The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts

Jennifer-Carmen Frey **Aivars Glaznieks** **Egon W. Stemle**
Institute for Specialised Communication and Multilingualism
EURAC Research
Bolzano/Bozen, Italy
{jennifer.frey,aivars.glaznieks,egon.stemle}@eurac.edu

Abstract

English. The DiDi corpus of South Tyrolean data of computer-mediated communication (CMC) is a multilingual sociolinguistic language corpus. It consists of around 600,000 tokens collected from 136 profiles of Facebook users residing in South Tyrol, Italy. In conformity with the multilingual situation of the territory, the main languages of the corpus are German and Italian (followed by English). The data has been manually anonymised and provides manually corrected part-of-speech tags for the Italian language texts and manually normalised data for German texts. Moreover, it is annotated with user-provided socio-demographic data (among others L1, gender, age, education, and internet communication habits) from a questionnaire, and linguistic annotations regarding CMC phenomena, languages and varieties. The anonymised corpus is freely available for research purposes.

Italiano. *DiDi è un corpus di comunicazione mediata dal computer (CMC), che raccoglie dati linguistici di area sudtirolese. Il corpus, multilingue e sociolinguistico, è composto da circa 600,000 occorrenze raccolte (previo consenso all'utilizzo dei dati) dai profili di 136 iscritti a Facebook e residenti in Alto Adige. Le principali lingue del corpus, tedesco e italiano (seguite dall'inglese), riflettono lo spazio plurilingue del territorio. I dati sono stati manualmente anonimizzati e i testi in lingua italiana sono corredati da etichette (manualmente corrette) per le parti del discorso. Inoltre, DiDi è annotato con dati sociodemografici forniti dall'utente (fra gli altri: L1, genere, età, istruzione e modalità di comunicazione via Internet) attraverso un questionario e contiene ulteriori annotazioni linguistiche relative a fenomeni legati alla CMC e agli usi di varietà linguistiche. Il corpus anonimizzato è liberamente disponibile a fini di ricerca.*

1 The DiDi Project

The autonomous Italian province of South Tyrol is characterized by a multilingual environment with three official languages (Italian, German, and Ladin), an institutional bi- or trilingualism (depending on the percentage of the Ladin population), and diverse individual language repertoires (Ciccolone, 2010).

In the regionally funded DiDi project,[1] the goal was to build a South Tyrolean CMC corpus to document the current language use of residents and to analyse it socio-linguistically with a focus on age. The project initially focused on the German-speaking language group. However, all information regarding the project, e.g. the invitation to participate, the privacy agreement, the project web site, and the questionnaire for socio-demographic data was published in German and Italian. Hence, we attracted speakers of both Italian and German. Accordingly, the collected data is multilingual, with major parts in German but with a substantial portion in Italian (100,000 of 600,000 tokens).

The collected multilingual CMC corpus combines Facebook status updates, comments, and private messages with socio-demographic data (e.g. language biography, internet usage habits, and general parameters like age, gender, level of education) of the writers. The data was enriched with linguistic annotations on thread, text and token level including language-specific part-

[1] For further information see www.eurac.edu/didi.

of-speech (PoS) and lemma information, normalisation, and language identification.

In this paper, we describe the corpus with respect to its multilingual characteristics and give special emphasis to the Italian part of the corpus to which we added manually corrected PoS annotations. Hence, it presents a continuation of Frey et al. (2015) which was restricted to German texts of the corpus, not taking into account the full variety of data collected for the total corpus.

2 Corpus Construction

For the purpose of the DiDi project, we collected language data from social networking sites (SNS) and combined it with socio-demographic data about the writers obtained from a questionnaire. We chose to collect data from Facebook as this SNS is well known in South Tyrol, hosts a wide variety of different communication settings, and is used over the whole territory by nearly all groups of the society.

Related research mainly draws on public data such as public Facebook groups, Twitter or chat data (e.g. Celli and Polonio (2013), Basile and Nissim (2013), Burghardt et al. (2016), Beißwenger (2013)), excluding the possibility to analyse discourse patterns of non-public everyday language use.

Collecting non-public and personal data for the DiDi corpus raised technical issues regarding Italian privacy regulations (which require user consent incl. privacy statement), the time-saving acquisition of authentic and complete language data, and the assignment of language data to questionnaire data. These issues have been solved by developing a Facebook application[2] that allowed for the gathering of all three sorts of data (user consent, language data, questionnaire data) at once. In addition, the application was easy to share via Facebook which helped to promote the project and to reach many potential participants. While data collection was solely managed by the Facebook application, we relied on Facebook's in-platform means (i.e. users' sharing and liking) to recruit participants. In order to reach older users (> 50 years) it was necessary to additionally resort to Facebook advertisment.[3]

With the consent of each participant, the data was downloaded via the Facebook Graph API[4] and from the used questionnaire service[5], and stored in a local MongoDB[6] data base. Both entities were linked via randomised unique identifiers. A python interface provided access points to retrieve user and text data from the data base in a linked and structured format, and also allowed to rebuild the conversational structure of threads by linking successive text objects together. This information can now be used to analyse turn-taking and language choices within threads.[7]

3 Corpus Annotations

This section describes the annotations added during the process of corpus construction.[8]

3.1 Socio-demographic Information about Participants

The corpus provides the following socio-demographic information about the participants obtained from the online questionnaire: gender, education, employment, internet communication habits, communication devices in use, internet experience, first language(s) (L1), and usage of a South Tyrolean German or Italian dialect and its particular origin.

