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 meta-data 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.

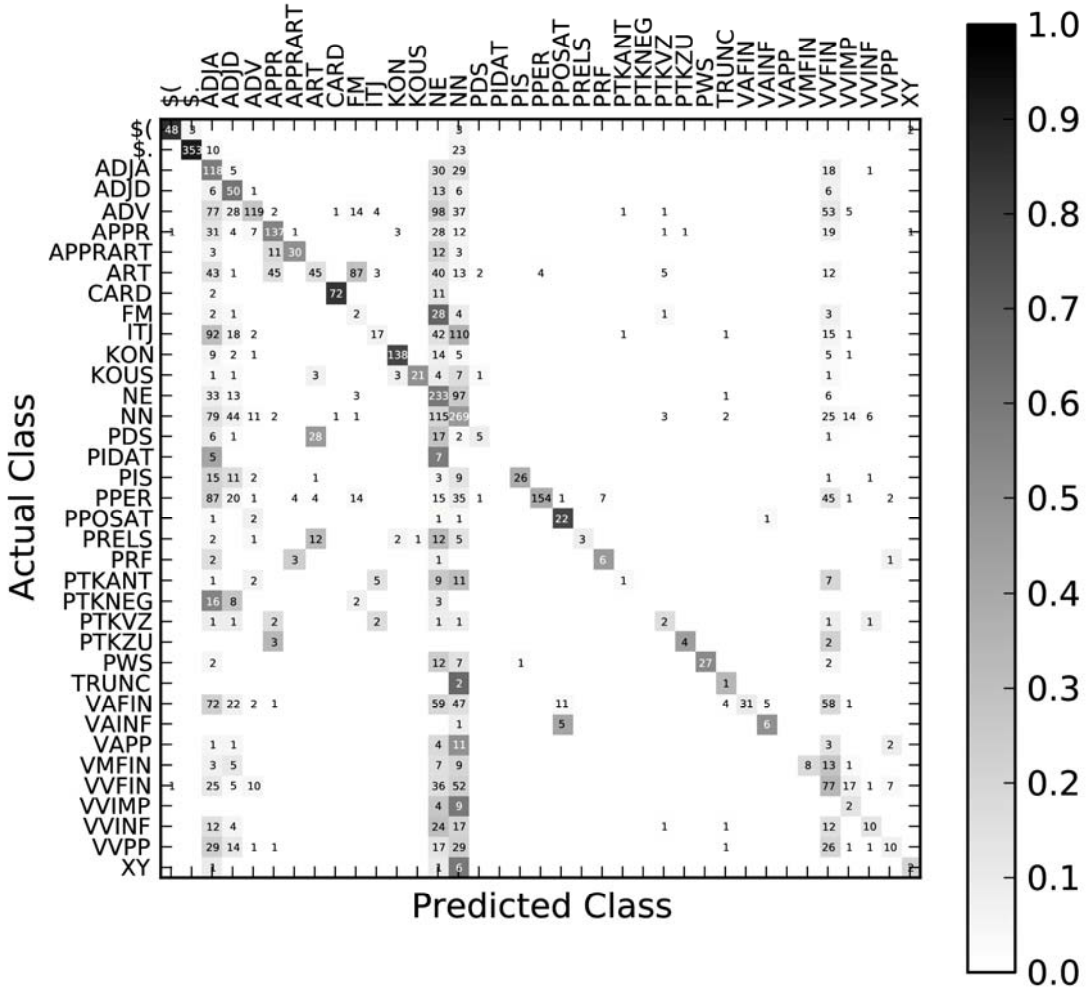


**Figure 1:** Confusion matrix for all wrongly tagged tokens (>5 instances) in ORG (see Schiller et al. (1999) for a glossary of all POS tags)

