The DiDi Corpus of South Tyrolean CMC Data

Jennifer-Carmen Frey **Aivars Glaznieks** **Egon W. Stemle**
Institute for Specialised Communication and Multilingualism
European Academy Bozen/Bolzano
Viale Druso 1, 39100 Bolzano
{jennifer.frey,aivars.glaznieks,egon.stemle}@eurac.edu

Abstract

This paper presents the DiDi Corpus, a corpus of South Tyrolean Data of Computer-mediated Communication (CMC). The corpus comprises around 650,000 tokens from Facebook wall posts, comments on wall posts and private messages, as well as socio-demographic data of participants. All data was automatically annotated with language information (de, it, en and others), and manually normalised and anonymised. Furthermore, semi-automatic token level annotations include part-of-speech and CMC phenomena (e.g. emoticons, emojis, and iteration of graphemes and punctuation). The anonymised corpus without the private messages is freely available for researchers; the complete and anonymised corpus is available after signing a non-disclosure agreement.

1 Introduction

The aim of the DiDi project was to build a text corpus to document the current language use of German native speakers from the multilingual province of South Tyrol. We collected a CMC corpus consisting of Facebook wall posts, comments on wall posts and private messages, as well as socio-demographic data of the writers (cf. Glaznieks and Stemle (2014) for more details). Thus, the corpus combines socio-demographic data of the investigated Facebook users such as their language biography, internet usage habits, general parameters such as age, gender and education level with texts on their Facebook profiles. This facilitates sociolinguistic analyses, which has been a secondary objective of the project. To investigate the languages and language varieties used and relate them to socio-demographic parameters (particularly focussing on the users' age and internet experience), a number of annotations have been added to the data. We annotated the predominant language and language variety used, and special CMC phenomena, added a standard transcription to non-standard words, applied part-of-speech tagging, and lemmatisation, and anonymised the corpus considering ethical and legal privacy issues.

2 Corpus

The DiDi corpus has an overall size of around 650,000 tokens gathered from 136 South Tyrolean Facebook users who participated in the DiDi project. It consists of 11,102 Facebook wall posts, 6,507 comments on wall posts and 22,218 private messages of the participants. All messages were written by the participants during the year 2013 (section 3). Although downloading messages from friends and other Facebook users on participant-initiated posts was possible[1], this data must not be used for privacy issues. Consequently, all extraneous data of this kind was removed from the corpus except for the number of replies to messages, the language and the time stamp. This was deemed appropriate in terms of privacy and will likely be relevant for conversational and discourse-centred linguistic analyses of the data, i.e. the corpus does not allow for textual analyses of conversational interaction. As every participant could offer either the private messages and/or the texts on the wall, we were given access to 130 wall profiles and 56 private inbox profiles; 50 participants granted access to both types of data.

3 Data Collection

According to our project design, we aimed for 3 types of data from at least 100 South Tyrolean Facebook users (all with German as L1 and equally spread over various age groups):

1. user consent and privacy agreement,

2. Facebook texts (wall and/or private messages) from the year 2013,

3. socio-demographic data of users' language biography and internet usage habits.

To acquire these types of data from every user in the most structured (for researchers) and simple (for participants) way, we developed a web application that provided an interface to recruit users, inform them about the project's aims and methodologies, allow them to subscribe and explicitly agree to the usage of their data, fill in the online questionnaire and give them the possibility to grant us access to their Facebook data via the Facebook API. Our web application set-up enabled us to download and merge all the necessary data while saving it into our internal document-oriented NoSQL database[2]. See (Frey et al., 2014) for an in-depth description of the process and its technical details.

The user recruitment was mainly accomplished by circulating the web application's URL using chain sampling within Facebook. Additionally, the link was posted in various South Tyrolean Facebook groups and other social media communities to draw further attention to the project. In order to reach more potential users, particularly in older age groups, was targeted Facebook advertising in which the link and some text were posted directly to the walls of South Tyrolean users matching the specific user group.

4 Corpus Annotation

All subsequently mentioned annotation tasks were carried out by three annotators according to a set of annotation guidelines. The tasks were carried out within our processing pipeline: annotators use their favourite spreadsheet program, e.g. Microsoft Excel or LibreOffice, and the pipeline converts between the spreadsheet representation and the structured representation of the database (and vice versa). The spreadsheet representation is a vertical file with one token per line, and individual (blocks of) columns represent annotation layers, which have to be edited according to our annotation guidelines[3]. For example, to merge multiple tokens into one (because they were misspelled) edit the appropriate column and write the proper normalisation in one field and the special token `'___'` in the column's next line(s).