3.2 Linguistic Annotation of Texts

The corpus was annotated on text and token level with a series of information.

- *Language identification:*
  The used languages of a text were identified in a semi-automatic approach: Firstly, using the language identification tool *langid.py* (Lui and Baldwin, 2012), and secondly, manually correcting short texts and texts with a low confidence score.

- *Tokenisation:*
  The corpus was tokenized with the Twitter tokenizer *ark-twokenize-py*[9] and subse-

[2] The source code is available at `https://bitbucket.org/commul/didi_app`.

[3] For details regarding the technical and strategical design of the data collection and methods of user recruitment see Frey et al. (2014).

[4] `https://developers.facebook.com/docs/graph-api`

[5] `http://www.objectplanet.com/opinio/`

[6] `https://www.mongodb.com/`

[7] The source code is available at `https://bitbucket.org/commul/didi_proxy`.

[8] See Frey et al. (2015) for detailed information on the anonymisation procedure and the normalisation and processing of German texts, including identification of languages and varieties.

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

quently corrected manually for non-standard language tokenisation issues.

- *Part-of-speech tagging and lemmatization:* (Corrected) tokens were annotated with PoS tags and lemma information considering the predominant language of the text at hand. We tagged Italian texts with the Italian tag set of the Universal Dependencies project[10] using the *RDR PoS Tagger* (Nguyen et al., 2014). Subsequently, we manually corrected PoS annotations to handle bad tagging accuracy for social media texts. Additionally, we used the *TreeTagger* (Schmid, 1994; Schmid, 1995) to assign PoS tags for German, English, Spanish, French and Portuguese texts applying the standard tagsets for each language. No manual correction was performed for these languages.

- *Normalisation:* So far, we have manually normalised non-standard language to word-by-word standard transcriptions only for German texts.

- *Variety of German:* We classified German texts as dialect, non-dialect or unclassifiable texts applying a heuristic approach based on the normalisation.

- *Untranslatable dialect lexemes:* We have created a lexicon for untranslatable dialect words encountered during manual normalisation. The dialect lexicon was used to post-process out-of-vocabulary (OOV) tokens in the corpus.

- *Foreign language insertions:* The most common OOV tokens that we manually classified as foreign language vocabulary have been annotated with information about their language origin.

- *CMC phenomena:* Emoticons, emojis, @mentions, hashtags, hyperlinks, and iterations of graphemes and punctuation marks were annotated automatically using regular expressions.

- *Topic of the text:* In order to investigate context factors of language choice we annotated texts as either political or non-political according to a list of politicians, political parties and political terms.

[10] http://universaldependencies.org/it/pos/index.html

3.3 Conversation-related Annotations

We rebuilt conversation threads by linking successive texts and created thread objects containing ordered lists of texts that are accessible via the Python interface. Thread objects contain information about the used languages and the number of active interlocutors and recipients of a message as well as the time passed between two texts.

As described in Frey et al. (2015), no text content of non-participants of the DiDi project was stored, but general information about the publishing time and the language of the text was kept. If all interlocutors of a thread were participants of the project, the whole conversation is available.

3.4 User-related Annotations

In addition to socio-demographic data, we added information about the users' (multilingual) communicational behaviour, i.e. their primary language, used languages and the number of interlocutors.

4 Corpus Data

4.1 Corpus Size

The DiDi corpus comprises public and non-public language data of 136 South Tyrolean Facebook users. The users could choose to provide either their Facebook wall communication (status updates and comments), their chat (i.e. private messages) communication or both. In the end, 50 people provided access to both types of data. 80 users only provided access to their Facebook wall and 6 users gave us only their chat communication. In total, the corpus consists of around 600 thousand tokens that are distributed over the text categories status updates (172 ,66 tokens), comments (94,512 tokens) and chat messages (328,796 tokens).

4.2 Multilingualism in the Corpus

The corpus is highly multilingual. Although the initial intention of the project was to document the use of German in South Tyrol, German language content comprises only 58% of the corpus. 13% are written in Italian and 4% in English (the remainder of the messages was either classified as unidentifiable language, non-language or other language). The distribution of the languages is

based on the language backgrounds of the participants and is comparable to the multilingual community of South Tyrol. The following tables show the distribution of profiles, texts and tokens (table 1) and text type (table 2) by L1.