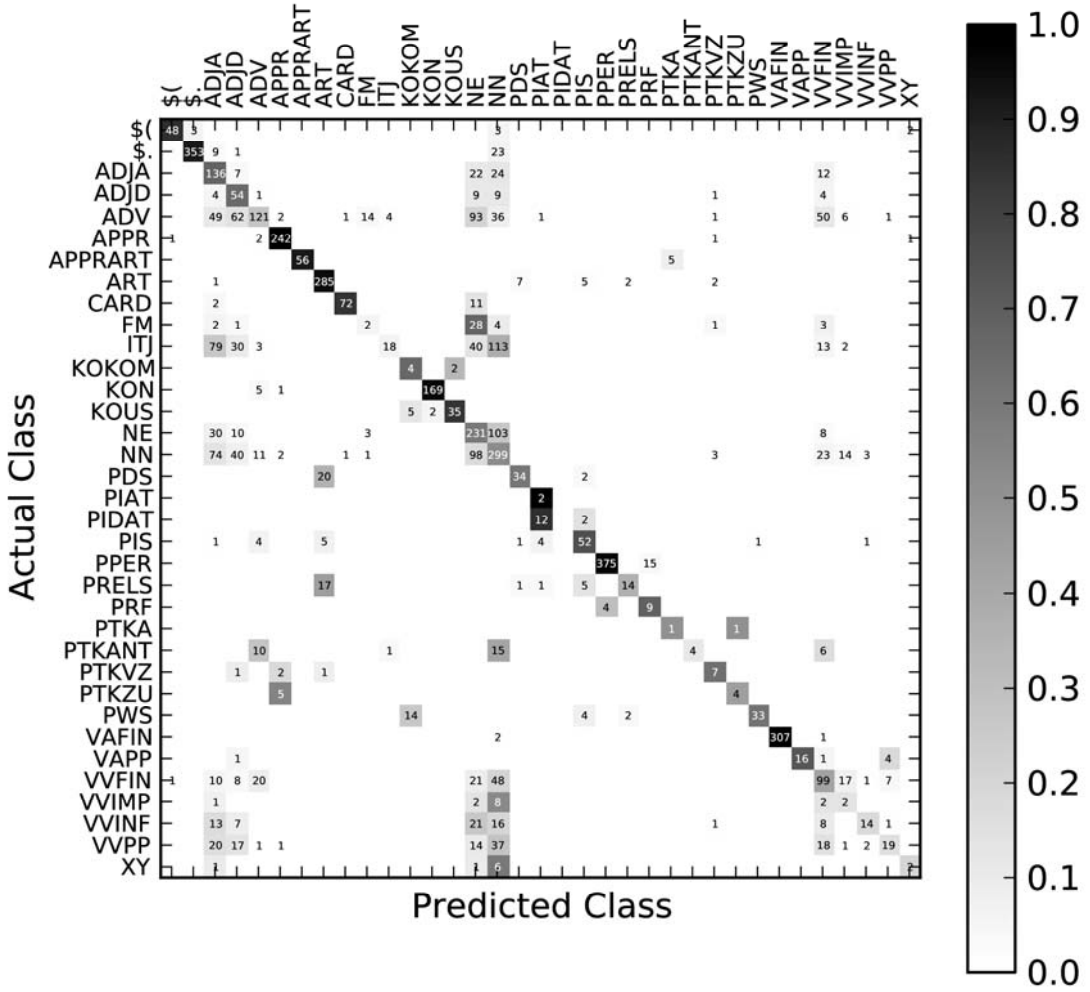


**Figure 2:** Confusion matrix for all wrongly tagged tokens (>5 instances) in CLSD

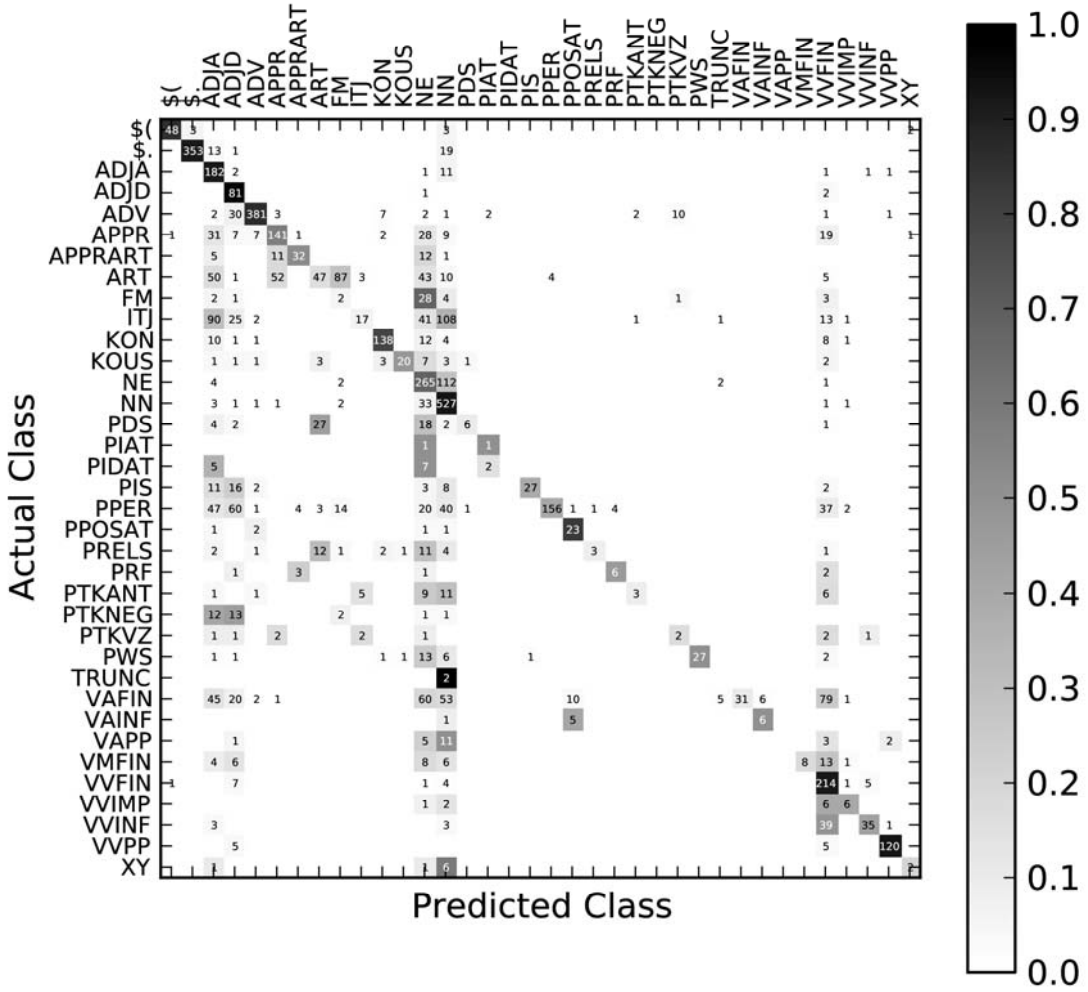