Problems of individual tasks were compared, and differences were discussed until a consensus was reached. If necessary, the annotation guidelines were updated and previous annotation work (sometimes) redone.

5 Corpus Processing

After the original data provided by the Facebook API and the data from the user questionnaire were downloaded and stored, the data went through various natural language processing (NLP) and annotation steps.

NLP of social media texts is still an unsolved problem. Social media corpora contain many short and noisy texts and the content is usually strongly contextualised; therefore, the corpora differ from each other in many ways and are very domain dependent. NLP algorithms are traditionally trained on news-based corpora and these differences affect their performance. (See, for example, Preotiuc-Pietro et al. (2012) and Baldwin (2012).)

For social media texts from South Tyrol, i.e. for our domain, Glaznieks and Stemle (2014) analysed tokeniser and part-of-speech (POS) tagger performance on non-normalised Facebook data of dialect writers, and they evaluated the added value of various levels of normalisation on the source data. In this *pre-test*, they showed that the poor base-line performance of non-normalised data for this domain can be considerably improved by normalisation.

5.1 Tokenisation

After testing a number of tokenisers for social media texts (most of which are tuned to a specific domain), we decided to use the Python version of the Twitter tokenizer `ark-twokenze-py`[4] as it showed the best results with our non-public Facebook data and could already deal with most of the CMC-related difficulties such as emoticons, hyperlinks and individual abbreviations. However, some problems highlighted by the *pre-test* such as incorrect splitting of various time and date formats or words written with special characters to express users' individual, artistic style were still tokenised poorly and therefore manually corrected.

5.2 Normalisation and tagging of privacy issues

As a result of the *pre-test* we invested most of the manual annotation work in normalising the texts. Keeping in mind the project goals, we only normalised German texts of L1 German speakers in

the corpus by using word-by-word transcriptions for each word that was not spelled in standard German be it because of diverge writing or the use of a dialect variety (see Ruef and Ueberwasser (2013) for more details). We used Duden online[5] as a reference to define the target standard spelling of words.

In this annotation task, compound words that were not written as one token in the original were merged together, and words that should have been split into multiple tokens according to standard German were inserted as separate tokens.

While adding this normalisation information, privacy issues were also indicated for later referencing and processing (section 5.5).

5.3 POS-tagging and lemmatisation

`TreeTagger`[6] for German (Schmid, 1995) and the Stuttgart-Tübingen-TagSet (`STTS`)[7] was used for POS-tagging and lemmatisation. The initial results on the previously normalised data were later improved by additional annotation work as the annotations allowed to assign fixed POS tags to previously error-prone tokens (cf. sections 5.4, 5.5, 5.6).

5.4 Handling of dialect lexemes and out-of-vocabulary tokens

During the manual normalisation (section 5.2), the annotators already indicated dialect words that had no equivalent and therefore no spelling in standard German. This information was then used to compile a list of dialect lexemes that are unique for South Tyrolean German. While the list was also linguistically interesting, the primary goal was to unify different spellings of the lexemes and provide an additional lexicon containing part-of-speech information for the POS-tagging and for other subsequent automatic procedures (e.g. classification of used variety). Furthermore, most of the common out-of-vocabulary (OOV) words were listed. Dialect lexemes, foreign language insertions, emoticons and abbreviations that occurred in large amounts were identified, classified as one of those categories and automatically annotated and processed afterwards. This also helped to further improve the POS-tagging of the corpus as in most cases fixed POS tags could be assigned to them (e.g. foreign language insertions received the POS tag *FM* of the `STTS`).

5.5 Anonymisation

The previously indicated privacy issues (section 5.2) were categorised as follows: personal names, group names, geographical names and adjectival references, institution names, hyperlinks, e-mail addresses, phone numbers, and a miscellaneous category containing other private information like numbers of bank accounts, and servers, postal codes, etc.

The original entities were then substituted with information-based type identifiers (PersNE, GruppeNE, GeoNE, GeoADJA, InstNE, link, mail, tel, XXX) that showed the anonymised category (cf. Panckhurst (2013)), keeping any inflectional affixes (e.g. “PersNEs Privatsphäreneinstellung”) and word formations (e.g. “InstNE-Zeltlager”). In addition, the categories determined the POS information on the POS layer (e.g. the POS tag *NE* was assigned to all tokens anonymised as *PersNE* of the `STTS`). This method allows for better readability of data that often consists of several private details, whereas a pure overwriting often leads to nonsense texts. Furthermore, it facilitates automatic analyses on the used private entities since categories and POS tags are defined and reliable.

