dudenverlag, 78-97.

Linke, A. (2003). "Senioren. Zur Konstruktion von (Alters-?)Gruppen im Medium Sprache." In: Häcki Buhofer, A. (ed.) (2003). Spracherwerb und Lebensalter. Tübingen/Basel: Francke, 21-36.

Mattheier, K. J. (1987). "Alter, Generation." In: Ammon, U., Dittmar, N. and Mattheier, K. J. (eds.) (1987). Sociolinguistics. An international handbook of the science of language and society. Berlin: Walter de Gruyter, 78-82.

Prensky, M. (2001). “Digital natives, digital immigrants.” In: On the horizon 9 (5). *http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf* (accessed 11 October 2013)

R Development Core Team (2011). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Ruef, B. and Ueberwasser, S. (2013). “The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages.”. In: Zampieri, M. and Diwersy, S. (eds.) (2013). Non-standard Data Sources in Corpus-based Research. Aachen: Shaker, 61-68.

Salomonsson, J. (2011). “Hamwa nisch … fragense mal da. Spiel mit Mündlichkeit und Schriftlichkeit in Diskussionsforen im Internet.” In: Networx Nr. 59 *http://www.mediensprache.net/networx/networx-59.pdf* (accessed 25 October 2013)

Schelling, H. R. and Seifert, A. (2010). Internetnutzung im Alter. Gründe der (Nicht-)Nutzung von Informations- und Kommunikationstechnologie (IKT) durch Menschen ab 65 Jahren in der Schweiz. In: Zürcher Schriften zur Gerontologie 7. University of Zurich: Zentrum für Gerontologie.

Schiller, A., Teufel, S. and Stöckert C. (1999): “Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset).” Universität Stuttgart: Institut für maschinelle Sprachverarbeitung. *http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf* (accessed 11 December 2013)

Schmid, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees.” In Proceedings of the International Conference on New Methods in Language Processing, Manchester, 14-16 September, 1994. *http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf* (accessed 14 October 2013)

Schmid, H. (1995). “Improvements In Part-of-Speech Tagging With an Application To German.” In: Proceedings of the ACL SIGDAT-Workshop. Dublin, 1995. *http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf* (accessed 14 October 2013)

Siebenhaar, B. (2006). “Gibt es eine jugendspezifische Varietätenwahl in Schweizer Chaträumen?” In: Dürscheid, C. and Neuland, E. (eds.) (2006). Perspektiven der Jugendsprachforschung. Frankfurt: Peter Lang, 227-239.

Siever, T (2005). “Von MfG bis cu l8er. Sprachliche und kommunikative Aspekte von Chat, E-Mail und SMS.” In: Der Sprachdienst 49, 137-147.

Siever, T. (2013). “Zugänglichkeitsaspekte zur Kommunikation im technischen Zeitalter.” In: Marx, K. and Schwarz-Friesel, M. (eds.) (2013). Sprache und Kommunikation im technischen Zeitalter. Wieviel Internet (v)erträgt unsere Gesellschaft? Berlin: de Gruyter, 7-25.

Storrer, A. (2012). “Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia.” In: Köster, J. and Feilke, H. (eds.):.Textkompetenzen für die Sekundarstufe II. Freiburg: Fillibach, 277-304.

Storrer, A. (2013). “Sprachstil und Sprachvariation in sozialen Netzwerken.” In: Frank-Job, B., Mehler, A. and Sutter, T. (eds.) (2013). Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften, 331-366.

Storrer, A. (2014). “Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde.” In: Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.

Tomasello, M. and Bates, E. (eds.) (2001). Language development. The essential readings. Oxford: Blackwell.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. (2003). "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." In: Proceedings of HLT-NAACL 2003, 252-259. *http://nlp.stanford.edu/downloads/tagger.shtml* (accessed 29 November 2013)

Ueberwasser, S. (2013). Non-standard data in Swiss text messages with a special focus on dialectal forms. In: Zampieri, M. and Diwersy, S. (eds.) (2013). Non-standard Data Sources in Corpus-based Research. Aachen: Shaker, 7-24.

Eliza Margaretha, Harald Lüngen

Building Linguistic Corpora from Wikipedia Articles and Discussions

Abstract
Wikipedia is a valuable resource, useful as a lingustic corpus or a dataset for many kinds of research. We built corpora from Wikipedia articles and talk pages in the I5 format, a TEI customisation used in the German Reference Corpus (Deutsches Referenzkorpus - DeReKo). Our approach is a two-stage conversion combining parsing using the Sweble parser, and transformation using XSLT stylesheets. The conversion approach is able to successfully generate rich and valid corpora regardless of languages. We also introduce a method to segment user contributions in talk pages into postings.

1 Introduction

Wikipedia is a large, multilingual and rich online encyclopedia covering a wide range of domains including medicine, sport and history in millions of articles and talk pages (discussions). As a language resource, Wikipedia is useful in multilingual natural language processing, knowledge extraction, linguistics studies, and other disciplines. Since the content of Wikipedia has not been written by a single author, but collaboratively by many users, it is also of interest in computer-mediated communication (CMC) studies.

While Wikipedia has many benefits, its content or wikitext, is rather difficult to access due to its complex structure. The structure is represented by a mixture of the *wiki markup* language[1] and HTML tags. Although there is an effort to create a wikitext standard[2] and a Wikipedia DTD[3], so far it has not been standardized. The HTML generated by the wiki software *MediaWiki*[4] (Barett, 2009) also contains structural errors (Schenkel et al., 2007; Dohrn and Riehle, 2011).

XML-based Wikipedia corpora are advantageous because Wikipedia content can be accessed by using the query language XPATH[5]. Moreover, they can easily be reused, because it does not require much effort to adapt them for other projects. Another advantage is that XML[6] can be converted into other standard formats such as (variants of) the TEI. An XML-based Wikipedia has been proved to be useful in various tasks such as semantic annotation (Atserias et al., 2008), information retrieval, and machine learning (Denoyer and Gallinari, 2006). In fact, only a few Wikipedia XML corpora are available, and commonly they only contain a small portion of Wikipedia.

[1] http://en.wikipedia.org/wiki/Help:Wiki_markup
[2] http://www.mediawiki.org/wiki/Wikitext_standard
[3] http://www.mediawiki.org/wiki/Wikipedia_DTD
[4] http://www.mediawiki.org/wiki/MediaWiki
[5] XML Path language, see http://www.w3.org/TR/xpath/
[6] Extensible Markup Language, see http://www.w3.org/TR/xml/

The aim of the work described in this paper was to implement a conversion of all German Wikipedia articles and talk pages into TEI-based corpora for integration in the *German Reference Corpus* (Deutsches Referenzkorpus - DeReKo). The conversion system should also be reusable over time and over different language versions of Wikipedia.

DeReKo is hosted by the Institut für Deutsche Sprache (IDS) in Mannheim and serves linguists as an empirical basis for research on contemporary written German. Currently it comprises more than 24 billion word tokens, distributed over many subcorpora with texts from genres as diverse as newspaper text, fiction, parliamentary debates, and specialised text (Kupietz and Lüngen, 2014). DeReKo texts are marked up for metadata and text structure according to the XML application *I5*, which is a TEI customization (Sperberg-McQueen and Lüngen, 2012) based on the Corpus Encoding Standard XCES (see Ide et al., 2000, and Section 4.1). I5 is also the internal format for data storage in the linguistic research system COSMAS II at the IDS.[7] Hence, I5 is the target format of the Wikipedia conversion described in this paper.

Our conversion approach is based on Bubenhofer et al. (2011)'s approach, who did not convert the Wikipedia wikitext directly into the IDS-XCES (the predecessor of I5) format, but divided the process into two stages: first, convert wikitext to an intermediate XML representation and second, convert the intermediate XML to IDS-XCES. The motivation behind the division was to create an intermediate corpus representing nearly all wiki markup elements used in the wikitext in XML. The WikiXML corpus is then filtered and transformed to the more constrained IDS-XCES format using XSLT[8]. Since XSLT is ideal for transforming XML into XML, it is naturally used for the second stage.

Unlike Bubenhofer et al. (2011) who used XSLT also for the first stage, we took a parsing approach using the recent Sweble parser (Dohrn and Riehle, 2011) implemented in Java. We argue that XSLT is not appropriate for the wikitext to WikiXML transformation, because the declarative nature of XSLT is not suitable for the complexity of wiki markup, which is not proper XML in the first place. On the other hand, Sweble can handle the complexity of wiki markup and generates a Java object model from wikitext. We also implemented an XML renderer to represent this object model in XML.

Beside providing the Wikipedia corpora in WikiXML and I5, our major contribution described in this paper includes the implementation of a system to convert Wikipedia articles and talk pages into a rich XML representation. For the talk pages, we introduce a posting segmentation method using delimiters and regular expressions. We also improved Bubenhofer et al. (2011)'s XSLT Stylesheets for converting WikiXML into IDS-XCES/I5.

The paper is structured as follows: In Section 2, wiki markup and the nature and structure of Wikipedia articles and talk pages are introduced. In Section 3, we present the state of the art of the development of Wikipedia corpora. In Section 4, we explain our approach to building Wikipedia corpora. We also describe the I5 target format and the posting segmentation method for Wikipedia talk pages. In Section 5, we discuss the

[7] `http://www.ids-mannheim.de/cosmas2/`

[8] Extensible Stylesheet Language, see `http://www.w3.org/TR/xslt`

Table 1: Wiki markup examples

Wiki markup	Function
`== heading level 2 ==`	heading
`----`	horizontal rule
`:indentation level 1`	indentation
`* item`	unordered list
`# item`	ordered list
`: definition 1`	definition list
`''italic text''`	italic
`'''bold text'''`	bold
`'''''bold italic text'''''`	bold italics
`<small>small text</small>`	small font-size
`O<sub>2</sub>`	subscripts
`[[target page name|link label]]`	internal link to another wiki page
`[[http://www.wikipedia.org Wikipedia]]`	external link
`[[File:Image.png]]`	image

conversion results and the evaluation of the posting segmentation. The paper ends with a conclusion in Section 6.

2 Wikipedia

2.1 Wiki Markup

Wikipedia content is not primarily written in a standard XML-based markup language such as HTML, but in a particular markup language for wikis called *wiki markup*. Text composed with this markup is called wikitext. Wikipedia uses MediaWiki as the software that runs the wiki and converts wikitext into HTML. Some examples of wiki markup are listed in Table 1. Wiki markup also includes HTML tags, for example, `<div class="center">` is used to center a block of text. Tables can be written both as HTML tables and in wiki markup format. An excerpt of wikitext describing a table in wiki markup is given in Figure 1. Some parts of a text such as quotations, poems and source code are marked separately using the `<blockquote>`, `<poem>` and `<syntaxhighlight>` tags.

Wiki markup contains *magic words*, which are special instruction words corresponding to parser functions, variables or behavior switches. For example, `{{lc:string}}` is a magic word corresponding to the parser function that converts the text "string" to lower case. Magic words of type variable are used to print the value of variables, for example `{{PAGENAME}}` will show the name of the current wiki page. The layout or behavior of a page is managed by behavior switches, for example a table of content is generated and replaces the magic word `__TOC__`.

[9] Wikitext of `http://de.wikipedia.org/wiki/Alkalimetalle` from July 27, 2013 dump

```
{| class=&quot;hintergrundfarbe2 rahmenfarbe1&quot; style=&quot;float:  right; clear:
right; margin:1em 0 1em 1em; padding:  0.1em; border-style:  solid; border-width:  1px;
empty-cells:  show&quot; | &lt;u&gt;[[Gruppe des Periodensystems|"'Gruppe"']]&lt;/u&gt;
| align=&quot;right&quot; | "'1"'
|-
| &lt;u&gt;[[Hauptgruppe|"'Hauptgruppe"']]&lt;/u&gt;
| align=&quot;right&quot; | "'1"'
|- align=&quot;center&quot;
| [[Periode des Periodensystems|"'Periode"']]
|- align=&quot;center&quot;
| [[Periode-1-Element|"'1"']]
{{Periodisches System/Element|serie=Nm|aggregat=g|protonen=1|name=Wasserstoff| symbol=H}}
|}
```

Figure 1: A table in wiki markup [9]

MediaWiki implements the notion of transclusion as a function to include some content from a *template page* into another page by using a reference in the wikitext. For example, `{{Archives}}` will include the page Template:Archives. Besides, MediaWiki allows for extending the built-in wiki markup with additional capabilities by creating *custom tags*. The `<nowiki>`, `<score>`, `<math>`, `<ref>` and `<references>` tags are examples of tag extensions for Wikipedia.

Due to the complexity of wiki markup and eventually the wikitext structure, Wikipedia content cannot be readily and easily used as a corpus. The plain text itself without all the heavy and rich structure cannot be easily accessed. We consider wikitext to be not "clean", because it combines HTML tags and wiki markup.

Although the tags in wiki markup should be contained in < and > symbols, many of the tags are "escaped" (i.e. written with < and >), for instance `&lt;small&gt;`. Some of the tags are interpreted by browsers, although many of them are not properly paired (i.e. some escaped elements lack either an opening or a closing tag). It is not always clear why they are present in a wikitext. Partly they seem to express genuine markup that is not or no longer available in HTML (like `&lt;small&gt;`, `&lt;del&gt;`, or `&lt;strike&gt;`), and partly they represent humorous pseudo markup, which will be rendered as markup on the webpage. In the discussions, for example, we found many escaped tags that are formed by interaction words (Beißwenger et al., 2012) as in `&lt;mutmaß&gt; Hängt die <i>Beaufsichtigung</i> ggf. mit der Märzrevolution zusammen ? &lt;/mutmaß&gt;`.

Overall wikitext also frequently contains ill-formed wiki markup such as the lack of a closing symbol for a table or improper line breaks, i.e. often there is no empty line between two paragraphs. These problems may lead to wrong element nesting in the generated HTML, thus creating malformed HTML. In short, considerable effort is needed to pre-process the wikitext before it can be used properly.

Figure 2: The structure of a talk page (representation adopted from Ferschke et al., 2012): a.) page title, b.) unsigned posting with insertion of IP-address, c.) signed posting d.) table of content, e.) thread heading f.) unsigned posting, g.) unstructured discussion thread, h.) discussion thread structured by indentation

2.2 Talk pages

A Wikipedia article may be associated with a *talk page* or *discussion*.[10] A talk page constitutes a piece of wikitext just like an article, i.e. wiki markup is used in it in the same way. On a talk page, users debate an article, often evaluating (parts of) its current content and arguing about whether and how it should be revised or extended, what references and images to include etc. Usually a new edit of an article is accompanied by a contribution on its talk page explaining and justifying the edit. The project described in Ferschke et al. (2012) exploits this fact to construct a corpus of discussions of the Simple English Wikipedia and provides it with dialog act annotations for research on collaborative authoring on the web.

A contribution to a Wikipedia talk page is similar to a *posting* in computer-mediated communication (CMC, see Section 3.2). A posting in CMC such as chat or discussion forums is a piece of text sent to the server by the author at a specific point in time. Postings about one particular topic typically form a thread structure (Beißwenger et al., 2012). Although a posting has a similar status as an utterance in a spoken conversation,

[10] There are also user-related "user talk pages" available, but this paper does not address them.

postings need not immediately follows each other. Together, they form a "written conversation".

In a Wikipedia talk page, a contribution ideally should be associated with its author and posting time information. In CMC corpora, it is essential to be able to distinguish the authors (in order to observe their patterns of interaction, for instance). Since a written conversation is likely to occur non-continuously, the posting time information is needed to keep track of the sequence relations in a thread. In Wikipedia talk pages, the author and posting time information are contained in the user signature. Nevertheless, users not always sign their postings, causing the boundary of a posting to become less clear or unclear.

Strictly speaking, a Wikipedia talk page contribution does not exactly correspond to a posting (as in a chat or forum communication), because in a wiki posting action, a new version of the whole wiki page is posted to the server. This means that users may edit the page in different places within one contribution. Still, a talk page is organised in dialogue structures, in which threads and sequentially ordered, posting-like dialogue turns can be identified. Following Beißwenger et al. (2012), we consider these units of dialogue as postings in our conversion.

3 Related work

3.1 Wikipedia Corpora

Although much effort has been invested in processing wikitext, not many Wikipedia corpora are publically available. Until now, only a few kinds of Wikipedia corpora with XML content have been distributed. Moreover, most of these corpora only include small selections from Wikipedia. Only a few corpora cover all articles and talk pages of a language. The corpora vary in their structures, particularly in the Wikipedia content representations and annotations.

Denoyer and Gallinari (2006) introduced an XML scheme for their Wikipedia corpus and created a multilingual corpus from Wikipedia articles in eight languages. The corpora do not cover many of the Wikipedia articles and exclude the talk pages. The largest corpus was developed for the English Wikipedia and contains about 650,000 articles, while the current English Wikipedia has more than four million articles. The resulting XML includes only a few details; nested sections, for instance, are not represented. Additional English corpora were developed and designed for use in information retrieval and machine learning purposes such as ad-hoc retrieval, categorization and clustering. The corpora were used in INEX (Initiative for the Evaluation of XML Retrieval) and WiQA (Question Answering using Wikipedia) 2006 (Jijkoun and de Rijke, 2006).

Schenkel et al. (2007) proposed YAWN, a system to automatically convert Wikipedia into an XML corpus with semantic annotations. Since the HTML tags in Wikipedia pages seem to generate malformed XML, all the HTML tags are eleminated in the pre-processing level. The elimination causes a loss of information about the text layout, which is tolerated because the focus of the corpus is the page contents. The conversion

includes section, list, tables, links, image links, and highlighting markup (i.e. bold and italic). To ensure that the corpus contains only well-formed documents, the resulting XML documents are checked for well-formedness. The corpus has a markup scheme similar to that of Denoyer and Gallinari (2006), but its structure is more detailed (for instance, nesting of sections is included). The page metadata and the page content are separated by `<header>` and `<body>` tags. Additionaly, the pages are annotated with concepts from WordNet (Fellbaum, 1998), where the concepts are identified by exploiting the lists in and the categories assigned to the pages.

The ILPS (Information and Language Processing Systems) group of the University of Amsterdam provides XML corpora based on Wikipedia [11]. The corpora were built from some portion of Wikipedia articles in different languages. The XML was particularly designed for information retrieval and natural language processing tasks in CLEF (Cross-Language Evaluation Forum). Their conversion tool is available for download.

Bubenhofer et al. (2011) built a comparable, multilingual, annotated XML corpus from all articles and discussions of the German, French, Italian, Polish, Hungarian and Norwegian Bokmål Wikipedia within the EurGr@mm project at the Institut für Deutsche Sprache in Mannheim. As mentioned in Section 1, we adapt their two-stage conversion approach and improve their XSLT Stylesheets.

Wikipedia was also used to build corpora for specific purposes, which are not necessarily formated in XML. Due to the lack of corpora for analyzing collaborative writing, Daxenberger and Gurevych (2012) built such a corpus in their study of the collaborative writing process of Wikipedia articles. First they created the Wikipedia Quality Assessment Corpus by selectively collecting featured and non-featured articles from the English Wikipedia. This corpus was used to compare the quality of the featured and non-featured articles. They also selected 891 revisions of these articles from the English Wikipedia Revision History. The revisions were compared and the difference between two adjacent revisions was defined as an edit. The number of edits was 1995, and the edits were annotated with a category, for example insert, delete or modify. Finally, the corpus contains a list of edits and its annotations.

3.2 Computer-Mediated Communication Corpora

Wikipedia discussions are an instance of computer-mediated communication (CMC). CMC exhibits peculiarities that are features of neither traditional written nor spoken communication, such as a document structure containing postings which are organised in threads or logfiles and also specific orthographic and lexical features such as the use of interaction words (e.g. inflectives **grins**) and symbols (e.g. smileys/emoticons) (Bartz et al., 2013).