**Figure 3:** Confusion matrix for all wrongly tagged tokens (>5 instances) in OPEN

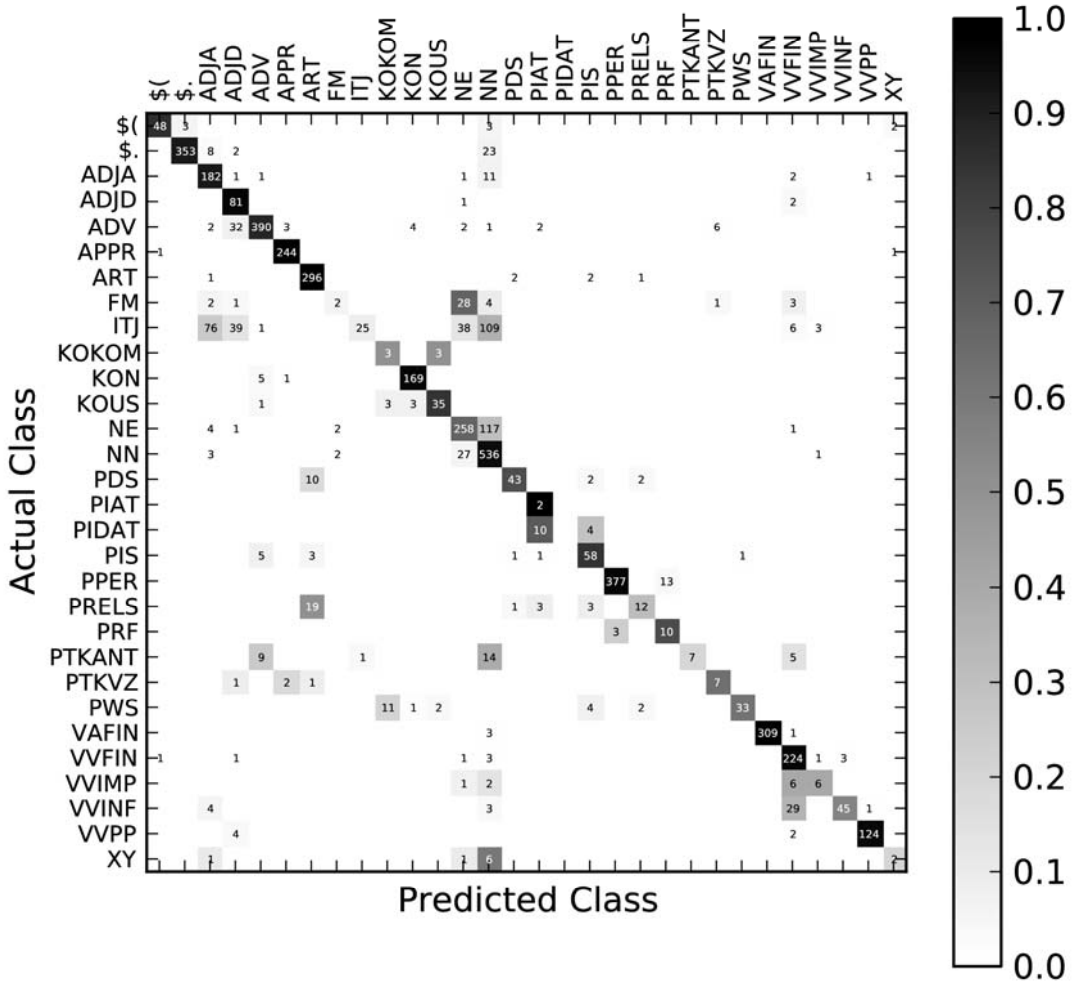


**Figure 4:** Confusion matrix for all wrongly tagged tokens (>5 instances) in