User L1	Profiles	Texts	Tokens
IT	9	4,260	80,368
DE	108	29,883	421,262
other	3	407	8,643
IT + DE	11	4,165	75,359
DE + other	5	1,110	10,642
Total	136	39,825	596,274

Table 1: Distribution of profiles, texts and tokens by L1.

User L1	SU	CO	PM
IT	1,682	1,063	1,515
DE	7,286	4,890	17,707
other	172	45	190
IT+DE	1,962	343	2,791
DE+other	1,031	166	13
Total	11,102	6,507	22,216

Table 2: Distribution of texts by text type (SU = status updated, CO = comments, PM = private messages) by L1.

While very few users wrote only in their first language, most users used at least two (88%), very often even three (73%) or more (51%) languages. Table 3 shows the number and proportion of German, Italian and English texts written as first or second/foreign language.

Text written	as L1	as L2
IT	4,761 (57%)	3,566 (42%)
DE	23,191 (99%)	170 (1%)
EN	166 (4%)	3,625 (96%)
All languages	28,120 (78%)	7,842 (22%)

Table 3: Distribution of text language by L1 or L2 use.

In terms of multilingual language use in the DiDi corpus, we observe a slight difference between Italian and German-speaking users. L1 Italian speakers stick more to their L1 compared to the German-speaking participants, who are characterized by a higher usage of L2 Italian. The comparison of L1 and L2 usage in status updates, comments and private messages (c.f. Table 4) shows that the respective L1 is preferred in all messages types. We find the highest percentage of second or foreign language use in status updates, whereas in comments and private messages around 75% of the texts are written in L1.

Text written	as L1	as L2
Status updates	6,774 (61%)	3,032 (27%)
Comments	5,089 (78%)	924 (14%)
Messages	16,257 (73%)	3,886 (17%)
Total	28,120 (71%)	7,842 (20%)

Table 4: Distribution of L1 and L2 use by text types.

Finally, we observed 4,295 code-switching instances on conversation level and at least 1,653 texts that contain multiple languages[11]. The average number of code-switching instances per user is 10%, meaning that every tenth text does not continue the language of the previous text in the thread (the maximum was around every second text, i.e. 42%). The average proportion of text with multiple languages per user is 4% (max. 25%).

5 Issues in Corpus Creation

In addition to general issues of working with social media texts (e.g. text processing on noisy, short texts as described for example in (Baldwin et al., 2013; Eisenstein, 2013)) , the high diversity in used languages and varieties in our corpus led to various restraints in corpus creation and processing as cross-lingual annotation and information extraction are still crucial problems in natural language processing. We tried to address the demands of a multilingual corpus by providing language specific PoS tagging and by applying language independent annotations. We are aware of the fact that this is by no means sufficient to deal with linguistic research questions that exceed language boundaries. Moreover, manual correction tasks occupied a significant part of the work on the corpus as automatic annotation (e.g. for language identification) does not yet provide the accuracy expected for linguistic studies (Carter et al., 2013; Lui and Baldwin, 2014).

[11]Texts were annotated as mixed-language texts during the correction of the language identification, therefore this annotation has not been done for the whole corpus. A further word-level identification of languages could detect even more mixed-language content(Nguyen and Dogruoz, 2013)

6 Conclusion and Future Work

In this paper we presented a freely available language corpus of Facebook user profiles from South Tyrol, Italy. The multilingual corpus is anonymised and annotated with sociodemographic data of users, language specific (and for Italian manually corrected) PoS tags, lemmas and linguistic annotations mainly related to used languages, varieties and multilingual phenomena. The corpus is accessible for querying via ANNIS[12] or can be obtained as processable data for research purposes on `http://www.eurac.edu/didi`.

Acknowledgements

The project was financed by the Provincia autonoma di Bolzano – Alto Adige, Ripartizione Diritto allo studio, università e ricerca scientifica, Legge provinciale 13 dicembre 2006, n. 14 "Ricerca e innovazione".

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164.

Manuel Burghardt, Daniel Granvogl, and Christian Wolff. 2016. Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. In *Proceedings of LREC 2016*, pages 2029–2033.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Fabio Celli and Luca Polonio. 2013. Relationships between personality and interactions in facebook. *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54.

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

Simone Ciccolone. 2010. *Lo standard tedesco in Alto Adige*. Il segno e le lettere. LED Edizioni Universitarie, Milan.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

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.

Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2015. The DiDi Corpus of South Tyrolean CMC Data. In *Workshop Proceedings of the 2nd Workshop on NLP4CMC at GSCL2015*.

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.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25, Gothenburg. Association for Computational Linguistics.

Dong-Phuong Nguyen and A Seza Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

Dat Quoc Nguyen, Dang Duc Pham Dai Quoc Nguyen, and Son Bao Pham. 2014. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49.

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