The recent initiative *DeRiK - German Reference Corpus of Computer-Mediated Communication* (Beißwenger et al., 2012) has set itself the task of compiling a balanced corpus of German CMC texts to provide an empirical basis for research on German CMC language use. In the long run, DeRiK is planned to include material from all

[11] `http://ilps-vm09.science.uva.nl/WikiXML/` accessed on August 26, 2013

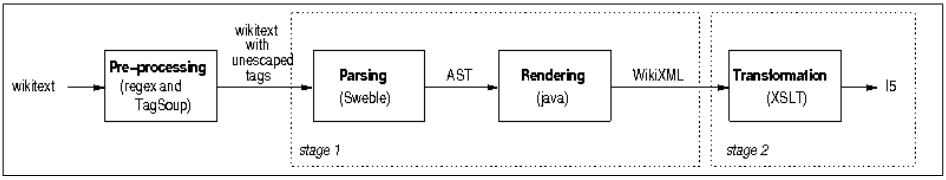

Figure 3: Wikitext to I5 conversion pipeline

major CMC genres such as email, social network communication, discussion forums, chat, Wikipedia discussion, and instant messaging.

The compilation of CMC corpora faces several obstacles. First, web content is subject to copyright restrictions, and in many cases the authors of CMC texts have never agreed that their writings be re-used in linguistic corpus projects, let alone be re-distributed among the linguistic research community, and it seems to be a hard if not impossible task to acquire such permissions retroactively. Fortunately the Wikipedia terms of use permit sharing and reusing of Wikipedia content under sufficiently free and open licenses[12].

Second, current corpus technology is in several ways not suited to the peculiarities of CMC text and CMC documents. For instance, Beißwenger et al. (2012) argue that the TEI P5 corpus encoding scheme lacks markup expressions for the specific macro and micro structure features of CMC mentioned above. To remedy the situation, they introduce a proposal for extending the TEI scheme with a module for CMC. Two encoding examples (Wikipedia talk pages and chat) according to the proposal are provided at http://www.empirikom.net/bin/view/Themen/CmcTEI.

Another CMC corpus that includes a Wikipedia subcorpus has been compiled in the the SoNaR project (van Halteren and Oostdijk, 2014).

4 Method

Our objective is to build Wikipedia corpora in the I5 format described in Section 4.1. The Wikipedia articles and talk pages are separated into two different corpora. To convert wikitext to I5, we use a pipeline architecture composed of four processing modules: pre-processing, parsing, rendering, and transformation. The pipeline is illustrated in Figure 3 and described in Section 4.3. Each Wikipedia article and talk page is processed through this pipeline.

The pipeline takes the entire wikitext of an article as an input. On the other hand, a wikitext of a talk page is segmented into postings, and the pipeline takes each posting as an input. The posting segmentation method is described in Section 4.3.2. In the first module of the pipeline, the input wikitext or posting is pre-processed by using regular expressions and TagSoup. The output of the pre-processing stage is a wikitext with

[12] CC-BY-SA (Creative Commons Attribution-ShareAlike License) and GFDLC (GNU Free Documentation License)

unescaped tags, which is given to the Sweble parser in the parsing module. The Sweble parser parses the wikitext into an abstract syntax tree (AST) that is then rendered as WikiXML. The conversion of wikitext to WikiXML is further explained in Section 4.3.1. The WikiXML content is finally transformed into I5 using XSLT stylesheets (see Section 4.3.3).

4.1 Target Markup I5

The IDS text model defines the hierarchical corpus and text structure of the German Reference Corpus (DeReKo). It was first introduced in 1992 as a home-grown, character-based format. It was recast and further extended in SGML as an adaptation of the Corpus Encoding Standard CES (Ide, 1998), and later, with the arrival of XML, in XCES (Ide et al., 2000). CES/XCES itself was based on the TEI P3 model, restricting TEI to an application to linguistic corpora. The IDS text model includes features that are not part of the TEI model, such as its tripartite corpus structure with the units *corpus - document - text*. In a corpus of literary texts, for example, each individual book constitutes a document, and each short story within a book of short stories, constitutes a text. Each of the three units has its own metadata section modeled on the teiHeader, but also including some bibliographic/metadata components that are not part of the TEI model, such as the time of creation (which may deviate considerably from the official date of publication). Besides, the text model contains elements that are present in the TEI model, but whose content model has been altered (mostly to restrict it more, but sometimes also to extend it). Because of this history, the document grammar used for the IDS text model was an adaption of the official XCES, i.e. IDS-XCES. Over the years, several new features were added when necessary (when new corpus material or analyses required it), but then care was taken that the new elements and attributes were based on the corresponding TEI specification if available.

With the advent of TEI P5 and the new ODD mechanism for TEI customisations, it became possible to specify formally how the IDS text model corresponds to the TEI and in exactly what points it deviates. Thus in 2012, a new document grammar called *I5* was introduced, specified as an ODD document which defines the IDS text model as a TEI P5 customisation (Sperberg-McQueen and Lüngen, 2012). Presently, the I5 model contains 178 elements. On the occasion of the Wikipedia conversion described in the present article, we have additionally introduced the posting structure for corpus documents as described in the proposal of (Beißwenger et al., 2012), because the available I5/TEI elements `<div>` or `<sp>` are not suitable for annotating CMC documents as also argued in Beißwenger et al. (2012). We introduced the `<posting>` element as a "divLike" element with its content model as in the TEI proposal, but we did not introduce the `<timeline>` and `<listPerson>` elements of the proposal because this information is not relevant for DeReKo, and we store it as external XML documents. Neither did we adopt the micro structure elements of the proposal because as yet we do not identify the tokens that are supposed to be annotated by the micro structure elements (i.e. inflectives, emoticons

```
[...]
<div complete="y" n="2" part="N" org="uniform" type="thread" sample="complete">
  <head type="cross">Totensonntag in der DDR</head>
  <posting indentLevel="0" who="WU00000000">
    <p>Hallo, weiß jemand ob es auch einen Totensonntag in der DDR Gab?? Danke</p>
  </posting>
  <posting indentLevel="1" synch="t00121163" who="WU00006525">
    <p>Warum sollte es den dort nicht gegeben haben? Auch in der DDR hörte das
       Kirchenjahr mit dem Ewigkeitssonntag/Totensonntag auf und das neue fing
       mit dem 1. Advent wieder an. --<autoSignature></autoSignature></p>
  </posting>
  <posting indentLevel="0" synch="t00121164" who="WU00031907">
    <p>Und weiß jemand, ob es den Totensonntag auch in Dänemark gibt??? DANKE!
      <hi rend="pt">(<hi rend="it">nicht
      <ref targOrder="u" target="de.wikipedia.org/wiki/Hilfe:Signatur">
      signierter</ref> Beitrag von</hi><autoSignature></autoSignature></hi></p>
  </posting>
</div>
[...]
```

Figure 4: Structure of discussion thread and postings according to the TEI-proposal by Beißwenger et al. (2012)

etc.). In Figure 4, we show the representation of a part of the Wikipedia talk page for *Ewigkeitssonntag*[13].

In the I5 target representation of the present Wikipedia conversion, the German Wikipedia articles collection and the German Wikipedia discussion collection each form a corpus (an XML document with the root element `<idsCorpus>`), all articles/discussions with the same initial letter of the headword form a document (element `<idsDoc>`), and each article/talk page forms a text (element `<idsText>`).

In comparison, the XML that we use for the intermediate representation between the two stages is less restricted, and no document grammar exists for it. For encoding the basic document structure, HTML tags such as `<div>` are used, but ad-hoc tag names derived from the wiki markup such as `<gallery>` also occur. Thus it mirrors more directly the features of the wiki markup, but, unlike the former, it is always well-formed.

4.2 The Sweble Parser

Dohrn and Riehle (2011) argue that the possiblities of processing wiki content are rather limited due to the fact that the wiki software running the wikis only generates HTML, which may be (and frequently is) not well-formed. The complexity of wiki markup makes it difficult for computer programs to access the wiki content. For this reason, they proposed the Sweble parser as a Java library that can handle the complexity of wiki markup and generates a machine-readable representation. The representation is an object model that can be further used to render well-formed HTML or XML. Sweble parses a wikitext using Parsing Expression Grammars (PEGS, Ford, 2004)

[13] `http://de.wikipedia.org/wiki/Diskussion:Ewigkeitssonntag`

into an object model in an abstract syntax tree (AST, Mogensen, 2011). An AST is a data structure representing the syntactic structure of a code implementation as a tree containing nodes for constants, variables, operators and statements. In Sweble, an AST is used to represent the parsing output, i.e. the objects generated from a wikitext.

Sweble takes a wikitext as input and processes it through a pipeline architecture composed of five processing steps: encoding validation, preprocessing, expansion, parsing and post-processing. In the encoding validation step, illegal characters that can harm the next processing steps are wrapped into certain entities. In the preprocessing step, redirect links, tag extensions, templates and unknown XML elements are handled and the AST nodes of the wikitext are created. Expansion is an optional step to extend the AST nodes by resolving the templates, magic words, parser functions, and tag extensions. Before starting parsing, the AST nodes are converted back to wikitext. The wikitext is subsequently analyzed by a PEG parser, and the parser generates an AST modeling the syntax and semantics of the wikitext. Finally, the post-processing step is applied to the AST. The apostrophes are interpreted and handled, the XML tags are matched, and the paragraphs are put together. The AST can be further processed by using the Visitor design pattern (Metsker, 2002), for example to generate a HTML page.

4.3 Processing Wikipedia: Parsing and Transformation Approach

4.3.1 Converting Wikitext to WikiXML

First, Wikipedia articles and talk pages are selected, and other pages, such as user, user discussion, file, and help pages, are filtered out. Since the page types correspond to their namespaces, the filtering is done by identifying the namespaces of the pages and selecting only those pages with article or talk namespaces. The redirect pages are also filtered out by identifying the redirect title in the page metadata.

Before going through the conversion pipeline, the wikitext of a Wikipedia page is separated from page metadata. The wikitext of talk pages also goes through the posting segmentation process described in Section 4.3.2. Subsequently, it becomes the input of the pre-processing stage including several tasks: unescaping tags, handling problematic symbols and correcting tags that are not well-formed.

As described in Section 2.1, wikitext contains many escaped tags. Since we intend to capture as much structure as possible, we unescape all the tags, except for the tags embedded in link markup. For example in `[[Datei:Sigmund Freud LIFE.jpg|miniatur|<center>Sigmund Freud auf einer Fotografie 1921 Aufnahme von [[Max Halberstadt]]<center>]]`, the `<center>` tags remain escaped. The left angle bracket symbols may cause problems in parsing because Sweble may falsely recognize them as a tag definition symbol, while they may alternatively be used to signify “lower than” (e.g. < 1 %). Such brackets are escaped to prevent false tag correction.

In this work, we use the Sweble parser version 2.0.0.alpha-2 for the parsing stage. The Sweble parser features a function that corrects the tags that are not properly

Table 2: Wikitext to XML conversion

Wiki Markup	XML
== Heading ==	Heading
----	<hr />
* Item	 Item
; Term	<dt>Term</dt>
: Definition	<dl>Definition</dl>
"""bold italic text"""'	bold italic text
[http://www.wikipedia.org Wikipedia]	Wikipedia

paired. However, its strategy to handle missing closing tags is rather verbose because it keeps adding closing tags until the end of the wikitext. To reduce this repetition, we use the TagSoup parser[14] (Cowan, 2002) which has a similar strategy as the Sweble parser. Instead of passing a complete wikitext, we segment the wikitext by each empty line and pass a paragraph-like segment of the wikitext as an input to TagSoup. This way, the repetition only affects a much smaller scope. Although this strategy does not ensure that the missing closing tags are placed properly, it ensures that the WikiXML is well-formed.

The wikitext output from the TagSoup parser becomes the input for the parsing stage. The Sweble parser models the wikitext into an AST. AST nodes represent wikitext elements such as paragraphs, links, and so on. Some wiki markup is not handled by Sweble, including file links.

For the rendering stage, we adapted the HTML renderer of the Sweble library and re-implemented it as an XML renderer. The XML renderer defines how the AST nodes should be expressed in WikiXML. Table 2 highlights some conversion rules from wikitext to WikiXML. Primarily, those wikitext elements which have HTML tag counterparts are converted to their HTML tag counterparts. Internal links referring to pages in Wikipedia, external links referring to sources outside Wikipedia, and image links are resolved to HTML links. Comments are removed. Templates and tag-extensions are wrapped in a tag. Other XML tags, such as <gallery> and <timeline>, are simply copied. Our WikiXML output has richer text structures than that of Schenkel et al. (2007) as it contains more tag types and Sweble parses more of the wiki markup. Like in Schenkel et al. (2007), each resulting WikiXML page containing the page metadata and the XML content is also checked for well-formedness.

4.3.2 Posting Segmentation

Considering the possible structures of posting threads in talk pages described in Section 3.2, the task of segmenting a talk page section into postings is not a trivial problem. To

[14] http://ccil.org/čowan/XML/tagsoup/

deal with this, we analyzed the structural and textual characteristics of the postings in the wikitext and created heuristic rules for segmenting them. Each posting is converted into XML by using the method described below. Moreover, the postings are annotated according to the TEI proposal for CMC corpora (Beißwenger et al., 2012). Like in the proposal, the user signature is anonymized, and the original information about the author and timestamp of a contribution is recorded in two separate XML documents. The idea behind this is to protect confidential information and to manage authorization for accessing this information.

User signature

```
Fachliche Anmerkung @Stern:  ARIS ist keine Architektur.  ARIS ist ein
Toolset zum Beschreiben von Architekturen, das ist ein Unterschied.
[[Benutzer:Ulrich.fuchs|Uli]] 14:11, 21.  Mär 2004 (CET)
```

Help signature

```
Umknicken mit der Fußsohle nach Innen bedeutet ja das
die Außenbänder beansprucht werden.  Mit logischem Denken
könnte man sich diese Frage auch selber beantworten.
&lt;small&gt;("nicht [[Hilfe:Signatur|signierter]] Beitrag von"
[[Spezial:Beiträge/212.7.177.129|212.7.177.129]] ([[Benutzer
Diskussion:212.7.177.129|Diskussion]]) 14:27, 26.  Jan.  2012 (CET))
&lt;/small&gt;
```

Unsigned Template

```
Wie werden SMD Elkos richtig gepolt?  {{unsigned|84.56.124.232|12:51, 19.
Sep.  2006 (CEST)|ALT=nicht}}
```

Horizontal line

```
Das wird glaub doch mit ß geschrieben.  Siehe dazu
[[diskussion:Rechtschreibung]]
----------
```

Heading

```
Außenband ist korrekt (ß nach langen Vokalen und Dipthongen) ss nach
kurzen Vokalen, z.B. Kuss, Fluss.  [http://www.duden.de/ Duden, Link
'Neue Rechtschreibung', Link 'Crashkurs'; Regel 4]

==&quot;Hoch/runter&quot;==
```

Figure 5: Posting delimiters [15]

For the posting segmentation, we defined a set of posting delimiters. A posting delimiter is some piece of text structure indicating the beginning or the end of a posting, thus separating one posting from another. We identified six wiki markup elements which

[15] Wikitext excerpt of `http://de.wikipedia.org/wiki/Diskussion:Architektur/Archiv`, `http://de.wikipedia.org/wiki/Diskussion:Außenbandruptur_des_oberen_Sprunggelenkes`, `http://de.wikipedia.org/wiki/Diskussion:Elektrolytkondensator/Archiv` from July 27, 2013 dump

```
Wer Mut hat, könnte vielleicht das Modell zwei der
Wikipedia-Begriffsklärung umsetzen - trau mich da nicht ran.
[[Wikipedia:Begriffsklärung]] [[Benutzer:Marc Tobias Wenzel|mTob]]
```

Posting level 1

```
:Bin mir da auch noch unsicher - wir haben noch irgendwie die
[[Baukunst]], die da mit rein muss.  Vielleicht doch besser Modell I
und unter [[Baukunst]] die &quot;eigentliche&quot; Architektur abwickeln?
[[Benutzer:Ulrich.fuchs|Uli]] 17:20, 12.  Jun 2003 (CEST)
```

Posting level 2

```
::Architektur/Baukunst gehört hier her.  Den Rest kann man unter
&quot;Architektur (Begriffserkl.)&quot; abhandeln.
```

Figure 6: Posting thread [16]

can be used as posting delimiters: user signature, help signature, unsigned template, heading, horizontal line and indentation markup. Figure 5 shows some postings in which these delimiters are used.

In Wikipedia, a user signature is an internal link to a user page followed by a posting timestamp. Ideally a posting should be signed with a user signature. A help signature is an internal link to the help page about signing a posting. When a posting is not signed, a help signature is added together with the user's IP address information and posting timestamp. Alternatively, the *unsigned* template is used to mark a posting without a signature. To detect the signatures in wikitext, we defined several regular expressions. We also defined regular expression for extracting the author and posting timestamp information from the signatures. Furthermore, the end of a posting can be identified by a horizontal line markup formed by multiple hyphen symbols, or a heading markup. Thus, any text after the last posting and before a horizontal line or heading markup is combined to form a posting.

A posting can be followed by other postings on different levels of indentation, thus creating a thread structure, which we call a posting thread. A posting thread has a smaller scope than a section thread. A section thread covers all the postings under a heading, while a posting thread covers all the postings within a list of indentations. The depth of an indentation determines the level of a posting, specified in the `@n` attribute. An example of a posting thread in wikitext is given in Figure 6. We consider each piece of wikitext with an indentation as a posting, i.e. the indentation markup is also used as a posting delimiter.

4.3.3 Converting WikiXML to I5

Each WikiXML page content is converted to I5 by means of XSLT stylesheets. We adapted the stylesheets first written by Bubenhofer et al. (2011). Firstly, the WikiXML

[16] Wikitext excerpt of `http://de.wikipedia.org/wiki/Diskussion:Architektur/Archiv` from July 27, 2013 dump

page content is split into sections by grouping the paragraph-like elements by the occurrence of headings. Since sections may have subsections, the grouping is done recursively. The level of a section is determined by the size of its heading, for example `<h2>` in XML indicates a `<div type="section" n="2">` in I5. The thread sections for postings, however, are rendered as `<div type="thread">`. Although Bubenhofer et al. (2011) implemented grouping of sections and subsections, the grouping for subsections is restricted only to subsections of the immediate sub-level. We improved the flexibity of the grouping by allowing a section to have subsections of any level. We also handle grouping of sections which are embedded in elements other than section elements.

Bubenhofer et al. (2011) simplified the conversion to IDS-XCES by including a conversion of some wiki markup to TEI elements already in their XML, such as the text highlighting elements indicating italic or bold printing. Therefore, the content of paragraphs and lists does not need to be transformed and can be simply copied in the conversion to XCES. Our approach, however, maintains a clear separation between WikiXML and I5 by using HTML tags instead of TEI elements in the WikiXML. Therefore, we substituted the corresponding TEI templates with templates handling HTML tags. We also added more templates for further WikiXML elements, which are not handled by Bubenhofer et al. (2011), such as `<dl>`, `<blockquote>` and `<caption>`.

Our WikiXML to I5 conversion involves many more transformation processes. It does not simply copy the content of paragraphs and lists, but carries out transformations for the elements inside the paragraphs and lists. First of all, the content of sections is tested for paragraph-like elements vs. phrase elements. Paragraph-like elements are elements that can appear between paragraphs, such as `<poem>`, `<list>`, and `<quote>`. Phrase elements are elements that can appear within paragraph-like elements, such as `<b>`, `<i>`,and `<sub>`. Since in I5, a section may only contain paragraph-like elements, all phrase elements are wrapped in a (new) paragraph element (`<p>`).

The `<posting>` element in WikiXML is defined based on Beißwenger et al. (2012), thus it has a structure similar to `<posting>` in I5. The author and time information are recorded in two additional XML documents. `<poem>` elements are specified in more detail in I5 with `<l>` elements designating poem line. The different kinds of lists in WikiXML are transformed into `<list>` elements.