5.6 Linguistic annotation

A number of annotations such as the predominant language of a text, the variety of German and predefined CMC phenomena were created in order to answer the project’s research questions (cf. Glaznieks and Stemle (2014)). Only those CMC phenomena (e.g. Bartz et al. (2013), Schlobinski and Siever (2013)) that are clearly distinguishable from dialect writing were used. For this reason, we annotated phenomena such as emoticons, emojis, @mentions, CMC-specific acronyms and abbreviations, iterations of graphemes, punctuations and emoticons, asterisk expressions (action words), hyperlinks, and hashtags as CMC phenomena. Other features that either originate from emulation of spoken language (e.g. assimilation, clitics, etc.) or represent deviations from standard German orthography (e.g. case insensitivity) were not categorised as CMC phenomena in order to avoid confusion with particularities induced by writing in dialect. So far, such phenomena were only normalised with standard German equivalents but not annotated with a specific tag. More details on the annotation procedure and results are given in section 6.

6 Corpus Data

There are two types of data in the corpus: (a) sociodemographic data for each participant who also shared Facebook texts, and (b) texts with their linguistic annotations. Both are described in the following.

6.1 User meta data

The meta data of each user provides the necessary demographic data to carry out sociolinguistic analyses with the given language data.

6.1.1 Data gathered by questionnaire

Within the web application, we asked for sociodemographic data of the participants that was mainly centred on the users' language and internet usage biography. Additionally, some standard parameters such as gender, age, level of education and current employment were gathered. Table 1 shows some of the questionnaire data from the DiDi corpus.

Meta data	**Texts L1 German**	**Texts Total**
female	18,615	20,273
male	16,545	19,554
14-19 years	5,807	5,807
20-29 years	5,225	5,289
30-39 years	7,215	7,514
40-49 years	5,258	8,377
50-59 years	9,519	10,016
60 years and older	2,136	2,824
university degree	8,728	11,972
matura	13,893	14,781
no matura	7,362	7,362
no data	5,177	5,712
employed	12,083	15,410
self-employed	9,653	9,806
in education	10,302	10,373
unemployed	2,946	2,946
no data	176	1,292
total	35,160	39,827

Table 1: Distribution of texts by user groups

6.1.2 Data gathered via Facebook

As the Facebook API provides a number of data fields for participating users such as gender, language preferences, etc., we merged these fields in the corpus as far as ethical and moral appropriateness was given.

6.2 Language data

Whereas the data provided by the Facebook API for each user was not exhaustive and mainly not publishable due to privacy issues, the language data was already enriched by various annotations. Some of the most important in terms of linguistic analyses could be named as: timestamp for creation and editing of text, privacy settings for the text, reactions in form of likes, comments, and shares of that text, attachments such as photos, videos or hyperlinks, recipients of private messages and the application used for publishing the text (e.g. *Facebook for Android/iPhone*, *Twitter*).
Annotations that were added to the data for the purpose of the linguistic analysis can be split into text level annotations and token level annotations.

6.2.1 Text level annotations

A number of additional annotations where made to enrich the data gathered from Facebook.

Language The language was annotated on text level using `langid.py` language identification tool (Lui and Baldwin, 2012). The basic automatic annotation was refined manually by validating and correcting every language annotation that was

- under a threshold of 0.8 confidence [8]
- shorter than 30 characters [9]
- identified as neither German, English, Italian, French, Portuguese nor Spanish [10].

Table 2 shows the language classifications for the gathered texts.

Language	**Texts**
German	23,258
Italian	8,216
English	4,344
Spanish	197
French	60
Portuguese	50
other [11]	236
not classifiable [12]	3,466
total	39,827

Table 2: Outline of languages in DiDi corpus

Variety of German The normalisation of the corpus data showed that our participants, when writing in German, used the regional dialect in transcribed form to a large extent, however there were also texts

that represented a standard-oriented variety of German. To analyse the differences and proportions of the used varieties we classified the German-tagged texts into 3 categories (see table 3 for details):

1. considered as South Tyrolean dialect,
2. considered as standard-oriented variety of German,
3. not classified.

For the categorisation, we used a rule-based approach based on previously compiled lists of untranslatable dialect lexemes and most common dialect-standard transcriptions as well as information on the quantity and quality of the token's divergence to the standard transcription (see table 3). All texts shorter then 30 characters were not classified for reasons of ambiguity. In addition, the subgroup of not classified texts represented text which included a mixture between standard and dialect varieties that did not allow for a valid classification to either category.