C&O

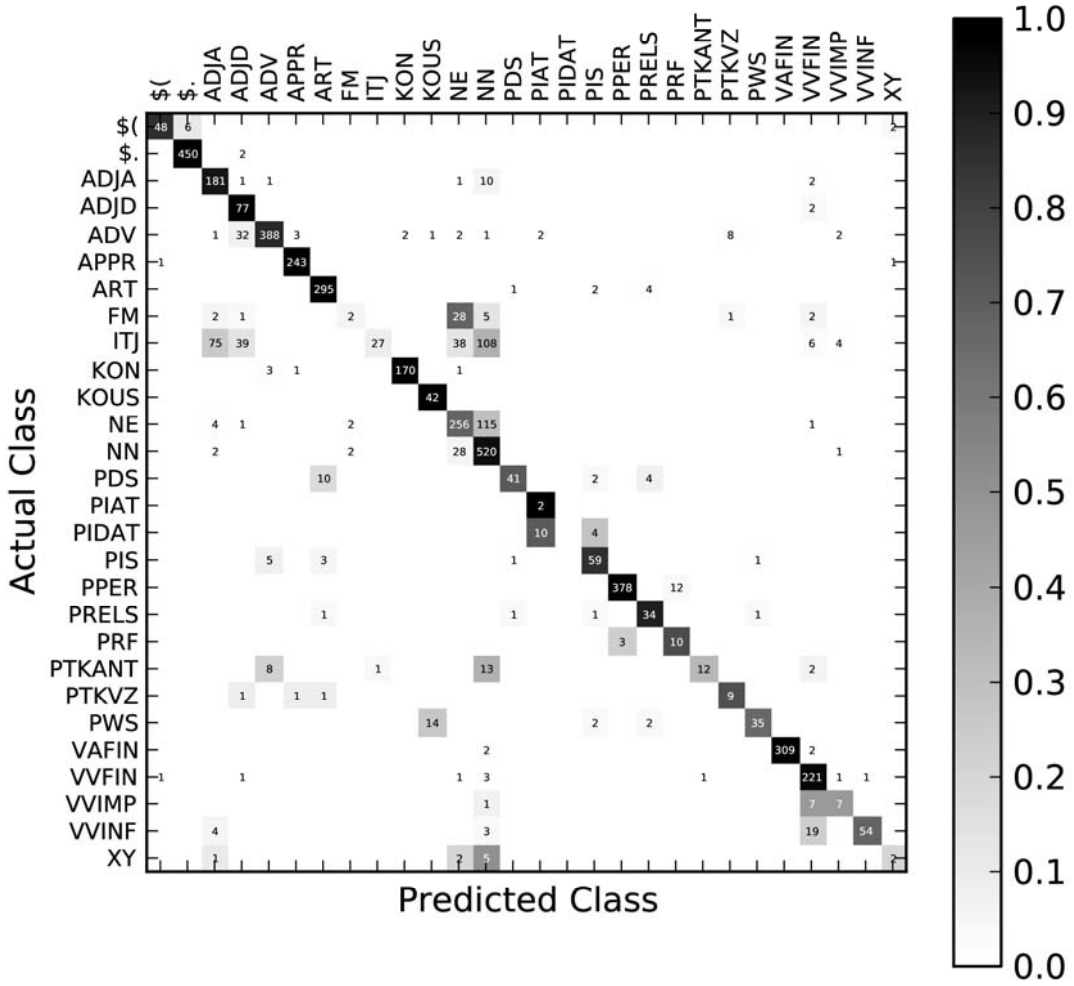


**Figure 5:** Confusion matrix for all wrongly tagged tokens (>5 instances) in ALL

## 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.