We counted the frequencies of occurrence of tags in the intermediate WikiXML corpora and implemented specific XSLT templates for the frequently occurring elements. In doing so, we implemented context-dependent transformations for certain elements such as `<div>`. Table 3 summarizes the conversion from WikiXML to I5 for the frequently appearing tags. We delete all less frequently occurring tags (< 500 times), but not their contents. Besides, we re-generate those tags that correspond to interaction words as escaped tags in I5 because they might be of particular interest in the linguistic analysis of CMC corpora.

Table 3: WikiXML to I5 Conversion

WikiXML Tags	I5
<p>	<p>
<posting>	<posting>
<poem>	<poem> with <l> elements
, , <dl>	<list>
, <dd>, or <dt>	<item>
<blockquote>	<quote>
<caption>,	<caption>
	<gap desc="*class name*"/>
whose class appears often	
with other classes	value of the element
<pre>	or value of the element when it appears inside a paragraph
	when appears inside a <text> or a element, otherwise value of the element
<table>, <timeline>, <references>, <gallery>	<gap desc="*tag name*"/>
<span class"=*tag-extension or template*">	<gap desc="*class name*"/>
<abbr>	<abbr>
 	<lb>
<link>	<ref>
<ref>, <Ref>, <REF>	xsl:apply-templates
<b> or	<hi rend="bo">
<i>	<hi rend="it" >
<u>	<hi rend="ul">
<small> or <big>	<hi rend="pt">
<sup>	<hi rend="super">
<sub>, <tt>, , <code>, <source>	<hi rend="*tag name*">
<font>	<hi rend="*font-style*">
<syntaxhighlight>	<hi rend="syntaxhighlight">
<s>, <strike>, <del>	value of element

5 Results and Discussion

We built WikiXML article and discussion corpora for the German Wikipedia dumps from July 27, 2013, originally containing almost 2.7 million articles and about 570,000 talk pages. We removed over 1.1 million redirect pages from the set of articles and over

50 redirect pages from the set of talk pages. We also removed about 15,000 empty pages from the set of talk pages. Eventually, we parsed approximately 1.6 million articles and 0.55 million talk pages. In the parsing stage, 142 articles and 24 postings failed to get parsed. Moreover, one parsed article and two parsed talk pages were not well-formed. All the parsed and well-formed pages are successfully transformed to I5. The resulting corpora were successfully validated against the I5 DTD. The I5 file containing the articles is 16GB and contains 678 million running words. The discussions file is 4.8GB and contains 264 million running words.

5.1 Evaluation of the Posting Segmentation

By using the posting delimiters described in 4.3.2, the posting segmentation program identifies over 5.4 million postings in the German Wikipedia talk pages. To evaluate the performance of the program, we calculated precision and recall of the postings segmented by the program against posting annotations of two human annotators. Furthermore, we compare the performance of the program with a baseline segmentation using only user signatures as the posting delimiter.

Many of the talk pages are very short containing only one or two postings. To represent the variation of talk page length, we selected 49 WikiXML talk pages (7.5KB average per page, 364KB in total) randomly but with a certain distribution of long, medium long, and short pages (12 pages >10kb, 7 pages 5-10kb, 20 pages <5kb) for the evaluation dataset. We removed the <posting> tags in the WikiXML talk pages and gave them to the annotators. The annotators consulted the corresponding webpages of the talk pages, thus judging by formal as well as textual indicators where a new posting starts, and annotated the talk pages with new <posting> tags using an XML editor. The first annotator annotated in total 646 postings and the second 602 postings in the dataset.

To measure the agreement between the two annotators, we calculated Cohen's Kappa based on boundary matches. By their annotations, the annotators had categorised each of the 1024 potential boundaries (given by all paragraph-like elements contained) into either a posting/posting boundary, a posting/non-posting boundary, or a none-boundary. Based on the resulting confusion matrix, the Kappa coefficient for the two annotators is κ=0.76. This suggests that the agreement between the two annotators is fairly good. It also suggests that there is a number of postings which are ambiguous, i.e. the boundary between some postings is not obvious.

In contrast to the annotators' segmentations, the program generates 817 postings, and the baseline 499 postings. These figures suggest that the program is somewhat overly constrained and generates too many postings. One reason for this is that sometimes the usage of the wiki markup is ambiguous. For example, the indentation markup is usually used for structuring a posting thread (see Section 4.3.2), however, Wikipedia authors also use it for other purposes, such as marking a quote. On the other hand, the segmentation of the baseline is rather sparse and generates too few postings, because the baseline does not capture the unsigned postings.

Table 4: Posting segmentation performance measures

	Annotator	micro average		macro average	
		P	R	P	R
Baseline, posting-based	1	53.51	41.33	41.96	37.03
	2	42.49	35.22	34.97	32.45
Program, posting-based	1	63.40	80.19	80.61	87.86
	2	58.75	79.73	76.80	82.85
Program, boundary-based	1	75.76	96.87	85.77	97.06
	2	70.75	96.49	86.77	96.56

For the evaluation of the posting segmentation program against the two manual annotations, we provide posting-based and boundary-based precision (P) and recall (R) in comparison with the segmentations by annotators 1 and 2 in Table 4. For the posting-based measures, we used posting text to match, whereas for the boundary-based measures, we counted only boundaries that were placed between two posting segments. The micro vs. macro scopes were averaged over the set of 49 test documents described above.

The results of the first four lines in Table 4 show that the program clearly outperforms the baseline. The micro precision values are significantly lower than the recall values in all evaluation modes, reflecting the above mentioned overgeneration of segments by the program. Compared to the manual annotations, the program is able to identify about 80% of the postings marked by annotator 1 and of the postings marked by annotator 2 (cf. the micro average recalls in lines 3 and 4 Table 4).

Measures of segmentation similarity are generally based on boundary matching, not segment matching, see Inkpen, Diana and Chris Fournier (2012) for an overview. However, we additionally evaluated segment matching because there are some limitations when evaluating boundary matching.

A boundary-based evaluation is limited by the fact that at least one boundary must exist to perform a comparison. In the case where a page is judged to contain only one posting, no boundary actually exists, because the beginning and end of a file are usually not included in the set of boundaries. Our evaluation dataset, however, contains pages annotated as containing only one posting. Thus, for the boundary-based evaluation, we created one boundary after each of these postings.

Besides, boundary matching commonly yields more matches than segment matching. For instance, suppose the program generated one posting where a human annotator segmented two postings (i.e. `|xx|` vs. `|x|x|`). Then the program got 0 out of 2 postings, but still 2 out of 3 boundaries correct. Because of this, the boundary-based evaluation measures (lines 5 and 6) are generally higher than the posting-based ones (lines 3 and 4).

5.2 Discussion

We consider a parsing approach as an improvement over the use of regular expressions as in Bubenhofer et al. (2011) to handle wiki markup, because regular expressions are difficult to maintain. The Sweble parser is an on-going project that specializes in handling the complexities of wiki markup. In fact, the Sweble parser is able to handle almost all wiki markup, including templates and tag extensions. It parses a wikitext into an object model, allowing a machine to conveniently access and manipulate wikipedia content. In our case, Sweble was able to parse almost all the German Wikipedia articles and talk pages. A few pages that exhibited very complex wiki markup failed to get parsed. In the example shown in Figure 7, a complex internal link to the *Römisch-katholische Kirche* is placed in a table, which is part of the caption of the link to an image file.

```
[[Datei:Vallásos és nem hívő közösségek Magyarországon.png|miniatur|hochkant=3.0|Die
regionale Verteilung der Konfessionen nach der Volkszählung 2001:
{| class=&quot;wikitable&quot; style=&quot;margin:  auto;cellspacing=&quot;0&quot;;
font-size=80%
!  Größte Religions-&lt;br /&gt;gemeinschaft
!  [[Römisch-katholische Kirche|Römisch-katholisch]] ...  |} ]]
```

Figure 7: Wikitext excerpt failed to be parsed by Sweble [17]

Besides its complex structure, wikitext is problematic because wiki markup is often not used properly. It contains many unmatched tags and in some cases, tags do not match because the opening or closing tag is wrongly placed as for instance in `<font color=&quot;#777777&quot;>{{NaviBlock</font>..}}` where the closing `<font>` tag occurs inside a template. We attempted to improve Sweble's current strategy to fix ill-formed wiki markup by employing the TagSoup parser (Section 4.3.1), but some wiki markup is not handled by TagSoup because TagSoup only parses pure HTML or XML. From remaining ill-formed wiki markup, Sweble generates awkward XML. Figure 8 shows an example of an unmatched wiki markup tag apparently intended to print "Niger-Kongo" in bold. Since the unmatched bold tag occurs in a list, a completed bold tag is repeated until the end of the list and then also for the rest of the wikitext.

A wikitext itself may contain awkward XML behavior. For example, a phrase element may contain a paragraph element. This behavior is not allowed in I5, and thus also makes the conversion from WikiXML to I5 difficult. Figure 9 shows a list wrapped by a `<small>` tag. This kind of structure is not properly handled by Sweble, because Sweble expects a tag to be matched per line in a list. Hence, Sweble will treat the `<small>` tag as an incorrect tag.

[17] Wikipedia excerpt of `http://de.wikipedia.org/wiki/Ungarn` from July 27, 2013 dump
[18] Wikipedia excerpt of `http://de.wikipedia.org/wiki/Ayere-Ahan` from July 27, 2013 dump
[19] Wikitext excerpt of `http://de.wikipedia.org/wiki/Denkmal_zur_Geschichte_der_deutschen_Arbeiterbewegung_an_der_Gedenkstaette_Eisenacher_Parteitag_1869` from July 27, 2013 dump

Wikitext

```
** '''Niger-Kongo
** Volta-Kongo
*** Süd-Volta-Kongo
**** Benue-Kongo
***** West-Benue-Kongo
****** '''Ayere-Ahan'''
******* Ayere (3.000 Sprecher, Kwara State, Gebiet Oyi, Kabba District)
******* Ahan (300 Sprecher, Ondo State, Gebiet Ekiti, Städte Ajowa,
Igashi, Omou)
```

Sweble output

```
    <ul>
        <li> <b>Niger-Kongo</b>
            <ul><b></b>
                <li><b> Volta-Kongo</b>
                    <ul><b></b>
                        <li><b> Süd-Volta-Kongo</b>
                            <ul><b></b>
                                <li><b> Benue-Kongo</b>
                                    <ul><b></b>
                                        <li><b> West-Benue-Kongo
                                                <ul>
                                                    <li> Ayere-Ahan<b></b>
                                                        <ul><b></b>
                                                            <li><b> Ayere (3.000 Sprecher, Kwara State, Gebiet Oyi, Kabba
District)</b></li>
                                                            <li> Ahan (300 Sprecher, Ondo State, Gebiet Ekiti, Städte Ajowa, Igashi,
Omou)</li>
                                                        </ul></li>
                                                </ul></b></li>
                                    </ul></li>
                            </ul></li>
                    </ul></li>
            </ul></li>
    </ul><b>
```

Figure 8: Unmatched wiki markup [18]

Originally, the WikiXML to I5 conversion using XSLT was supposed to straightforwardly map the elements of the intermediate, unrestricted WikiXML on I5 elements, and also to filter out tags (and sometimes their content) that are not relevant in linguistic corpora in general and have no equivalent in I5. However, we eventually used the stylesheets to handle also the awkward XML structure, either because of the over-generation of tags introduced in the tag-correction process, or the wikitext behavior itself.

Our conversion system can be used for new German Wikipedia dumps in the future and also for dumps of other languages. We have tested the system by converting the French Wikipedia from Sep 4, 2013. Only a very small number of articles and postings failed to get converted or have invalid I5 (far less than 1%). However, we hope that Sweble will be further improved to deal with unstructured wiki markup esp. the problem of unmatched tags.

The treatment of the escaped XML tags in wikitext described above in the second stage is not entirely language-independent because the tags themselves are not. We have seen examples of infrequently occurring but linguistically interesting interaction words used as escaped tags in the German talk pages, but there are also German-specific frequently occurring tags such as `<div class="BoxenVerschmelzen">` occurring in the

```
&lt;small&gt; :DER EINZELNE HAT ZWEI AUGEN
:DIE PARTEI HAT TAUSEND AUGEN.
::DIE PARTEI SIEHT SIEBEN STAATEN
::DER EINZELNE SIEHT EINE STADT.
:::DER EINZELNE HAT SEINE STUNDE,
:::ABER DIE PARTEI HAT VIELE STUNDEN.
::::DER EINZELNE KANN VERNICHTET WERDEN,
::::ABER DIE PARTEI KANN NICHT VERNICHTET WERDEN.
:::::DENN SIE IST DER VORTRUPP DER MASSEN
:::::UND FÜHRT IHREN KAMPF
::::::MIT DEN METHODEN DER KLASSIKER, WELCHE GESCHÖPFT SIND
::::::AUS DER KENNTNIS DER WIRKLICHKEIT. &lt;/small&gt;
```

Figure 9: A list wrapped by a phrase element [19]

intermediate WikiXML. Our treatment of such tags is based on a derived frequency list of tags in the German Wikipedia and thus not readily applicable to the Wikipedia of other languages. Using our present tools, language-specific tags of other languages will be simply deleted and their content will be analysed for further tags by XSLT templates in the second stage of the conversion, which is a desirable behaviour as long as specific XSLT templates have not been defined for them.

Compared with other Wikipedia corpora, our I5 corpora contain richer text structures as the conversion covers more wiki markup. Unlike many corpora that only samples from Wikipedia articles, our corpora contain almost all Wikipedia articles (99.9%). Moreover, we also provide large discussion corpora from Wikipedia talk pages, including markup for postings, which have not been available before.

6 Conclusion

Our conversion system implementing the two-stage approach, i.e. first converting wikitext to unrestricted WikiXML and then filtering and transforming the WikiXML to valid I5, has proved adequate for building linguistic corpora from Wikipedia. We have shown that the resulting corpora cover almost all German Wikipedia articles and talk pages, and that the system can also be used for languages other than German. For the second conversion stage, we have improved Bubenhofer et al. (2011)'s approach and included more structures in the Wikipedia corpora.

For the discussion corpus, we have introduced a posting segmentation, the results of which highly correspond to annotations made by humans. Moreover, we have extended the document grammar I5 to accommodate thread and posting structures as occurring in Wikipedia talk pages according to the TEI proposal for CMC corpora by Beißwenger et al. (2012). Consequently, the discussion corpus can also be used in the linguistic analysis of CMC data.

The Wikipedia corpora (the article and discussion corpora described in this paper) will be available as a subcorpus in the COSMAS II corpus search and analysis system under http://www.ids-mannheim.de/cosmas2/ as of 2014. The XML corpus files (the article and discussion corpora in I5 but also in the intermediate WikiXML versions) are available for download from the Institut für Deutsche Sprache in Mannheim under the license CC-BY-SA. The I5 corpora also contain structured metadata in the TEI header format for each article and discussion page, and the running texts are also provided with markup for sentence boundaries (`<s>`) from our sentence splitter. Additionally, we offer POS tagging of the I5 corpora from the Tree Tagger (Schmid, 1994) in separate files, represented as stand-off annotations. For download and more information, see: http://www1.ids-mannheim.de/kl/projekte/korpora/verfuegbarkeit.html.

Literatur

Atserias, J., Zaragoza, H., Ciaramita, M., and Attardi, G. (2008). Semantically Annotated Snapshot of the English Wikipedia. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Barett, D. J. (2009). *MediaWiki*. O'Reilly Media, Inc., USA.

Bartz, T., Beißwenger, M., and Storrer, A. (2013). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics*, 28(1):157–198.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative [Online]*, 3.

Bubenhofer, N., Haupt, S., and Schwinn, H. (2011). A comparable Wikipedia corpus: From Wiki syntax to POS Tagged XML. In Hedeland, H., Schmidt, T., and Wörner, K., editors, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, volume 96B of *Working Papers in Multilingualism*, pages 141–144. Hamburg University.

Cowan, J. (2002). TagSoup: A SAX parser in Java for nasty, ugly HTML.

Daxenberger, J. and Gurevych, I. (2012). A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 711–726.

Denoyer, L. and Gallinari, P. (2006). The Wikipedia XML Corpus. *SIGIR Forum*, 40(1):64–69.

Dohrn, H. and Riehle, D. (2011). Design and implementation of the Sweble Wikitext parser: unlocking the structured data of Wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 72–81, New York, NY, USA. ACM.

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012). Behind the article: recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 777–786, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ford, B. (2004). Parsing expression grammars: a recognition-based syntactic foundation. In *ACM SIGPLAN Notices*, volume 39, pages 111–122. ACM.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In *Proceedings of the First International Language Resources and Evaluation Conference (LREC)*, page 463–470, Grananda, Spain.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, page 825–830, Athens, Greece.

Inkpen, Diana and Chris Fournier (2012). Segmentation similarity and agreement. In *Proceedings of Human Language Technologies: The 2012 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 152–161, Stroudsberg, PA.

Jijkoun, V. and de Rijke, M. (2006). Overview of the WiQA Task at CLEF 2006. In *CLEF*, pages 265–274.

Kupietz, M. and Lüngen, H. (2014). Recent Developments in DeReKo. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Metsker, S. J. (2002). *Design Patterns Java Workbook*. Pearson Education, Inc., US.

Mogensen, T. A. (2011). *Introduction to Compiler Design*. Springer, London.

Schenkel, R., Suchanek, F. M., and Kasneci, G. (2007). YAWN: A Semantically Annotated Wikipedia XML Corpus. In *12th GI Conference on Databases in Business, Technology and Web (BTW 2007)*, Aachen,Germany.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Sperberg-McQueen, C. and Lüngen, H. (2012). A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative (jTEI)*, 3.

van Halteren, H. and Oostdijk, N. (2014). Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens. *Journal for Language Technology and Computational Linguistics*, 29.

Wilbert Spooren, Tessa van Charldorp

Challenges and experiences in collecting a chat corpus

Abstract

Present day access to a wealth of electronically available linguistic data creates enormous opportunities for cutting edge research questions and analyses. Computer-mediated communication (CMC) data are specifically interesting, for example because the multimodal character of new media puts our ideas about discourse issues like coherence to the test. At the same time CMC data are ephemeral, because of rapid changing technology. That is why we urgently need to collect CMC discourse data before the technology becomes obsolete. This paper describes a number of challenges we encountered when collecting a chat corpus with data from secondary school children in Amsterdam. These challenges are various in nature: logistic, ethical and technological.

1 Introduction

Present day access to a wealth of electronically available linguistic data creates enormous opportunities for cutting edge research questions and analyses. The data used in such analyses often come in systematically collected collections of texts that are, in one way or the other, representative of the population of discourses from which they are taken. Computer-mediated communication (CMC) data are specifically interesting for a number of reasons (cf. Herring 2004; 2013 for an extensive discussion of possibilities and pitfalls of so-called Computer-Mediated Discourse Analysis, or CMDA). A much-discussed issue is how classical distinctions like that between written and oral data, as discussed for example in Chafe (1982), break up in CMC, as witnessed by Baron's (2008) analyses of email and instant messages. For example, the conceptualization of how coherence is established in CMC is based on language use and routines from written language (such as coherence relations and their markers) as well as from spoken language (such as adjacency pairs and other interactional 'devices') (cf. Sanders and Spooren 2013). Similarly, technological advancements allow users to express multimodal content in ways not imagined two decades ago. This type of multimodality creates new forms of communication that may not be similar to the traditional ways in which we 'write' and 'speak.'

The rapid technological developments bring into existence both new media and new genres that are sometimes short-lived. These developments make the analysis of CMC a discipline in which we operate in a permanent laboratory for the study of genre and the role of language in it. At the same time, the rise and fall of technologies and genres like Second Life gaming and MSN chat show that it is imperative to collect these materials for scientific study before they become obsolete. A large and varied collection of CMC material allows us to answer questions not only about language change across generations or language use across various CMC modes, but also about the creativity and adaptation of language use in technologically advanced or restricted environments (for example, see Herring 2004). Furthermore, sociolinguistic questions about age, gender and technological experience in relation to language use can be explored when such data are available alongside the corpus. Once we have collected this type of discourse systematically it allows us to study the use

and construction of computer-mediated language use over time and throughout various types of communication.