Variety	**Texts**
Standard-oriented German	10,227
South Tyrolean dialect	9,570
Not classified	3,461
Total German texts	23,258

Table 3: Outline of the varieties used in German texts

6.2.2 Token level annotations

The corpus contains the following token level annotations. *Original token:* tokenised automatically and manually corrected. *List of normalisation tokens:* standard transcription of misspelled or dialectal words. *Part-of-speech tag:* on normalised standard transcriptions. *Lemma:* on normalised standard transcriptions. *Foreign language insertions:* according to list of most common OOV tokens classified as foreign language vocabulary. *Untranslatable dialect lexemes:* according to list of dialect lexemes compiled during manual annotation and post-processing of OOV tokens. *CMC phenomena:* list of CMC phenomena rendered relevant for the linguistic analysis of the project's research questions:

- Emoticons
- Emojis
- @Mentions
- Most common CMC acronyms and abbreviations (*cmq*, *thx*, *glg*, ...)
- Iteration of graphemes, punctuation or emoticons
- Asterisk expressions
- Hyperlinks
- Hashtags

7 Conclusion and Future Work

The DiDi corpus provides an insight into private, or at least non-public, informal written language use of people in a multilingual environment. The corpus combines the peculiarities of computer-mediated communication with the sociodemographic data of the writers in question and allows for a detailed investigation of current communicational strategies and language usage. A profound evaluation of the DiDi corpus is needed to ensure the quality of further investigations. Nevertheless, the corpus already offers a vast range of research opportunities not only for linguists interested in CMC, multilingual language use, the use of regional varieties, etc., but also for researchers interested in the technical processing of such textual content.

Further information regarding downloading the corpus data and querying it via ANNIS[13] is available at `http://www.eurac.edu/didi`.

Acknowledgements

The project was financed by the Autonome Provinz Bozen Südtirol, Abteilung Bildungsförderung, Universität und Forschung, Landesgesetz vom 13. Dezember 2006, Nr. 14 „Forschung und Innovation"

[1] Replies to comments, for example, are interwoven into the original content in such a way that it is impossible *not* to download them.

[2] `http://mongodb.org`

[3] See `http://www.eurac.edu/didi` for details.

[4] `https://github.com/myleott/ark-twokenize-py`

[5] `http://www.duden.de`

[6] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[7] `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`

[8] According to the confidence value stated by langid.py

[9] We defined this as a minimal length as shorter texts were too ambiguous to obtain a reliable classification with automatic tools.

[10] These languages were the most common results from the language identification tool, are partly taught in schools, or have been stated as native languages by participants and were therefore expected to show up in the corpus, whereas

References

[Baldwin2012] Timothy Baldwin. 2012. Social media: Friend or foe of natural language processing? In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 58–59, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

[Bartz et al.2013] Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2013. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *JLCL*, 28(1):157–198.

[Frey et al.2014] Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2014. Collecting language data of non-public social media profiles. In Gertrud Faaß and Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 11–15, Hildesheim, Germany, October. Universitatsverlag Hildesheim, Germany.

[Glaznieks and Stemle2014] Aivars Glaznieks and Egon W. Stemle. 2014. Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project. *JLCL*, 29(2):31–57.

[Lui and Baldwin2012] Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

[Panckhurst2013] Rachel Panckhurst. 2013. A Large SMS Corpus in French: From Design and Collation to Anonymisation, Transcoding and Analysis. *Procedia - Social and Behavioral Sciences*, 95:96 – 104.

[Preotiuc-Pietro et al.2012] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on real-time analysis and mining of social streams*.

[Ruef and Ueberwasser2013] Beni Ruef and Simone Ueberwasser. 2013. The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages. *Non-standard Data Sources in Corpus-based Research*, pages 61–68.

[Schlobinski and Siever2013] Peter Schlobinski and Torsten Siever. 2013. Microblogs global: Deutsch. In Torsten Siever and Peter Schlobinski, editors, *Microblogs global. Eine internationale Studie zu Twitter & Co. aus der Perspektive von zehn Sprachen und elf Ländern*, pages 41–74. Peter Lang.

[Schmid1995] Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

other languages such as Turkish, Danish or Chinese where less probable and therefore manually checked.

[11] The group *other* was used for all manually classified texts that did not belong to any of the previously stated languages.

[12] Category for texts containing solely non-verbal graphs (e.g. emoticons, links, etc.) or ambiguous or multi-language expressions that can not be classified as a single language (e.g. interjections, international greetings or other internationally used words as "super" or "bravo")

[13] `http://annis-tools.org/`