Before corpus linguistic analysis can be done, a corpus first needs to be created and annotated. High quality written and spoken corpora have been available for decades, and within the field of corpus linguistics standardization of corpus annotation has been the main objective of the last decade or so. However, CMC is a relatively new phenomenon, and despite its many interesting research opportunities, relatively few CMC corpora are available, especially for Dutch (cf. Oostdijk et al. 2013, for a recent corpus comprising some forms of CMC data, among which the sub corpus described in this paper).

After providing a short overview on past corpus creation literature and the technicalities of our corpus, we present the challenges we met when we put together a chat corpus. The descriptions are intended as a very practical account of what we encountered when building a chat corpus in 2004-2006.

2 Literature

The ultimate goal of any corpus research is to have naturally occurring data available in which phenomena of interest can be found to a degree that is representative of the way these data occur in the type of language use that the researcher is interested in. For quantitative corpus studies this means that the occurrence of the phenomena of interest resembles that of the population. This will only occur if the corpus is based on random sampling and if the corpus is balanced, i.e., the size of the sub registers in the corpus reflects that of the language use the researcher is interested in (cf. Gries and Newman 2013, and the references cited there). For qualitative corpus analysis representativeness means that all the varieties of language use that the researcher is interested in should occur in the corpus. It also means that we should have as much data available about the context of language use as possible. Both types of requirements pose serious challenges to building a CMC corpus. It is generally acknowledged that these requirements are ideals, and that in actual practice corpus builders often deviate from them, partially due to pragmatic reasons (also see pragmatic challenges when building an SMS corpus as explained by Tagg 2009).

CMC corpora to date – for example, the Deutsches Referenzkorpus zur internetbasierten Kommunikation (Beißwenger et al. 2013), the Dortmund Chat Corpus (Beißwenger 2013), the Queer Chat-Room corpus (King 2009), or the Netlog Corpus (Kestemont et al. 2011) – scarcely describe the challenges they met when collecting CMC materials. Most authors focus on the next phase: format and annotation challenges when building corpora (cf. Beißwenger and Storrer 2008, section 3 as an example). Ethical issues are mentioned in passing, but authors are quick to suggest that it is "unrealistic to obtain a declaration of consent for the recording and subsequent use of users' statements for research purposes." (Beißwenger and Storrer 2008, section 3.1.8). This is especially the case when CMC corpora are based on publicly available communication on the internet, for example public chatrooms or public fora.

This article elaborates on the often overlooked phase of corpus creation: collecting data. We describe the challenges we met when building a CMC corpus and demonstrate that CMC material can be collected with consent. We made decisions that balance between the

ideal requirements for a representative corpus and the pragmatic reality. Before setting out the seven challenges, a short overview of the corpus characteristics is given.

3 The ChatIG corpus

3.1 Background

The corpus project was part of the "Vrolijke School" ["Happy School"] project – a collaboration between VU University Amsterdam and Ignatius Gymnasium Amsterdam (an Amsterdam-based grammar school) – aimed at creating awareness of the pleasures and complexities of research amongst secondary school students. In this project various faculties of VU University Amsterdam formulated subprojects for collaboration. One such subproject was ChatIG, a collaboration between the Faculty of Arts and the Ignatius Gymnasium. This project had the objective to 1) let the pupils build a corpus of chat data; 2) make the corpus available for scientific research; and 3) formulate and answer research questions based on the corpus.

3.2 Participants

In 2004-2005 four classes of Ignatius pupils participated, two classes from grade 1 (age 12/13) and two classes from grade 3 (age 14/15). In 2005-2006 three classes participated, all 3rd grade students (age 14/15). In total 188 pupils participated. The pupils were supervised by their own Dutch language teachers as well as both authors. Technical assistance was provided by the university.

3.3 The chat experiment

The pupils chatted with each other through the chat program within Blackboard, an online environment used in various schools and universities for handing in papers, storing files, emailing, having forum discussions, etc. During the chat experiment, the pupils participated in seven short chat sessions of five minutes each. The first session was a practice session to get used to the chat environment. The other six sessions were devoted to various topics, two of which were deemed 'involved' (i.e., it was expected that the pupils would experience involvement with the topic, e.g. the MTV awards which were recently aired on television), two of which were expected to be 'non-involved' (e.g. the election of a new pope), and two sessions were free choice topics. The experiment was also set up in such a way that each participant took part at least once in a session with two, three and four pupils. There was no moderator present and students did not have the option to integrate other types of media into the chat sessions (i.e. whiteboards, material on external platforms, movies, etc.).

3.4 Metadata

Apart from partaking in various chat sessions, all pupils filled in a questionnaire, which allowed the authors to collect metadata on the pupils' gender, background of their parents, languages spoken, computer and chat usage, circle of close friends, etcetera. Both pupils and parents filled in a consent form allowing the authors to use the collected materials, both the chat output and the questionnaires, for scientific research.

3.5 The final corpus

Once the pupils had participated in the chat sessions, we had a large collection of chat interactions as archived by the Blackboard chat system. No messages were adapted or deleted. All the sessions were exported to the NoteTab Light editor and the data were cleaned up manually.

We decided that time tags after each chat contribution would not help analyses since the system was quite slow. We therefore removed the time tags, with the exception of the time tags that showed when individuals entered and left the chat session. Furthermore we removed all irrelevant information related to archiving the sessions. We also replaced the pupil information at the beginning of each submission with a more detailed line of information. In the final corpus this information has been replaced with the following string of information: 'unique pupil ID_male or female + schoolyear_pupil ID throughout chat sessions.' A new ID will look like this: '64_m1_20', meaning that this pupil has received 64 as their unique ID in the Meta Database, the pupil is male, a first grader, and used number 20 throughout the chat sessions.

The total corpus has been incorporated in the SoNaR corpus (Oostdijk et. al. 2013) in order to make it accessible for linguistic research. Here the corpus has been standardized and annotated using the SoNaR standard (cf. Sanders 2012 for details and references). The size of the ChatIG corpus is 83,806 tokens (Sanders 2012). The corpus is available through VU University or via TST Centrale (http://tst-centrale.org).

4 Challenges

4.1 Challenge 1: Finding chat participants

Since chat communication generally takes place in a private setting (whether this is in a chat room or through a chat messenger program), we required participants who would chat in similar, private conditions. At the same time we wanted to store the chat logs. Rather than visiting people in their homes while chatting or asking people to store their own private chat logs from their own computers we decided to control the chat sessions and set up a chat experiment. This way we could also be sure that the chat data were unedited. Since chat communication in 2004 was most popular amongst Dutch youth (0-24 year olds), we wanted to have access to this audience. A school seemed like the best way to reach Dutch teenagers. We found a school to collaborate with through the "Happy School" project.

Creating the chat corpus therefore depended very much on the collaboration with this Amsterdam grammar school. On the one hand this allowed us to get a unique set of data, on the other hand this created a number of practical issues that are difficult to deal with and may well hinder the creation of a large scale corpus of this type. First of all, it is imperative that the pupils are capable of making their contribution. In our case this required that appointments were made with the teachers of the various classes, so that the pupils could come to the university and chat in our computer rooms where we had the required hardware and software. In order to make this work this required very precise arrangements with the school, who are ultimately responsible for the pupils (for example, the pupils' use of the tram, checking the presence of all of the pupils, etcetera).

The benefits of working together with the school are that the chat sessions were part of the regular school program. Pupils were therefore required to attend and participate. Furthermore, the pupils were always supervised by their own teacher and a lot of the practical matters were dealt with by the teacher. However, even though the chat experiment was made part of the regular school lessons, one cannot make partaking in a university experiment a requirement for passing a Dutch class at grammar school. One cannot force anyone to partake in an experiment, especially not minors. It is therefore essential to have a good relationship with the partner school and to have consent from all relevant parties.

4.2 Challenge 2: Privacy and consent issues

The pupils' chat contributions needed to be available to the scientific community. In principle there are two ways to make these data available. The first is to use the opt-out strategy that is used for example by Google Books: the data are available unless participants explicitly request that their data be withdrawn from the corpus. We consider this option unethical, in that we feel that participants should be aware of the fact that their data are the object of research, not post hoc, but before the fact. The second option is to ask the participants beforehand for consent to participate in the research. Lewis (2002) distinguishes between consent (there is formal approval for participation) and assent (the participant is willing to participate). In the case of adult participants these two forms of approval usually coincide. In the case of pupils from the Gymnasium age this requires approval from the caretaker (usually the parent). That is why we asked both parents and pupils to sign an informed-consent form beforehand.

Informed consent implies that participants know beforehand that their data will be recorded. This is what we believe to be the most ethically responsible way to collect research data. Although this option is ethically transparent, it does cause participants to possibly behave differently than they would under 'natural' circumstances. Our corpus data show that the participants are very much aware that their data are being read by the researchers.

(1) 6 leerling: alles wordt gescreendt
6 pupil: everything is being screened

At the same time, the students treat this fact as a joke:

(2) 15 leerling: zouden ze al deze gesprekken kunnen nalezen
14 leerling: ja
11 leerling: Tuurlijk
15 leerling: shit
11 leerling: Ach
15 leerling: pas op met wat je zegt
15 leerling: !!!
11 leerling: Whahaha

15 pupil: will they be able to re-read all these conversations?
14 pupil: yes

11 pupil: Of course
15 pupil: shit
11 pupil: Owh
15 pupil: be careful with what you say
15 pupil: !!!
11 pupil: whahaha

Just like participants who are being recorded with a tape recorder forget within a few minutes that the tape recorder is actually present in the room (ten Have and Komter 1982), these pupils also quickly forget that their chat conversations are being recorded. This becomes apparent when pupils are talking about smoking drugs, a topic that will most likely not be appreciated by their school teachers. Even when one of the pupils reminds the others that the conversation is being recorded, the pupils continue talking about drugs.

(3) 18 leerling: pam[1] heeft kk veel geblowt dit weekend ehh
17 leerling: haha
18 leerling: gwn bij der huis, kapot gek
17 leerling: ja?
16 leerling: ehm.. dit wordt opgenome he.. dat jullie da ff wete\
17 leerling: heb ik ook wel is gedaan
18 leerling: ja ze is para tog
17 leerling: ma toen waren mn ouders weekend weg

*18 pupil: pam [name] f*ing smoked up a lot this weekend ehh*
17 pupil: haha
18 pupil: jst at her house, mad crazy
17 pupil: yeah?
16 pupil: ehm.. this will be recorded.. just so you guys know
17 pupil: I've done that before
18 pupil: yeah, she's para right
17 pupil: but my parents were away for the weekend

This example also shows that we need to be very careful with anonymizing the data. Pupils at this age cannot be held accountable for their actions or language use at a later stage in life. Although they do all sign a form which states that their data can be used for research, they will most likely not be aware of the implications of their talk about drugs in future circumstances. It is for this reason that the data were anonymized in all publications. Last names were never recorded or required.

4.3 Challenge 3: Technological challenges

For our data collection we made use of the Blackboard 6.2 and later 6.5.1 (full participation mode) chat system. This text-based tool is especially designed for live, synchronous interaction. This provided the opportunity to archive the data and to prevent the pupils from chat-

[1] Names in the examples have been changed.

ting with people outside of the chatroom. We also turned off the private-message setting so that students could not message each other privately, outside of the experimental setup. At the same time the use of this technology proved to be challenging. The interface was highly unfamiliar for the pupils, which we accommodated for by letting the students familiarize themselves with the system during the first five minute session. The interface was also relatively basic, compared to that of MicroSoft Network messaging service program (MSN), which was the dominating chat program at the time of the data collection. For example, the pupils could not make graphical smileys and the like in the way they were used to.

Another technical issue had to do with the sizes of the groups of pupils coming to the university. As our computer rooms cater for 18 participants the groups had to be divided over two rooms, on different floors. The synchronization between the two classrooms was not optimal during the first chat session. This meant that some pupils were waiting for their chat partner to arrive in the chat room. Especially when the pupil only had one chat partner, this meant that pupils were annoyed or waiting for too long, missing out on minutes in which they could have been chatting.

The choice for using the Blackboard system provided us with maximal control over the situation: pupils could chat only with fellow pupils of a known age and background who had also consented to participate in the data collection. Drawbacks to this approach were already mentioned above (unfamiliar interface, lack of speed, etcetera). In a later chat data collection for the benefit of the SoNaR project a dedicated tool was used to collect chat data from similar age groups (cf. Sanders 2012 on the Bonhoeffer data). In general such a dedicated tool seems preferable, but of course it implies the availability of the relevant technology.

4.4 Challenge 4: Imitating a natural chat situation

By choosing to have pupils take part in a controlled chat experiment, we also chose to work in an unnatural chat situation. The university computer facilities as well as the number of supervisors involved in the experiment allowed us to use two different classrooms. However, pupils would sometimes chat with people in the other classroom and sometimes with people in the same classroom.

In order to maintain the experimental set up but at the same time imitate a home situation, individual cabinets might offer a better solution. At the same time, it will remain difficult to resemble the home situation (van Charldorp 2005). In a home situation pupils might be listening to the radio, watching TV, be involved in a face-to-face conversation or otherwise be multitasking. Nowadays, pupils may be chatting or apping on their mobile phones and may even be on the road, at school or in the bus while chatting online with friends.

The speed of the internet connection and the chat program also contributed to the unnaturalness of the chat situation. At the time of creating this corpus, pupils were used to MSN which operates not only speedier, but also has a different interface. The speed of the Blackboard chat system was a frequently discussed topic:

(4) 12 leerling: sssssssssssssssssslllllllllllllllllllllllllooooooooooooooooooooo-
oooooooooooooooooooooooooooooooooooooooooooommmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmm

12 student: sssssssssssssssssslllllllllllllllllllllllllllooooooooooooooooooooooooooo-
oooooooooooooooooooooooooooooooooowwwwwwwwwwwwwwwwwwwwww-
wwwwwwwwwwww

The Blackboard chat system is also specifically compared to MSN in the example below:

(5) 11 leerling: dit is vaag man, en sloom
17 leerling: isset leuk daaro?
11 leerling: jaah, man
17 leerling: jah kweet
17 leerling: msn is btr
11 leerling: waatttuh\, neej, tis saai maar beter dan sgool
17 leerling: dat zei ik dus egt een ur gelee
17 leerling: ja keej
17 leerling: maar kan neit tege sloomheid
11 leerling: jaaah,
17 leerling: van de comp

11 student: this is vague man, and slow
17 student: isit fun o there?
11 student: yeah, man
17 student: yah I know
17 student: msn is btr
11 student: whaaaat\, naah, its boring but better than scool
17 student: that's what I said like an hour ago
17 student: ya ok
17 student: but I can't take slowness
11 student: yeaaah,
17 student: of the comp

In the MSN chat program you can see who you are chatting with. Since we decided to anonymize all pupils by assigning numbers (pupil 1, etc.), pupils' first question would usually be: who are you?

(6) 16 leerling: who r u???
16 pupil: who r u???

Or, in the example below where there are more than two chat participants, pupil 10 asks the others present who pupil 16 is.

(7) 10 leerling: wies 16
10 leerling: ?????????

10 pupil: whos 16
10 pupil: ?????????

These examples show that identity is generally known beforehand when chatting with each other. This is an important issue given that identity management may crucially determine chat communication (Becker and Stamp 2005).

Soon the pupils also started playing with this anonymity issue. Pupils could for example pretend to be someone else, like in the example below:

(8) 1 leerling: he ik ben ook sacha
7 leerling: das raar

1 pupil: hey I'm also sacha
7 pupil: thats weird

This type of activity then shows that pupils quickly become aware of the technological restrictions but also opportunities that it provides for interaction.

A similar activity occurred when pupils realized that emoticons did not work the same way in the Blackboard chat program as they did in the chat programs they were used to at home.

(9) 23 leerling: wij zijn annaaa(8)
19 leerling: anna is sexy naam (HAHAHAHA)
23 leerling: oh.. geen emoticons =_=

23 pupil: we are annaaa(8)
19 pupil: anna is sexy name (HAHAHAHA)
23 pupil: oh.. no emoticons =_=

In the first line, pupil 23 uses an emoticon: (8), which should create an emoticon with sunglasses. However, the Blackboard system does not create emoticons based on characters, and thus the pupils just see the (8) on their screen. The pupil realizes that her emoticon did not work and shares her disappointment: "oh.. no emoticons" and adds the characters that create a 'bored face' emoticon. Again, the use of characters to display an emoticon, while knowing that the emoticon will not work in this chat program, shows the ability of the pupils to adapt their language use to the technological restrictions. However, students still complained about the lack of emoticons and buzzers, not only in the evaluation of the chat sessions afterwards, but also to each other in the chat sessions themselves:

(10) 8 leerling: echt rot dat hier gteen emoticons zijn en buzxzers die zijn het leukst \
8 pupil: it sucks that there arbent any emoticons and buzxzers they are the best \

Although the natural chat situation was not exactly replicated, the data show other interesting language activities that provide insight into how pupils deal with technological restric-

tions or differences. As such the data provide extremely interesting information. It remains to be seen to what extent our data resemble those of naturally occurring chat data.

There was one last aspect of our experimental setup that specifically deviates from the natural situation, which were the topics that students had to chat about. For sociolinguistic purposes we wanted to see if students chatted differently when talking about an involved or a non-involved topic. We furthermore wanted to have a comparable dataset in which pupils were allowed to talk about anything at all. During the evaluation afterwards it appeared that students often did not stick to the given topics. Therefore we believe that having such topics did not specifically add to the unnaturalness of the situation:

(11) 12 leerling: we hebben het niet echt over tmf awards maar kan mij het schelen
12 student: we didn't really talk about tmf awards but I don't care

For some research communities (e.g., conversation analysts) collecting data in an experimental situation and under suboptimal technical circumstances affects the validity and hence the usefulness of the collected materials. Although we have no emprical evididence yet to support the claim, we believe that many of the phenomena that linguists look for in these type of CMC data are represented in the corpus and hence the corpus will be of use for many researchers.

4.5 Challenge 5: Ethical dilemmas

As can be expected amongst teenagers, not all pupils get along with each other. One of the classes even dealt with regular bullying which had been discussed at school with the students and parents. The fact that the pupils realize that they are being monitored does not prevent them from bullying while chatting:

(12) 3 leerling: Leerling 4 is een lul:P
[...]
3 leerling: het stinkt naar lleeerling 4
[...]
3 leerling: miriam ruitk vies
[...]
3 leerling: jaah leerling 4 g0re l*ul

3 pupil: Pupil 4 is a dick:P
[…]
3 pupil: It smells of ppupil 4
[...]
3 pupil: miriam stinks
[...]
*3 pupil: yeah pupil 4 dIrty d*ick*

Pupils dare to say a lot of things in the chat room, perhaps more so than face to face in the classroom, or when writing papers. Especially the boys-only conversations are filled with

slang and curses. There are entire sessions where boys just curse at each other and use curse abbreviations. Pupils like to make fun of each other but they also make fun of (or bully) pupils who are not partaking in the session. Even though the pupils are aware that the conversations are being recorded and that their input is being used for research, they do not seem to mind using vulgar language or talking about taking drugs or crime related topics, as in example (3).

Such language use creates ethical issues. Researchers saw these data but also the pupils' own teachers. Even though the data are anonymized, should we not protect the pupils by excluding these materials from the corpus that in principle is expected to last for a long period of time? We decided to leave all of the materials in the corpus, as the pupils and their parents had consented to the data collection, but we realize that such a choice is up for debate. We are interested in how other corpus analysts look upon this issue.

4.6 Challenge 6: Understanding in-group language

The results of the questionnaire show that 77% of first graders and 88% of third graders believe chat language to be like spoken language (vs written language or both). Although this remains an unanswered and much discussed question for linguists (see the introduction), pupils themselves seem to have a more unified view on this matter. Knowing that pupils themselves feel like their language use resembles spoken language when chatting, might make understanding the data a little bit easier. For example, in the data we see contributions such as written out animal sounds:

(13) 26 leerling: Kukelekuuuuuuuuuuuuh
26 pupil: Cock-a-doodle-doooooooo

instrument sounds:

(14) 22 leerling: toeteretoeteretoet
22 pupil: tooootooootootoot (trumpet sound)

extreme use of interpunction symbols:

(15) 10 leerling: ?????????
10 pupil: ?????????

phonetic spelling:

(16) 27 leerling: nahja maar hoezo ik wist egt neit dattiej bij soon club ofzo zat…
27 pupil: nooow but why i relly didt know thathej was part of such a club or smth...

Furthermore, pupils quote songs, chat in different languages (English, Italian, German), frequently misspell words, shorten words, use abbreviations and create creative language (also see van Charldorp 2006). When cleaning up the data we learned one important lesson:

do not throw away data that look unfamiliar. The data proved to be enormously rich. It provides a great insight into the creative use of language by Dutch teenagers in a CMC environment that can be used for a great variety of research topics. At the same time, this creativity creates the challenge of how to normalize and standardize the spelling variants to make the data searchable (cf. Oostdijk and van Halteren 2013 for a similar issue in Twitter).

4.7 Challenge 7: Sustainability of the corpus

In developing our corpus its sustainability proved to be a challenge in several respects. Firstly, there is the issue of the rapid technological developments, which can make systems and even complete genres obsolete almost overnight. MSN is a good example: while being the dominant chat system in the period of our data collection, presently it has disappeared and has been succeeded by systems with different technological affordances like Whatsapp, Facebook chat and Twitter. In a sense this means that corpus linguists interested in the relationship between language use and medium should be aware of their role as archeologists of language, before the phenomenon of interest has disappeared.

A second type of sustainability involves the extension of the materials. The creation of the ChatIG corpus depended completely on the cooperation with the Ignatius Gymnasium Amsterdam, funding from the VU University, and our contacts with enthusiastic teachers. That makes data collection also a vulnerable type of operation. After funding stopped and one of the teachers involved took a different position, our data collection project was discontinued. Fortunately, researchers from the Language and Speech Technology Group at Radboud University Nijmegen have set up a chatbox to collect data and have thus added to the collection of chat data in SoNaR considerably.

A third type of sustainability is related to making the data accessible for linguistic research. To that end our data were transferred to the Language and Speech Technology Group at RU Nijmegen, which has standardized and annotated the data following the SoNaR standard (cf. Sanders 2012 for details and references).

5 Conclusion and discussion

CMC data contain a wealth of information for corpus linguists, and hence it is imperative that we collect such data before the technology by which they were produced becomes obsolete. In this paper we described the challenges we encountered in our collection of chat data in the ChatIG project. We have shown that access to the right group of people to produce the data can be difficult, and depends very much on the social network of the researcher. Especially when we are dealing with a specific target group such as secondary school pupils, data collection creates issues of privacy and consent. We firmly believe that those issues should not be taken lightly and that only a full consent (or, more specifically, consent and assent) of the contributors of the corpus is ethically acceptable.

Other ethical issues occur when contributors display socially unacceptable behavior during the chat sessions. As researchers we should realize that a vulnerable group like secondary school pupils may not be aware of the consequences when displaying such behavior in a data collection project that intends to generate sustainable data for future research; should we protect the students and eliminate the data from the corpus? But that would eradicate a

type of language behavior that might very well be typical for the type of language use under investigation and hence affect the validity and usability of the resulting corpus. As mentioned above, we decided to leave these data in the corpus, thus prioritizing the validity of the materials. We welcome a debate about this issue.

Data collection also creates linguistic and logistic challenges. How do we know for sure that we understand the speech-like utterances that are full of slang and language games? Of course, this issue is not unique for collecting CMC data (viz. the analysis of so-called *straattaal* ('street language') by Appel 1999 and Nortier 2001), but given the speech-like character of CMC language use by this age group, we may well need an anthropological take on learning to understand this type of in-group language. Logistically, we need to devote attention to a timely collection of the data, before the technology is out of date, in such a way that it leads to a substantial amount of data that is suitable for corpus linguistic analysis, and hence links on to established formats. That is why we believe that our case study has important implications for the collection of data from newer technologies like Whatsapp: we urgently need to collect sufficient amounts of those data and store them in a standardized format before it is too late.

References

Appel, R. (1999). Straattaal; De mengtaal van jongeren in Amsterdam. *Toegepaste taalwetenschap in artikelen*, *62*, 39-55.

Baron, N. (2008). *Always on: Language in an online and mobile world*. Oxford etc.: Oxford University Press.

Becker, J.A.H. and Stamp, G.H. (2005). Impression management in chat rooms: A Grounded Theory Model. *Communication Studies, 56*(3), 243-260, DOI: 10.1080/10510970500181264

Beißwenger, M. and Storrer, A. (2008). Corpora of computer-mediated communication. In: A. Lüdeling and M. Kytö (Eds), *Corpus Linguistics. An International Handbook. Volume 1*. (pp. 292-308). Berlin / New York: De Gruyter.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. and Storrer, A. (2013). *DeRiK*: A German reference corpus of computer-mediated communication. *Literary and linguistic computing, 28*(4), 531-537.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In: D. Tannen (Ed.), *Spoken and Written Language: Exploring Orality and Literacy* (pp. 35-54). Norwood, NJ: Ablex.

Charldorp, T.C. van (2005). Building a chat corpus. Unpublished manuscript, available from http://www2.let.vu.nl/documenten/corpora/Building_a_Chat_Corpus_2005.pdf

Charldorp, T.C. van (2006). Dutch teenage chat communication: a sociolinguistic perspective. Unpublished thesis VU University Amsterdam.

Gries, S.T. and Newman, J. (2013). Creating and using corpora. In Robert J. Podesva and Devyani Sharma (eds.), *Research methods in linguistics* (pp. 257-287). Cambridge: Cambridge University Press.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. A. Barab, R. Kling, and J. H. Gray (Eds.), *Designing for Virtual Communities in the*

Service of Learning (pp. 338-376). New York: Cambridge University Press. Preprint: http://ella.slis.indiana.edu/~herring/cmda.pdf.

Herring, S. C. (2013). Discourse in Web 2.0: Familiar, reconfigured, and emergent. In D. Tannen and A. M. Tester (Eds.), *Georgetown University Round Table on Languages and Linguistics 2011: Discourse 2.0: Language and new media* (pp. 1-25). Washington, DC: Georgetown University Press. Prepublication version: http://ella.slis.indiana.edu/~herring/GURT.2011.prepub.pdf

Kestemont, M., Peersman, C., de Decker, B., de Pauw, G., Luyckx, K., Morante, R., Vaassen, F., van den Loo, J. and Daelemans, W., (2011). The Netlog Corpus. A Resource for the Study of Flemish Dutch Internet Language, in: *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 23-25). Istanbul: European Language Resources Association (ELRA).

King, B. (2009). Building and analysing corpora of computer-mediated communication. In: Baker, P. (ed.), *Contemporary Corpus Linguistics* (pp. 301-320). London: Continuum.

Lewis, A. (2002). Accessing, through research interviews, the views of children with difficulties in learning. *Support for Learning, 17*(3), 1467-9604. DOI: 10.1111/1467-9604.00248.

Nortier, J. M. (2001). *Murks en straattaal: Vriendschap en taalgebruik onder jongeren.* Amsterdam: Prometheus.

Oostdijk, N.H.J. and Halteren, H. van (2013). Shallow parsing for recognizing threats in Dutch tweets. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* (pp. 1034-1041). New York: IEEE. http://dx.doi.org/10.1145/2492517.2500271

Oostdijk, N.H.J., Reynaert, R., Schuurman, I. and Hoste, V. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns and J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Theory and Applications of Natural Language Processing* (pp. 219-247). Berlin: Springer.

Sanders, E. (2012). Collecting and analysing chats and tweets in SoNaR. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani et al. (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: European Language Resources Association (ELRA). Available from http://www.lrec-conf.org/proceedings/lrec2012/pdf/416_Paper.pdf

Sanders, T.J.M. and Spooren, W.P.M.S. (2013). Exceptions to rules: a qualitative analysis of backward causal connectives in Dutch naturalistic discourse. *Text & Talk, 33*(3), 399-420.

Tagg, C. (2009). A corpus linguistic study of SMS text messaging. Unpublished PhD thesis.

ten Have, P. and Komter, M. (1982). De angst voor de tape. Over bezwaren tegen het gebruik van de bandrecorder voor onderzoek. In C. Bouw, F. Bovenkerk, K. Bruin, and L. Brunt (Eds.), Hoe weet je dat? Wegen van sociaal onderzoek (pp. 228-242). Amsterdam: Uitgeverij de Arbeidspers & Wetenschappelijke Uitgeverij.

Hans van Halteren, Nelleke Oostdijk

Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens

Abstract

In this paper, we attempt to estimate which proportion of the word tokens in Dutch tweets are not covered by standard resources and can therefore be expected to cause problems for standard NLP applications. We fully annotated and analysed a small pilot corpus. We also used the corpus to calibrate automatic estimation procedures for proportions of non-word tokens and of out-of-vocabulary words, after which we applied these procedures to about 2 billion Dutch tweets. We find that the proportion of possibly problematic tokens is so high (e.g. an estimate of 15% of the words being problematic in the full tweet collection, and the annotated sample with death-threat-related tweets showing problematic words in three out of four tweets) that any NLP application designed/created for standard Dutch can be expected to be seriously hampered in its processing. We suggest a few approaches to alleviate the problem, but none of them will solve the problem completely.

1 Introduction

With the advent of the social media, communication has changed drastically. Where before the public mass media (radio, television, newspapers) were dominated by a relatively small group of communication professionals, the social media have provided a platform for the masses through which they can voice their experiences and opinions and interact with other users. Nowadays the volumes of user-generated content that are produced on a day to day basis exceed by far whatever expectations existed when the first services were launched.[1] Access to the social media is unrestricted and generally services are widely and freely available. Communication is fast and as such extremely suitable for spontaneous, almost instant communication.

User-generated data have attracted the attention not only of linguists and communication experts, but also businesses, governmental and non-governmental organizations. The data are being exploited for a wide range of purposes, from linguistics and communication research to developing marketing strategies or evaluating policy issues. Linguists and communication experts have taken an interest particularly in the conversational data that are available from chats and discussions lists. Where the chat data are typically real time (private) conversations between a small number of people, with discussion lists the communication is usually more public but slower, and therefore often also more edited. In recent years, the interest for user-generated data has received an immense boost as Twitter was adopted by the masses. Twitter combines more instant communication and public availability which make it a very valuable source also for information mining.

One of the problems with user-generated data is that users take the liberty of expressing themselves as they see fit, without necessarily adhering to spelling conventions, grammati-

[1] In 2007, the fledgling Twitter boasted a meagre 5,000 tweets per day (Weil 2010). In 2013, this has grown to 500 million tweets per day (Krikorian 2013).

cal rules, etc. As a result we find that texts display a great deal of variability as regards typography, orthography, syntax, semantics, and discourse. The variability has several dimensions. The variability may be

- medium-related, that is, the medium may impose limitations on the length of the texts, while authors may also experience that they are under pressure of time (e.g. in chats) as their interlocutors may claim their turn prematurely while they are yet to finish. Moreover, there may be an effect of the (im)possibilities of the text entry mechanism used.
- author-related, that is, the language use of each author is characterized by its own idiosyncracies
- use-related, that is, depending on whether texts are used for professional or social use they adhere more or less to more widely accepted conventions or standards. Thus it appears that news feeds and government communications are quite conventional in the language they use, whereas texts exchanged between pals especially by sms or whatsapp but also on Twitter may be almost incomprehensible to people outside their peer group.

Clearly, it would be naive to think that all deviations from the norm set by the language used in the conventional media are errors. Some obviously are, but in many other cases the author made a deliberate choice to use some variant form.[2] As a result, we find that processing user-generated data is severely hampered as standard tools such as tokenizers, part-of-speech taggers, lemmatizers, morphological and syntactic parsers, and named entity recognizers cannot handle the variability very well. Han and Baldwin (2011: 369) report that they "found Twitter data to have an unsurprisingly long tail of OOV [out-of-vocabulary; HvH/NO] words, suggesting that conventional supervised learning will not perform well due to data sparsity. Additionally, many ill-formed words are ambiguous, and require context to disambiguate." Thus the variability is found to be prohibitive when it comes to successfully using applications such as text-to-speech systems (for example when wanting to have a text message read aloud), search and retrieval systems, and machine translation systems. In previous research we have found that in the n-gram based recognition of threatening tweets modeling spelling variation alone increased recall with a further 2.7-5.8% (Oostdijk and van Halteren 2013).

In this paper the research question we attempt to answer is: What proportion of the word tokens occurring in Dutch tweets is not covered in lexical resources designed/created for standard Dutch, either because tokens are not included at all or because tokens are (unrelated) homographs of the tokens listed there.[3] Related questions here are (a) is it possible to estimate the proportion automatically on the basis of a sample? and (b) to what extent does the proportion vary between various authors and topics? The question underlying our main

[2] See also the comprehensive study by Tagg (2009) who on the basis of a corpus of sms texts investigates the many different strategies that people use when the medium imposes severe restrictions on the length of a text and text entry is hindered by the entry device.

[3] Examples include clitics, such as *dak* (normally "roof") for *dat ik* ("that I") or *int* (normally "collects") for *in het* ("in the"), and the abbreviation *eik* (normally "oak") for *eindelijk* ("finally").

research question and the motivation for undertaking the research described here is that we want to know whether or not, and if so, to what extent the variability in Dutch tweets hinders automatic processing

Starting-point for our approach is the collection and manual annotation of a pilot corpus of tweets in which a small number of specific hashtags are represented. The pilot corpus is annotated for different types of token, both words and non-words, after which the annotated corpus can be used for a detailed investigation of the variability at this scale. Furthermore, we use the annotation as a benchmark for an automatic estimation procedure with which much larger amounts of tweets can be processed. After conformation of the validity, we apply the estimation procedure to a large – almost 2 billion tweets – collection of Dutch Tweets.

The structure of the paper is as follows. First, in Section 2 we introduce the pilot corpus that constitutes the experimental material that we use as a basis for obtaining estimates of the proportion of problematic tokens. Next, we describe the annotation of experimental material (Section 3). In Section 4, we discuss our findings as regards the different types of word and non-word tokens in the manually annotated data. The automatic estimation of the proportion of problematic cases is the topic of Section 5, while the estimates for the whole collection of tweets considered in this paper are given in Section 6 together with an analysis of our findings. We conclude this paper with a brief summary and our plans for future work.

2 Selection of experimental material

Most work on out-of-vocabulary words in tweets has been done on the basis of type lists (e.g. Han and Baldwin 2011; Sidarenka et al. 2013). However, such lists lack vital information. They do not show us how the words are distributed over the tweets, so that we cannot estimate which percentage of the tweets is affected, or whether there are differences between users and/or topics. They also do not show the context, so that we cannot know whether an in-vocabulary word is in fact known or whether the form is merely a homograph of another word in a lexicon used by some NLP application. For a proper investigation, we will have to look not at type lists, but at the underlying tweets. We will start with a modestly sized pilot corpus, so that a full manual annotation is feasible. For this corpus, we have selected ten hashtags, intended to provide a reasonable spread in topics and language use. Obviously, no exact predictions can be made about the language use for any topic, as even the most professional topics will occasionally attract emotional or humorous comments.

- **#aardbevingen (A)**[4] Tweets carrying this hashtag discuss earthquakes in the province of Groningen, that are caused by extracting gas from below the surface. The tweets are expected to be mostly official communiqués and attacks by interest groups, and therefore rather clean language.
- **#doodsbedreiging (D)** Tweets with this hashtag contain (death) threats and reactions to them. They tend to be very emotional and regularly contain street language.
- **#file (F)** These are tweets concerning traffic jams, generally people reporting on new traffic jams and their reaction. The level of emotion is less high than might be expec-

[4] We will be using single letter abbreviations for the various hash tags in tables and figures.

ted, possibly because one is used to being in traffic jams. A special type of token here are the various names of cities and roads.

- **#houdoe (H)** This hashtag does not refer to any specific topic, as *houdoe* is a dialect word for goodbye. We included this hashtag in order to find uses of dialect in tweets.[5]
- **#irri (I)** This hashtag like #houdoe, #jaloers and #omg does not refer to a spefic topic. *irri* is a short form for Dutch *irritant* (English: irritating). Given this hashtag, we expect high emotion levels with concomitant effects on language use.
- **#jaloers (J)** With *jaloers* meaning 'jealous', we again expect some emotion, although less strong than with *#irri*.
- **#miljoenenjacht (M)** Tweets carrying this hashtag relate to a Dutch tv game show. These tweets are expected to be mostly from the average user rather than (semi-) professional authors.
- **#ns (N)** *NS* is short for *Nederlandse Spoorwegen* (Dutch Rail). Tweets with this hashtag will regularly contain official communiqués, but also quite a lot of train traveler reports and, sometimes vehement, complaints. A special type of token here is formed by the various names of train stations and routes.
- **#omg (O)** *OMG* is short for "Oh, my God". Here we expect quite emotional tweets.
- **#syrie (S)** Tweets with this hashtag discuss the situation in Syria. These tweets mostly contain reports and comments, and they regularly refer to and quote from foreign media. Although the tweets have been marked as Dutch, we find a surprising number of tweets here in a foreign language, mostly French.

For each of these hashtags, we randomly sampled the tweets in the data collection available from the Dutch eScience Centre (Tjong Kim Sang and van den Bosch 2013) with a date stamp from January 1, 2011 to June 30, 2013. Sampling for each hashtag continued until at least 1,000 word tokens (see below) contained in Dutch tweets were found.[6]

3 Annotation of experimental material

In order to get a more precise overview of the (word) tokens in tweets that are found to be problematic for standard tools, we have manually annotated all tweets in our pilot corpus. The text was first tokenized automatically, after which we marked all word and hash tokens in Dutch-language tweets which were either out of vocabulary with regard to the OpenTaal[7] word list or that did occur in the list but only because the variant form just happened to be a

[5] If we wanted to find a lot of dialect, we should have selected tweets from the province of Limburg. However, Limburg dialects are often closer to being another language entirely.

[6] With the annotators deciding whether tweets were in Dutch and how many word tokens were present..

[7] OpenTaal is a project directed by the Dutch Language Union which aims to make available for free (written) Dutch language resources for use in open source projects (e.g. OpenOffice.org). One of the resources is the OpenTaal word list that was compiled for use with for example spelling checkers and grammar checkers. The word list we used for the research described in this paper is version 2.10g. It includes some 350,000 word forms, including many frequently used abbreviations and common Dutch proper names. For more information see http://www.opentaal.org/opentaal.

homograph of an item in the list (e.g. the token *na* used as the preposition *naar* instead of the preposition *na*). Where necessary, we corrected the automatic tokenization.

We tokenized all text samples with our own specialized tokenizer for tweets.[8] Apart from normal tokens like words, numbers and dates, it is also able to recognize a wide variety of emoticons. The tokenizer is able to identify hashtags and Twitter user names to the extent that these conform to the conventions used in Twitter, i.e. the hash (#) resp. at (@) sign are followed by a series of letters, digits and underscores. URLs and email addresses are not completely covered. The tokenizer counts on clear markers for these, e.g. http, www or one of a number of domain names for URLs. Assuming that any sequence including periods is likely to be a URL proves unwise, given that spacing between normal words is often irregular. And actually checking the existence of a proposed URL is computationally infeasible for the amount of text we intend to process. On the current sample, the tokenizer performs adequately, except that it still misses a number of emoticons, e.g. *:$* and *o.o*. In addition, the samples include one missed URL, *dlvr.it/10hgmg*. Finally, for words and hashtags, the tokenizer assigns a classification INVOC (in-vocabulary) or OOV (out of vocabulary), by looking up the token, decapitalized and stripped of diacritics, in the abovementioned word list (similarly normalized).

When annotating, we also found several tweets which were written completely in a different language than Dutch. The eScience Twitter corpus was collected by searching for tweets with any of a number of probably Dutch words, after which a character n-gram language filter was applied (Tjong Kim Sang and van den Bosch 2013). For older sections of the corpus, only tweets clearly marked as Dutch are included. Later sections include more tweets, together with an indication of the language proposed by the filter. Where this is another language, such as French or English, or where this is UNKNOWN, we automatically exclude the tweet from our sample. However, there is also a marker *notdutch*, which we find for both foreign language tweets, e.g. English, and for Dutch language tweets containing multiple non-Dutch tokens. For the manual annotation, we remove tweets entirely or mainly written in a foreign language by hand.[9]

The tokenizer distinguishes between the following types of token:

- **<word>** A normal word, as can be expected to be found in a dictionary. Apart from letters, a word may include digits (e.g. *A4*) and punctuation (e.g. *dag-/nachtlicht*).
- **<rt>** The sequence RT, used in tweets to indicate a retweet.
- **<num>** A numerical token. This includes numbers, but also dates, times, phone numbers etc.
- **<hash>** A hashtag as prescribed by Twitter.
- **<@>** A Twitter user name which is addressed, marked as such, in the tweet. The name of the author of the tweet is not annotated.

[8] We intend to merge our tokenizer in the near future into the open source tokenizer Ucto (http://ilk.uvt.nl/ucto/).

[9] In the automatic estimation procedure, we will also exclude notdutch, leaving only tweets that are likely to be Dutch.

- **<url>** An included URL.
- **<eaddr>** An included email address.
- **<emo>** An emoticon which is built including symbols. Emoticons built completely from letters, e.g. *xd*, are tokenized as <word>.
- **<symb>** All other symbols or symbol sequences. Often, but not always, symbols are punctuation marks.

In addition there are a few minor types for rare special cases. We only annotated <word> and <hash>. In this paper, though, we will focus only on the tokens of type <word>.

For the annotation of the <word> and <hash> tokens we applied the following markers for OOV tokens and INVOC tokens where these were used in a non-standard manner:

- **Missing.** The token is a correctly spelled word, but proves to be absent from the OpenTaal word list. We subclassify these as neologisms (missing-neo) or traditionally known words (missing-trad). Many of the neologisms are related to the new media, e.g. *facebookaccount*, *smst*, *tweet*, and *appt*.
- **Diminutive.** The token is a correctly spelled diminutive form of an INVOC word, but the form is OOV, e.g. *drinkmaatje*, *zenuwtrekje*.
- **Compound.** The token is a correctly spelled compound, which is not present in the list, e.g. *verkeershel*, *hamsterwangen*, and *schildpaddennek*
- **Hyphenation.** The token is spelled correctly, except for the hyphenation which is incorrect, e.g. *leeg-halen* and *bom-aanslag*.
- **Complex.** The token contains punctuation for some special effect, e.g. *huis/leerwerk* which combines *huiswerk* and *leerwerk*.
- **Proper name.** The token forms, by itself or in combination with adjacent tokens, a proper name. We will discuss these separately below.
- **Spelling.** The token is spelled in a non-standard manner. This could be due to a typographical error (e.g. *funcitoneert* instead of *functioneert*), but could also be intentional (e.g. *regenbooog* instead of *regenboog*, in a tweet mimicking the lyrics of a song 'Vlieg met me mee naar de regenboog!'). We subclassify these into lexicalized spelling variants (spelling-lex; e.g. *me* instead of *mijn* for the first person possessive pronoun) and productive ones (spelling-prod; e.g. *zukkels* instead of *sukkels* or *wilt* rather than *wil*).
- **Abbreviation.** As spelling, except that the variant spelling is clearly meant to shorten the word. Again, we differentiate between lexicalized (abbreviation-lex; e.g. *tv* for *television* and *ff* for *even*) and productive (abbreviation-prod; e.g. *is* for *eens* and *gewn* for *gewoon*) instances.
- **Dialect.** The token is a dialectal form. Here we distinguish between out-of-vocabulary forms (dialect-oov; e.g. *houdoe* is a dialect word originating from Brabants, one of the dialects spoken in the south of the Netherlands) and forms which are confused with a word in the standard word list (dialect-conf; e.g. *ons* as dialectal possessive form where standard Dutch would have *onze*).
- **Street language.** As dialect, except that this is the "street dialect" rather than a regional dialect. Again we distinguish street_language-oov (e.g. *wollah* and

djoeken) and street_language-conf (e.g. *kantelen*, street language for 'to kill', where in standard Dutch its meaning is 'to turn over').

- **Foreign.** A word from another language, which we consider not to be lexicalized yet in Dutch, e.g. *party, ciao, jihad*.
- **Interjection.** The token is an interjection. These tokens are often variants of invocabulary interjections (e.g. *hahaha*), with sometimes extreme repetition to stress the degree of emotion.
- **Emoticon.** The token is a recognized short letter combination expressing some emotion, e.g. *xd*.
- **Formula.** A non-linguistic combination, often containing measurements, e.g. *1u30* for an hour and a half.
- **Part of multiword.** The token forms a multi-token expression together with one or more adjacent tokens, and at least one of the tokens in the expression is not present in the word list, e.g. *in feite*, where *feite* is absent from the list (both tokens are marked).
- **Clitic.** The token is a concatenated, and usually shortened, combination of a pronoun and a verb, e.g. *kzal* for *ik zal*.
- **Merge.** The token is another concatenated combination of words. We distinguish between instances where the concatenation is used to form a hashtag (merge-hash; e.g. *#zieligpersoon*) and other concatenations produced by the author (merge-aut; e.g. *ofzo* for *of zo*).
- **Split.** The token is the beginning of a sequence of tokens which together form a word. We subclassify as to the reason for splitting. Just as for merge, we see splits for reasons of hashtag formation (split-hash; e.g. *trein #storing* in which the word *treinstoring* has been split to be able to use the hashtag *#storing*) and other author produced splits (split-aut; e.g. *ex politici, stop gezet*). Only when it is clear that the split was not intended do we mark it as such (split-typo; e.g. *moete n* for *moeten*).[10]
- **Insplit.** The token is a follow-on of the preceding split-token.
- **Clipped.** The token has been clipped because the text was cut off by Twitter or by a retweeting author, e.g. *probl, werel, reizig*.
- **Unknown.** The annotators are unable to determine the intended form and meaning of the token and can therefore not assign it one of the above classes.

Proper names form a rather special class of tokens.[11] They are usually only partly covered in lexical resources and therefore one can expect that in any text a proportion of OOV words can be explained in terms of proper names. In an NLP context proper names are often handled not by relying on a lexicon, but by some heuristics or separate module dedicated to their identification. In Dutch as in many other languages, proper names can often be recognized

[10] The reason for the indication aut for merge and split is that there are also splits caused by tokenization errors (split-tok; e.g. *o. o* for the emoticon *o.o*); however, these are corrected during annotation and will therefore no longer be found in the frequency counts below. Note that merge-tok does not currently occur, but this could change if a future tokenizer attempts to recognize multi-token units.

[11] We include under this class also derived forms such as adjectives, e.g. *Turkse*.

because they are capitalized. In Twitter, unfortunately, capitalization is often not regular. As a result, the identification of proper names is even more problematic than with more conventional text types. In order to be able to estimate the size of the problem, we differentiate between proper names written with (cap) or without (decap) a starting capital letter. However, there is another complication. It may be that a proper name, stripped of case and diacritics, coincides with another word which is in the list, and is therefore confused with it, e.g. minister *Kamp* is not in the word list, but the common noun *kamp* is. All in all, we distinguish seven cases:

- **Oov-cap**. The stripped form is not in the stripped list and the form was capitalized in the tweet, e.g. *Kerry*, and might therefore be recognized as a proper name.
- **Oov-decap**. The stripped form is not in the stripped list and the form was not capitalized in the tweet, e.g. *kilkowski*, and would most likely be processed incorrectly.
- **Inlex-decap**. The stripped form occurs only in the stripped list as proper name, and it can therefore be recognized as such. However, the form is not capitalized, e.g. beatrix.
- **Inlex-capdia**. The stripped form occurs in the stripped list as proper name. The form is capitalized, but deviates in the use of diacritics and/or use of capitals with regard to the standard spelling, e.g. *SYRIE* instead of *Syrië*, and *PVDA* instead of *PvdA*.
- **Inlex-decapdia**. The stripped form occurs only in the stripped list as proper name. The form is not capitalized, and deviates in the use of diacritics and/or use of capitals with regard to the standard spelling, e.g. australie instead of Australië, and *ipod* instead of *iPod*.
- **Conf-cap**. The stripped form occurs (also) in the stripped list due to a non-proper-name word. However, the form is capitalized, e.g. *Ban* (in *Ban Ki-moon*), where *ban* ("ban") is in the list as a common noun.
- **Conf-decap**. The stripped form occurs (also) in the stripped list due to a non-proper-name word. This problem is aggravated by the form being written in the tweet without capitalization, e.g. *robben* (instead of *Robben*), where *robben* ("seals") is in the list as a common noun.

The annotation process is not yet completely streamlined. Currently, we have all tokenized samples in Excel, with the rows each pertaining to one token and the columns to the various fields of information. When annotating, we regularly switch between the original order, so that we have a good view of the context, and sorting on specific column combinations, so that we can consistently annotate specific types of tokens. If we would ever want to annotate more material, an instruction manual for the annotators would obviously be useful.

4 Estimates from the annotated material

The pilot corpus that we compiled, after tokenization, comprises 14,783 tokens. A breakdown of the tokens into the different types of non-words and words is shown in Table 1. The

spread in the total proportion of non-word tokens is sizable. #doodsbedreiging has the highest with 35.6% non-word tokens. This is not surprising seeing the high number of retweets (mostly by people commenting on the tweet, either by an involved person to bluff back in the threat discussion or as an outsider to complain about the awfulness of this kind of tweet), each leading to an <rt>, sometimes a <symb> (the colon) and an <@>, plus often an <@> for the threatened person. At the low end, we find 20.2% for #irri, which is more surprising. Apparently, irritation is mostly uttered in words and is shouted to the world rather than addressed to specific people. In general, the differences we find are unexpected. The emotional groups (#irri, #jaloers and #omg), e.g., are very different, which suggests that each emotion has its own means of expression. The only grouping where we do find similarity is the two transport hashtags, #ns and #file. Here only the number of hashes shows a real difference; if we examine these, we see that #file has many more hashtags for the location and the reason for the traffic jam, where #ns tends to just list this information in the text. If we look at similar proportions of specific types of tokens, we often see unexpected rather than expected combinations. Take <symb>, where the highest numbers are seen for #doodsbedreiging (210 <symb>) and #aardbevingen (205 <symb>). Death threats and discussions of earthquakes are hardly comparable topics. On further examination, differences come out: in #doodsbedreiging the high symbol count is again due to the retweets which are expressed with a colon, whereas in #aardbevingen we see more proper punctuation than for most other hashtags and relatively many quotes. Just as for traditional text types, we can conclude that each domain has its peculiarities.

Table 1. Frequency and distribution of the different types of token in the various samples. The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

	A	D	F	H	I	J	M	N	O	S
<@>	56	122	32	12	18	101	18	26	43	47
<emo>	5	0	5	11	8	19	9	4	5	0
<hash>	181	110	220	249	127	149	159	138	173	142
<num>	17	18	27	14	7	18	28	33	16	14
<rt>	33	83	10	1	2	7	12	6	16	28
<symb>	205	210	162	114	92	160	141	192	163	190
<url>	26	9	17	3	0	4	4	12	9	19
<word>	1007	1003	1008	1002	1010	1016	1005	1001	1014	1007
Percentage non-word tokens/token	34.2	35.6	32.0	28.8	20.2	31.1	27.0	29.2	29.6	30.4

When we look at the proportion of OOV words in the different samples (Table 2), we find a better distinction between our preconceived groups. The emotional hashtags all show a high OOV proportion (#irri 10.8%, #jaloers 11.0% and #omg 10.9%). They are joined by #houdoe (11.1%) which is not necessarily emotional, but the familiar greeting does indicate a more personal involvement, so that more informal language should not be unexpected. Lower proportions are found for the more factive and/or newsrelated hashtags (#syrie 3.9%, #aardbevingen 4.3%, #ns 4.5% and #file 5.6%). That #miljoenenjacht (7.5%) is somewhere

in between is also to be expected, as part of the tweets reflect personal, sometimes emotional reactions to what is happening in the tv show. Only #doodsbedreiging with a rather low proportion of 9.0% is somewhat surprising, but of course this sample also includes tweets with less emotional commentary on the threats.

Table 2. Frequency and distribution over various samples of the problematic (PROBLEM) word tokens, i.e. the out-of-vocabulary (OOV) and of in-vocabulary word tokens that are used in an alternative way (INVOC-ALT), i.e. other than the use foreseen for their entry in the word list. The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

	A	D	F	H	I	J	M	N	O	S
OOV	43	90	56	111	109	112	75	45	110	39
Percentage OOV/word token	4.3	9.0	5.6	11.1	10.8	11.0	7.5	4.5	10.9	3.9
INVOC-ALT	34	51	40	58	34	41	27	49	46	17
Percentage INVOC-ALT/word token	3.4	5.1	4.0	5.8	3.4	4.0	2.7	4.9	4.5	1.7
PROBLEM	77	141	96	169	143	153	102	94	156	56
Percentage PROBLEM/word token	7.7	14.1	9.5	16.9	14.2	15.1	10.2	9.4	15.4	5.6

When we consider the proportion of tweets that contain one or more OOV word tokens, we also see (Table 3) a sizeable variance, from 31.5% for #syrie to 65.5% for #doodsbedreiging. Furthermore, the proportion of affected tweets is not necessarily correlated with the proportion of affected words. #houdoe, that shows the highest proportion of OOV words (11.1%) has an OOV tweet proportion of only 33.3%, the second lowest. This implies that #houdoe is a mix of reasonably clean tweets and tweets that contain a lot of OOV words.

Table 3. Proportion of problematic tweets in the various samples: overall (PROBLEM TWEET/tweet), containing OOV word tokens (OOV TWEET) or containing in-vocabulary words used in an alternative way (INVOC-TW). The single letter column headings represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

	A	D	F	H	I	J	M	N	O	S
Percentage OOV-TWEET/tweet	36.6	65.5	45.2	33.0	61.2	57.3	45.6	35.1	57.8	31.5
Percentage INVOC-TWEET/tweet	26.8	35.7	29.0	26.1	27.2	23.3	16.8	41.9	31.0	16.4
Percentage PROBLEM-TWEET/tweet	48.8	75.0	55.9	47.9	72.8	65.0	53.6	58.1	70.7	37.0

Continuing on to the in-vocabulary word tokens that are used in an alternative way to the one foreseen for their entry in the word list (INVOC-ALT), we see (Table 2 and Table 3) fairly high proportions (1.7% to 5.8% for the words and 16.4% to 41.9% for the affected tweets), especially considering that these tokens are hardly ever recognized as problematic because the focus is generally on OOV words. On average, there are about half as many INVOC-ALT words as there are OOV words, but the two proportions are not correlated (correlation factor of only 0.465).

If we look at both types of problematic words together, we see proportions ranging from 5.6% (#syrie) to 16.9% (#houdoe) at the word level (Table 2) and 37.0% (#syrie) to 75.0% (#doodsbedreiging) at the tweet level (Tabel 3). The most important finding of the annotation and its analysis may well be that tools that have to rely on their built in lexicon will have trouble with at least one in three tweets, for more extreme topics even three in four. We think this can justly be called a serious problem.

In order to judge how easy it might be to solve this problem, we need to look at the individual types of problematic words (see Tables A and B in the appendix for a detailed overview of the frequency and distribution of the various types of problematic word tokens). The easiest solution would be to add frequently occurring problematic words to the lexicon. These would mostly be the words that can be considered to be lexicalized (either in general or at least on Twitter), i.e. spelling-lex and abbreviation-lex, together with the often social-media-related neologisms (missing-neo). These three groups together make up about one third of the problematic cases. This means that this simple solution considerably alleviates the problem, but by no means solves it completely. A next partial solution might be to add pattern matching techniques to recognize specific classes of productive tokens. Emoticons are an example of this, but then we are currently considering only the words. Here, only (most of) the OOV interjections form such a class. Assuming we could recognize all interjections, this would resolve about one tenth of the problematic cases, not impressive but still worthwhile. A final substantial class of problematic words is formed by the proper names, about one fourth of the problematic cases. Here we would have to adapt existing named entity recognition techniques to the kind of text found on Twitter. This will certainly not be easy, as the two main information sources for NER are both corrupted. As for capitalization, only about half of the problematic proper name tokens are written with a capital letter. As for context, the system would have to be able to cope with spelling variation as well as deviant syntax in the surrounding text. As for other types of problematic cases, no easy solutions come to mind.

5 Automatic estimates of proportions of problematic tokens

The manual annotation of data for spelling variation is rather labour intensive. Obviously, if we want to investigate whether and how the proportion of spelling variants varies per user or per topic, and we need a sufficiently large amount of data to be able to draw conclusions, manual annotation is out of the question. We will therefore have to look to automatic means to estimate such proportions.[12]

So far we have considered two types of problematic cases. First, there are the out-of-vocabulary words. These are in principle relatively easy to find. We only need proper tokenization and then lexicon lookup. However, tokenization, as mentioned above, is not flawless. Furthermore, there is the problem of the presence of non-Dutch tweets in the material. Still, we will show below that we can automatically derive adequate estimates for OOV

[12] Note that, in this paper, we are not concerned with finding the actual problematic cases, but merely in estimating how many there are as a proportion of the total number of tokens, which means that a comparison at the level of an overall number is sufficient and that we do not have to use e.g. precision and recall to measure that we are finding the correct cases.

words. The in-vocabulary words with alternative uses, however, are much more problematic. Attempting to find them will involve language models, be it grammars or n-gram statistics, and although we are working on this (van Halteren and Oostdijk 2012), this work is not yet at a stage where we could sensibly attempt this task. Also, the frequency of these words is not strongly correlated with the frequency of the OOV words: looking at our ten samples, the two show a Pearson product-moment correlation of only about .46 Therefore, extrapolation from the OOV frequency is not possible. For now, we have decided to concentrate on estimating the proportion of OOV words.

There is yet a third source of possible problems for NLP applications. Such applications are designed to work with words and punctuation, possibly including numbers and abbreviations. All the other types of tokens that we encounter in tweets will severely hamper applications such as the already mentioned text-to-speech and translation systems. For this reason, we will also estimate the proportion of non-word tokens for various tweet types.

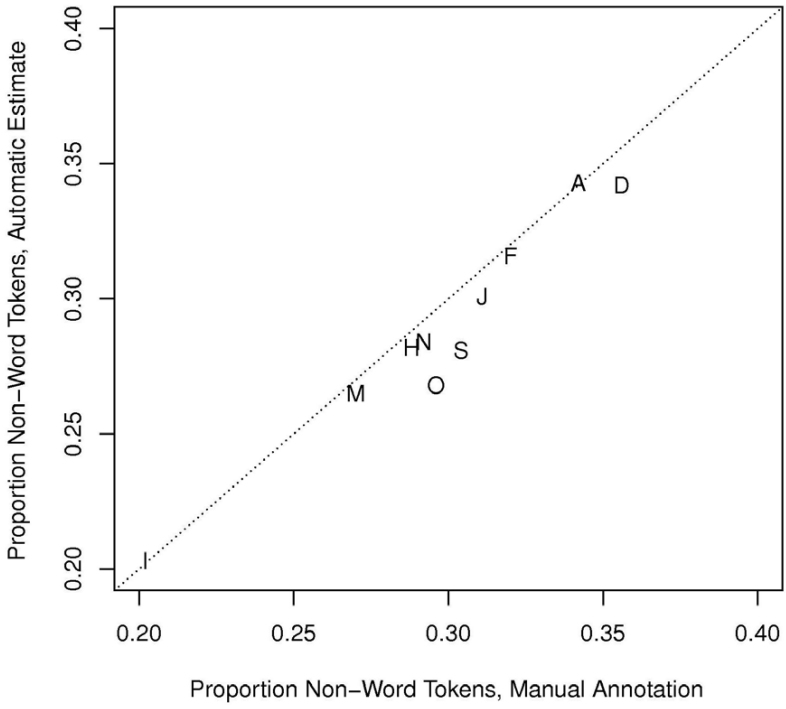

Figure 1. Benchmarking the automatic estimate of the proportion of non-word tokens on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

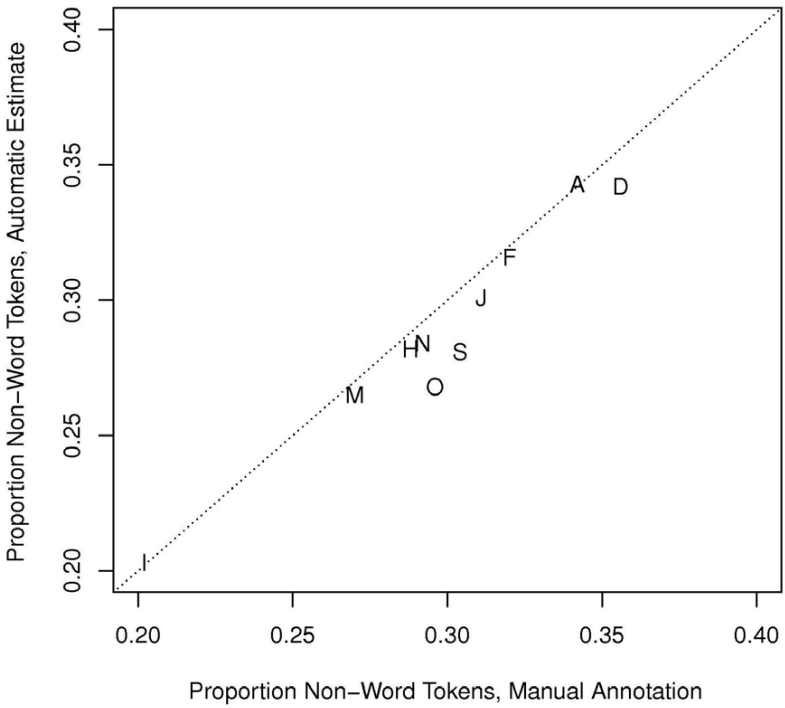

Figure 2. Benchmarking the automatic estimate of the proportion of non-word tokens on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

For both automatic estimates we will do just what we described above: tokenize and look up the tokens in the lexicon. We will ignore the fact that the tokenizer is known to make occasional mistakes. We do try to compensate for the presence of non-Dutch tweets in the material. All tweets marked either for another language or as UNKNOWN or notdutch are left out.

We tested the estimation process on the manually annotated material, starting from the raw rather than from the manually corrected version. Figures 1 to 3 show the comparison between the estimates derived from the manually annotated data and those derived automatically.

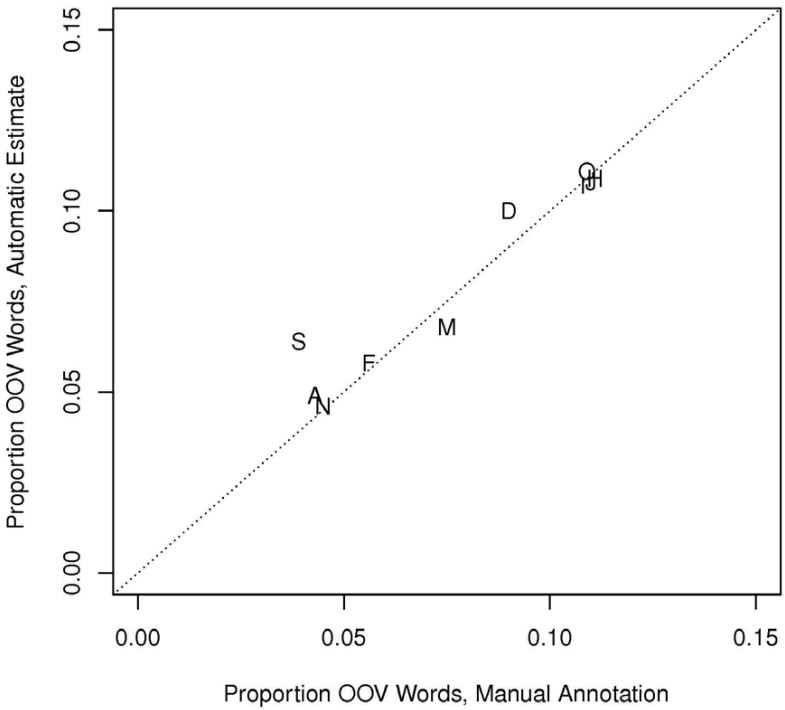

Figure 3. Benchmarking the automatic estimate of the proportion of OOV words on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

For the proportion of non-word tokens (Figure 1), the estimate is clearly adequate. This is confirmed by a correlation of .974 (confidence interval .890-.994). The same can be said for the proportion of OOV words within all words (Figure 2), although it has a slightly lower correlation of .959 (confidence interval .833-.991). The procedure is less effective when it tries to estimate which proportion of the tweets contain OOV words (Figure 3). Given the lower correlation of only .828 (confidence interval .415-.958), and with the errors concentrated in specific samples, we will refrain from using this estimate in the investigations below.

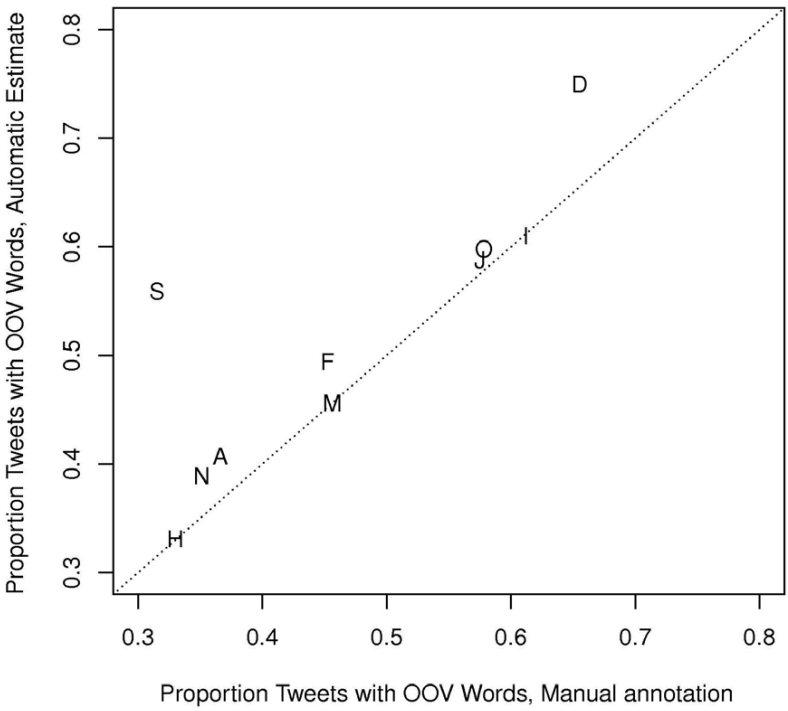

Figure 4. Benchmarking the automatic estimate of the proportion of tweets including OOV words on the annotated pilot corpus. Each letter represents one hash-tag-based sample in the pilot corpus, as listed in Section 2.

6 Automatic estimates for the whole tweet collection

We applied the estimation procedure described above to all tweets in the eScience Centre Twitter corpus with date stamps between January 1, 2011 and June 30, 2013. In all, almost 2 billion tweets marked with the language indication 'dutch' were processed, comprising about 23 billion tokens of which 17.5 billion are words (76.6%). Overall, the procedure finds about 1.8 billion OOV words (10.2%).

If we look at individual users, we see quite some variation. In Figures 4 to 9, we show measurements for all 1.7 million users producing at least 1,000 words in the given time period.[13] The full lines indicate the measurements when taken over all tweets, and the dashed lines when taken over all tweets produced by users who did not reach the word thres-

[13] The maximum number of words was not restricted and the full production of each user is being measured.

hold (i.e. the less active users). In all, there were more than 38 million active user names during this period, which is remarkably high, given that the Netherlands and Flanders together have only about 24 million inhabitants.

Starting with the OOV words (Figure 4), we first observe a large main cluster ranging from close to 0% up to about 30% OOV, which is likely to be the core Dutch speaking Twitter population. Higher up, especially around 50% and 60%, we find secondary clusters. Looking at the user names involved, we get the impression that these clusters at least partly stem from foreign tweets erroneously marked as Dutch. This does of course affect the overall estimate somewhat, but not too drastically we expect, as only about 1% of the users shows a proportion of OOV higher than 30%, and they contribute only 0.2% of the examined words. Furthermore, we also find Dutch users in the higher regions, with OOV proportions well over 90% for several contact ad feeds which consist of URLs, hashtags and compacted information fields. At the other end of the spectrum, we find users who manage to produce tens of thousands of words without any recognized OOV word. An examination here shows various automatic text generation systems, varying from ads linking to websites, to a solar power driven work of art reporting whether it is awake. Looking at the other two plots (Figures 5 and 6), we see that the proportion of non-word tokens varies more predictably, mainly between 0% and 40%. In Figure 6, the secondary clusters again show up, but mostly on the OOV dimension, so that it would seem that the proportion of non-word tokens is similar, although possibly a bit higher, for the other languages involved.

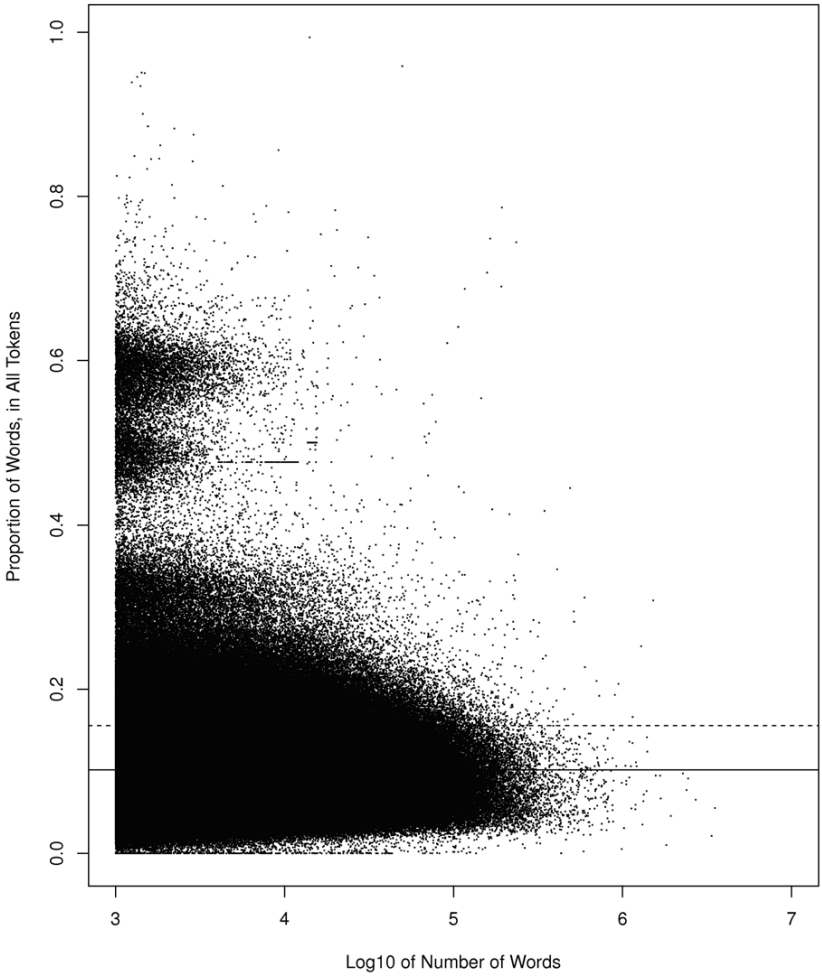

Figure 5. The estimated proportion of OOV words as a function of the produced number of words for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

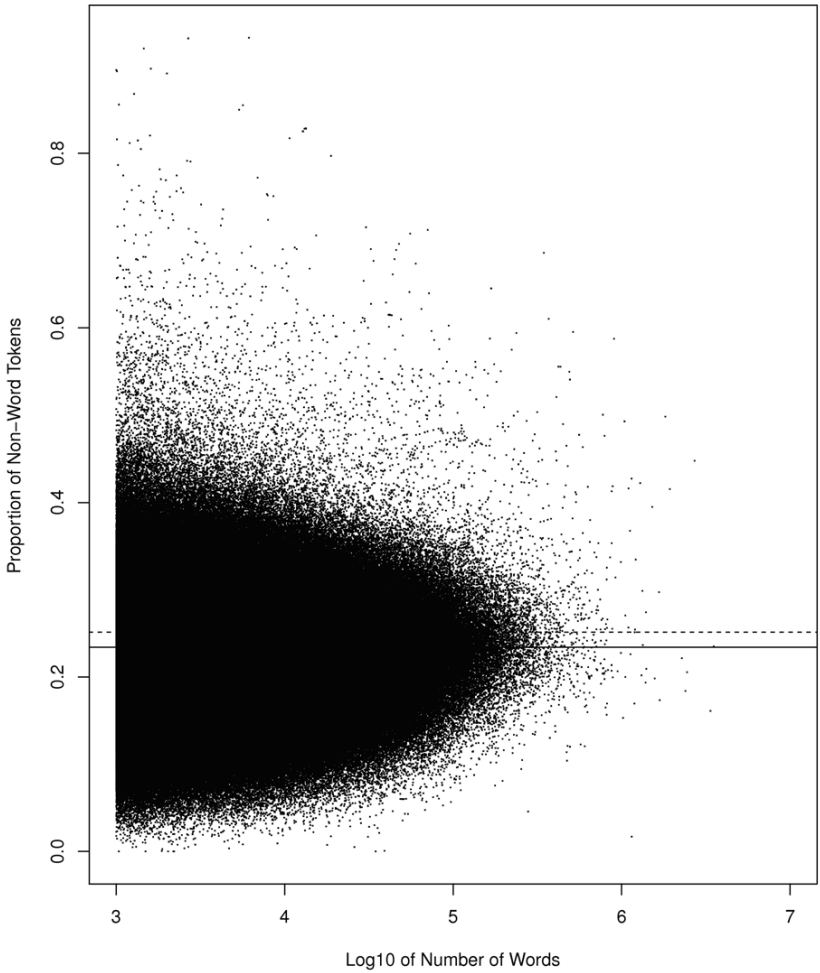

Figure 6. The estimated proportion of non-word tokens as a function of the produced number of words for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

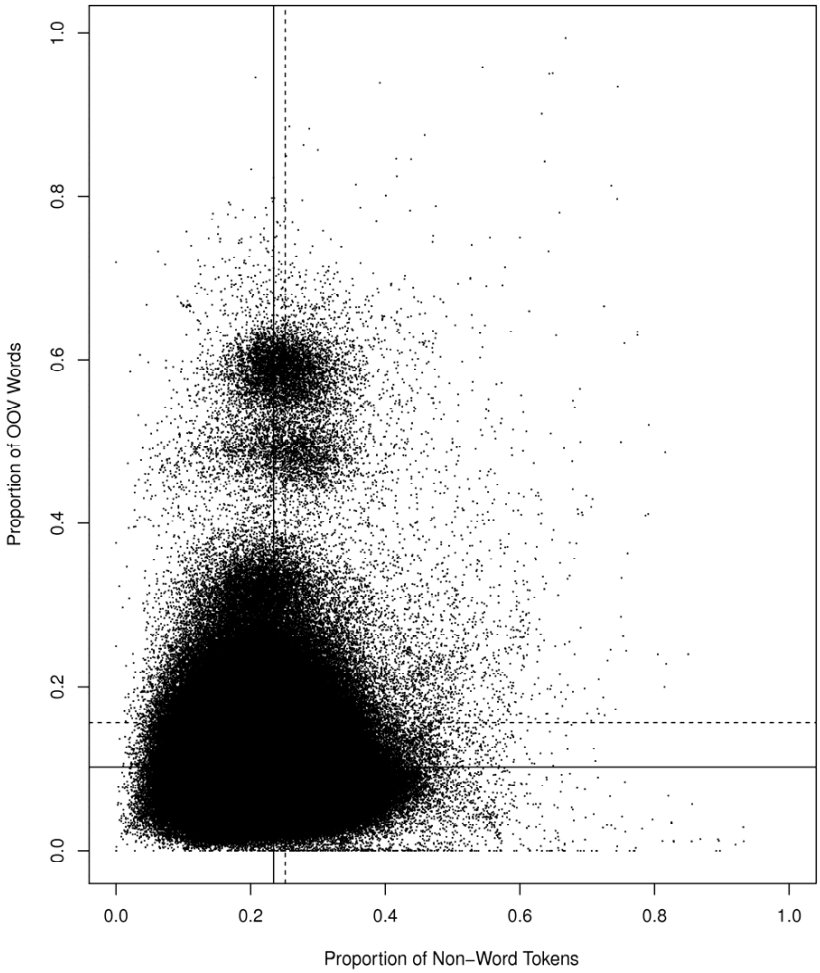

Figure 7. The estimated proportion of OOV words as a function of the estimated proportion of non-word tokens for all users with a production of at least 1,000 words. Each data point represents one user. Lines show the overall scores for all tweets (full) and lower volume users (dashed).

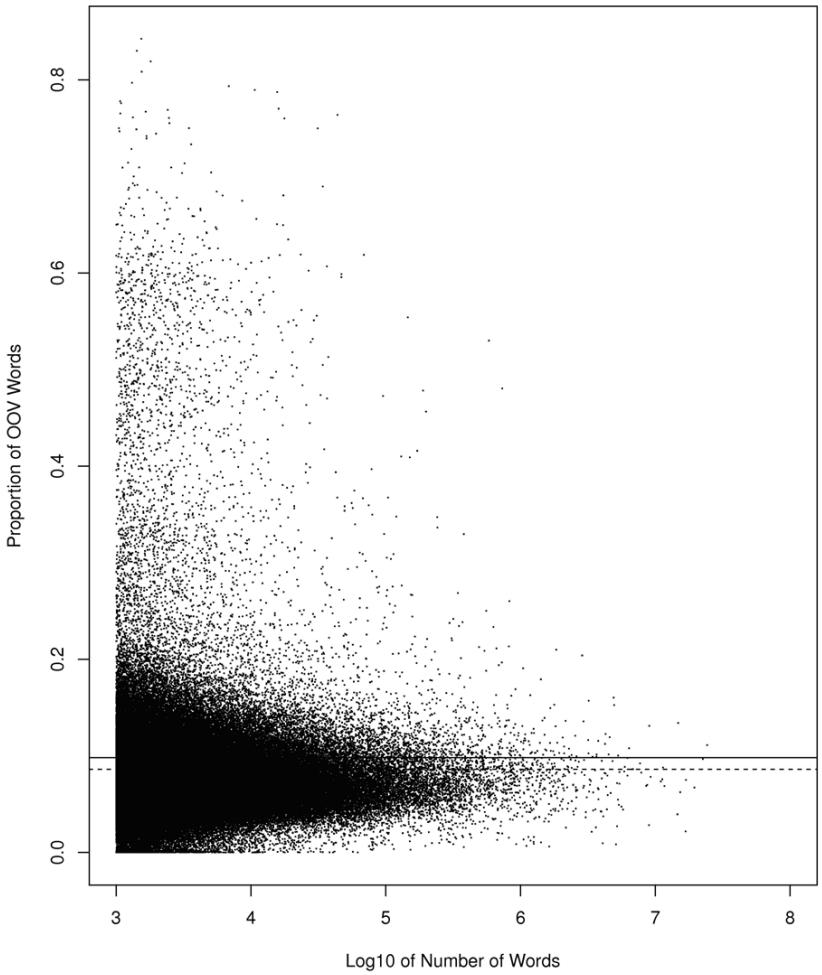

Figure 8. The estimated proportion of OOV words as a function of the produced number of words for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

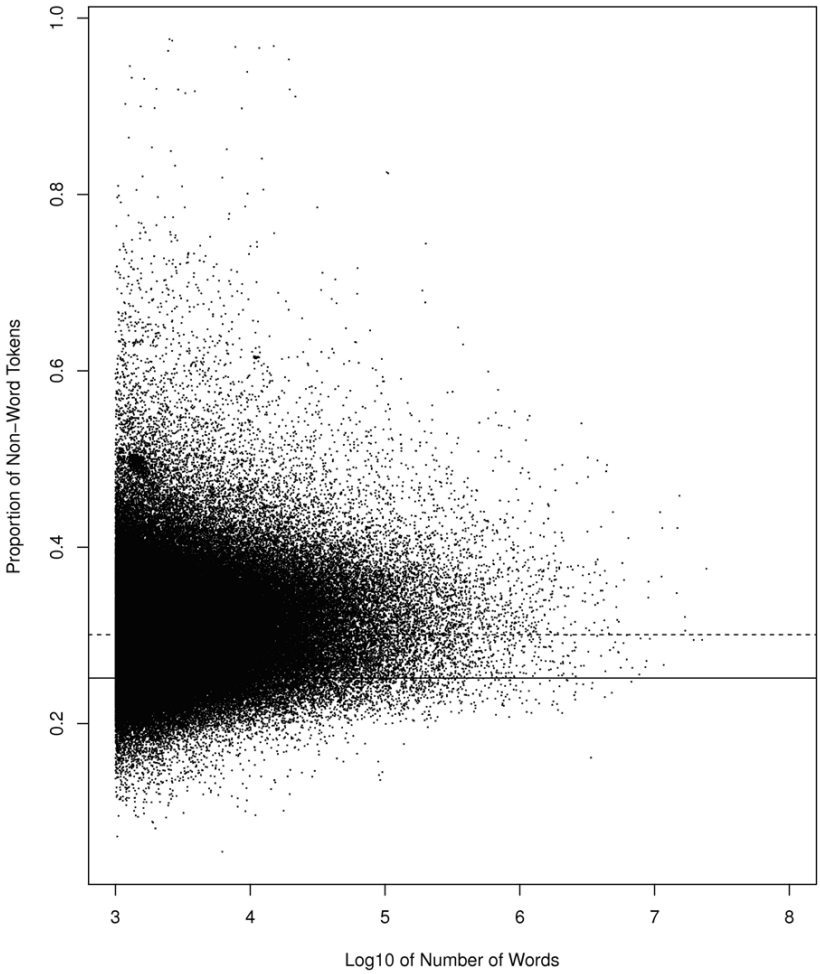

Figure 9. The estimated proportion of non-word tokens words as a function of the produced number of words for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

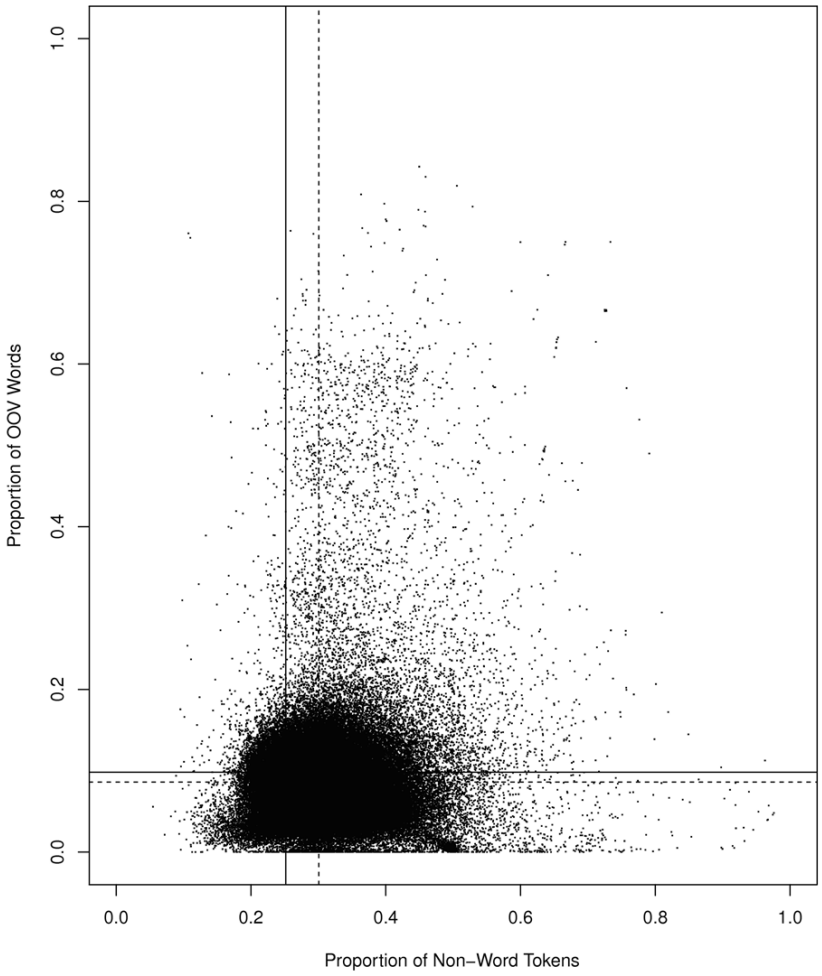

Figure 10. The estimated proportion of OOV words as a function of the estimated proportion of non-word tokens for all hash tags with a production of at least 1,000 words. Each data point represents one hash tag. Lines show the overall scores for all tweets (full) and lower volume hash tags (dashed).

We repeated this investigation for hashtags. Of the approximately 23 million hashtags used in the time period in question, only about 200,000 produced at least 1,000 words. If we examine the same plots as for users (Figures 7 to 9), we again see most activity in the main cluster. For OOV words (Figure 7), the cluster now ranges upto only about 20%. This would imply that tweets with hashtags on average contain fewer OOV words. We also get this impression from the position of the line for the overall estimate. In order to confirm our impression, we separately measured the OOV words for tweets with and without hashtags. Those with a hashtag (2.7 billion words) showed a clearly lower OOV count of 8.1% than those without a hashtag (14.8 billion words) at 10.5%. We do not see any secondary clusters in this plot, but there is a large sparse cloud extending up to around 80% OOV, which on closer examination is dominated by German hashtags. The influence of the overall estimate should be even lower than for the users, as the hashtags with a higher than 20% OOV comprise 1.7% of the hashtags, but contribute less than 0.1% of the words. As for non-word tokens (Figure 8), we see a similar spread, except that the cluster moved up slightly. The average for the tweets with hashtags is 30.1% versus 22.0% for those without hashtags. This difference is easily explained as it is probably completely accounted for by the presence of the hashtag required. The combination plot (Figure 9) does show an interesting small secondary cluster around 50% Non-Word and 0% OOV. This turns out to be caused by a single spam feed, each time asking whether you are looking for a specific specialist profession (hashtag) in a specific location (hashtag) and giving a URL.

7 Conclusion

In this paper, we investigated the proportion of tokens in tweets which might cause problems for automatic processing. We were able to look in detail at a pilot corpus containing for each of ten selected hashtags a random selection of tweets containing around 1,000 words (Sections 2 to 4). We also automatically estimated the proportion of non-word tokens and OOV words on almost 2 billion Dutch tweets (Section 6), after having shown that the automatic estimation procedure is adequate for these two measurements (Section 5).

In our annotated samples, we see that the proportion of non-word tokens ranges from 20% to 36%. The proportion of OOV words ranges from 4% to 11%, whereas forms judged to be in vocabulary because they are homographs of listed words range from 2% to 6%. Especially the latter class calls for our attention, as these tokens tend to go undetected at first but are bound to have a negative effect when attempting to automatically process texts with standard resources. In all, there are problematic words in 37% to 75% of the tweets in the examined ten samples. In an automatic investigation of all tweets, we see that, for the bulk of tweet types, the proportion of non-word tokens ranges between 0% and about 40%, with an average of around 23%, and of OOV words between 0% and about 30%, with an average of around 10%. In the automatic investigation, we did not attempt to investigate the proportion of confused words and the proportion of tweets with problematic cases; however, if the average of about half as many INVOC-ALT words as OOV holds for the whole collection, there should be about 15% of problematic words in total. The differences between the measurements for annotated and automatically processed material are consistent with our fin-

dings that tweets with hashtags generally have a higher proportion of non-word tokens, but a lower proportion of OOV words.

We found pronounced differences between tweet samples focusing on different topics. In the annotated material, there is a clear gap between the OOV proportions for the hashtag with a higher expected emotion load and those with a lower one. We are somewhat surprised that the tweets related to Dutch Rail are found to be in the non-emotional cluster, but then there is a large number of official tweets among them. The difference due to emotional load is not visible in the automatic estimates, but this may well be because of the plot denseness caused by showing measurements for 200,000 hashtags. For non-word tokens, neither analysis shows a clear pattern. There is also extreme variation between users, as can be expected, but we observe no recognizable clusters. Outside the main bulk of users and hashtags, we find several deviant types of tweets, such as spam. Also, we find foreign language tweets, many of them German, even though the language filter marked them as Dutch. Although these do not appear to seriously impact our main findings, we would like to investigate how the material can be cleaned up in this respect.

Where the automatic estimates only provide overall percentages, the analysis of the annotated material gives us more insight in the types and distribution of problematic tokens. The main conclusion from our analysis is that at least some of the problems can be solved with relatively simple means, e.g. addition of lexicalized items to the lexicon (about one in three of the problematic cases) or patterns matching techniques (about one in ten), or possibly more involved ones, e.g. adaptation of named entity recognition to the Twitter environment (about one in four).

With these findings, we can now also address our underlying research questions. Given the observed proportions of problematic cases, it would seem unwise in most cases to attempt to process the bulk of Dutch tweets with NLP tools developed for standard Dutch. Especially the proportion of tweets in which problems arise is too high for this. However, it might be possible to process tweets authored by specific (types of) users or containing specific hashtags. And there are clear approaches to alleviate the problem by adjusting the tools. On the other hand, we should remember that we have only addressed the lexicon in this paper, and that other problems such as deviant syntax are also present.

Now that it has been confirmed that Dutch tweets need special means for proper processing, we will continue our work in this area. Apart from the work on spelling variation (van Halteren and Oostdijk 2012), this is likely to include improved language filters so that we can better focus on Dutch. Also, we are inspired to initiate a deeper investigation into classes of users and topics, as it appears that the best approach to processing might well vary per class, and that, for most types of research, it is advantageous to focus on specific types of tweets, e.g. those either more or less emotional. Furthermore, for most types of research one probably would like to remove spam tweets. Finally, seeing the proportion of problematic cases we found, and seeing for example the surprising finding that tweets with hashtags are somehow different from those without, it might also be interesting to investigate how these facts could have affected previous or existing research.

Bibliography

Han, B. and T. Baldwin (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In Proceedings of the 49th Annual Meeting of the Assoc. of Computational Linguistics. Portland, Oregon, June 19-24. ACL. 368-378.

Krikorian, R. (2013). New Tweets per second record, and how! Twitter Offical Blog. August 16, 2013.

Oostdijk, N. and H. van Halteren (2013). N-gram-based Recognition of Threatening Tweets. A. Gelbukh (ed.) CICLing 2013, Part II, LNCS7817, pp. 183-196. Springer-Verlag Berlin/Heidelberg.

Small, H., K. Kasianovitz, R. Blanford and I. Celaya (2012). What Your Tweets Tell Us About You: Identity, Ownership and Privacy of Twitter Data. The International Journal of Digital Curation. Vol. 7(1): 174-197.

Sidarenka, U., T. Scheffler and M. Stede (2013). Rule-based normalization of German Twitter messages. Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology.Sept. 25-27, 2013, Darmstadt, Germany. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/gscl2013/workshops/sidarenka_scheffler_stede.pdf

Tagg, C. (2009). A Corpus Linguistics Study of SMS Text Messaging. http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf

Tjong Kim Sang, E. and A. van den Bosch (2013). Dealing with Big Data: The case of Twitter. CLIN Journal Vol. 3:121-134.

van Halteren, H. and N. Oostdijk (2012). Towards Identifying Normal Forms for Various Word Form Spellings on Twitter. CLIN Journal Vol. 2: 2-22.

Weil, K. (2010). Measuring Tweets. Twitter Offical Blog. Februari 22, 2010.

Appendix

In Tables A and B a more detailed overview is given of our findings in the pilot corpus as regards the frequency and distribution of the different types of out-of-vocabulary words (OOV; Table A) and in-vocabulary words that are used in an alternative way, i.e. other than the use foreseen for their entry in the word list (INVOC-ALT; Table B).

The different samples in the pilot corpus are referred to by means of capital letters as follows: A=#aardbevingen; D=#doodsbedreiging; F=#file; H=#houdoe; I=#irri; J=#jaloers; M=#miljoenenjacht; N=#ns; O=#omg; S=#syrie. Each of the hashtags has been described briefly in Section 2. For a description of the different types of OOV and INVOC-ALT word tokens, we refer to Section 3.

Table A. Detailed overview of the frequency and distribution of the different types of out-of-vocabulary words. The single letter column headingcs represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

	A	D	F	H	I	J	M	N	O	S
Total OOV	43	90	56	111	109	112	75	45	110	39
abbreviation-lex	13	21	5	24	21	15	16	10	18	10
abbreviation-prod	2	3	1	6	4	4	1	1	2	1
clipped	4	0	0	0	0	0	1	2	0	1
clitic	0	0	1	2	0	0	1	0	1	0
complex	0	0	0	0	1	0	0	7	1	0
compound	1	5	5	3	3	2	2	2	1	3
dialect-oov	0	0	0	4	0	0	0	0	0	0
diminutive	0	0	0	0	1	1	0	0	0	0
emoticon	0	0	0	0	2	2	1	0	3	0
foreign	5	4	4	15	2	10	6	3	11	6
formula	0	0	1	1	0	1	0	0	0	0
hyphenation	1	1	0	0	0	0	0	0	0	0
interjection	1	6	7	3	10	36	21	1	19	0
merge-aut	1	3	3	7	10	10	5	2	7	0
missing-neo	2	3	1	1	6	3	4	2	8	2
missing-trad	1	2	1	0	0	0	2	0	0	0
part_of_multi-word	1	0	2	0	0	0	0	0	0	0
proper_name-oov-cap	11	13	17	1	4	8	11	16	5	19
proper_name-oov-decap	1	9	8	12	9	9	9	4	3	8
spelling-lex	1	9	2	5	9	5	2	2	6	0
spelling-prod	3	10	4	30	28	15	7	5	31	1
split-aut	0	0	1	2	0	0	1	0	0	0
split-typo	0	0	0	0	1	0	0	0	0	0
street_language-oov	0	7	0	1	0	1	0	0	0	0
unkn	0	0	0	5	1	2	0	0	3	0

Table B. Detailed overview of the frequency and distribution of the different types of in-vocabulary words. The single letter column headingcs represent the hash-tag-based samples in the pilot corpus, as listed in Section 2.

	A	D	F	H	I	J	M	N	O	S
Total INVOC-ALT	34	51	40	58	34	41	27	49	46	17
abbreviation-lex	6	13	15	15	11	8	6	29	10	6
abbreviation-prod	2	1	0	4	2	0	0	0	2	1
clipped	2	2	0	0	0	0	0	1	0	1
clitic	0	0	0	0	1	1	1	2	0	0
compound	0	0	0	0	0	0	0	0	0	1
dialect-conf	0	0	0	3	1	0	0	0	0	0
foreign	2	0	4	2	0	6	2	0	3	2
interjection	0	2	0	1	0	2	0	0	0	1
merge-aut	0	0	0	2	0	1	0	0	0	0
part_of_multi-word	1	0	2	0	0	0	0	0	0	0
proper_name-inlex-decap	1	6	4	6	4	8	1	3	5	0
proper-name-conf-cap	14	3	2	2	1	0	2	4	6	2
proper_name-conf-decap	1	9	8	5	3	4	3	2	1	0
proper_name-inlex-capdia	2	0	0	0	0	0	0	1	1	1
proper_name-inlex-decapdia	0	0	0	0	1	1	0	0	0	0
spelling-lex	0	9	2	9	4	5	5	0	5	0
spelling-prod	0	2	2	7	5	5	5	6	11	2
split-aut	4	4	2	2	1	1	2	2	0	1
split-hash	0	0	0	0	0	0	0	2	0	0
street_language-conf	0	0	0	0	0	0	0	0	0	0
unkn	0	0	0	3	0	0	0	0	2	0

Author Index

Georges Antoniadis
Université Stendhal Grenoble 3
Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles
Georges.Antoniadis@u-grenoble3.fr

Michael Beißwenger
TU Dortmund
Institut für deutsche Sprache und Literatur
michael.beisswenger@tu-dortmund.de

Thierry Chanier
Clermont Université
Laboratoire de Recherche sur le Langage
thierry.chanier@univ-bpclermont.fr

Aivars Glaznieks
European Academy of Bolzano/Bozen (EURAC)
Institute for Specialised Communication and Multilingualism
aivars.glaznieks@eurac.edu

Linda Hriba
Université d'Orléans
Laboratoire Ligérien de Linguistique
linda.hriba@yahoo.fr

Julien Longhi
Université de Cergy-Pontoise
Centre de recherche textes et francophonies
julien.longhi@u-cergy.fr

Harald Lüngen
Institut für Deutsche Sprache (IDS), Mannheim
Programmbereich Korpuslinguistik
luengen@ids-mannheim.de

Eliza Margaretha
Institut für Deutsche Sprache (IDS), Mannheim
Programmbereich Forschungsinfrastrukturen
margaretha@ids-mannheim.de

Nelleke Oostdijk
Radboud University Nijmegen
Centre for Language and Speech Technology
n.oostdijk@let.ru.nl

Celine Poudat
Université Paris 13, Sorbonne Paris Cité

Laboratoire Langues, Textes, Traitements informatiques, Cognition
cpoudat@gmail.com

Benoit Sagot
Inria & Université Paris-Diderot
Analyse Linguistique Profonde A Grande Echelle
benoit.sagot@inria.fr

Djamé Seddah
Université Paris Sorbonne & Inria
Analyse Linguistique Profonde A Grande Echelle
djame.seddah@paris-sorbonne.fr

Wilbert Spooren
Radboud University Nijmegen
Dpt Dutch Language / Discourse studies
w.spooren@let.ru.nl

Egon W. Stemle
European Academy of Bolzano/Bozen (EURAC)
Institute for Specialised Communication and Multilingualism
egon.stemle@eurac.edu

Angelika Storrer
Universität Mannheim
Seminar für Deutsche Philologie
astorrer@mail.uni-mannheim.de

Tessa van Charldorp
VU University Amsterdam
Faculty of Humanities
t.c.van.charldorp@vu.nl

Henk van den Heuvel
Radboud University Nijmegen
Centre for Language and Speech Technology
h.vandenheuvel@let.ru.nl

Hans van Halteren
Radboud University Nijmegen
Centre for Language and Speech Technology
hvh@let.ru.nl

Ciara R. Wigham
Université Lumière Lyon 2
Laboratoire Interactions, Corpus, Apprentissages et Représentations
ciara.wigham@univ-lyon2.fr