eurac research

DiDi Corpus

French-German colloquium on standards for corpora of computer-mediated communication

egon w. stemle
<egon.stemle@eurac.edu>

Overview of the DiDi Project (facts)

- The goal of the regionally funded project 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 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 collected data is multilingual, with major parts in German but with parts in Italian (and some other languages)
- The corpus is accessible for querying via ANNIS and *will*[1] be made available as processable data for research purposes

[1] likely, in TEI format

Overview of the DiDi Project (figures)

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: Distribution of profiles, texts and tokens by 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: Distribution of L1 and L2 use by text types.

Language technology / Linguistic annotations I

- Which types of annotations are included in the resource?

p	layer for paragraphs (using '\n' from original texts)
token	original token layer
normtok	(orthographically) normalised token layer
lemma	layer for lemma annotation based on the normalised token layer
pos	layer for part-of-speech annotation based on the normalised token layer
cmc	layer for CMC phenomena (cf. Description of the annotations on the layer cmc)
stir	layer that marks South Tyrolean dialect lemmas

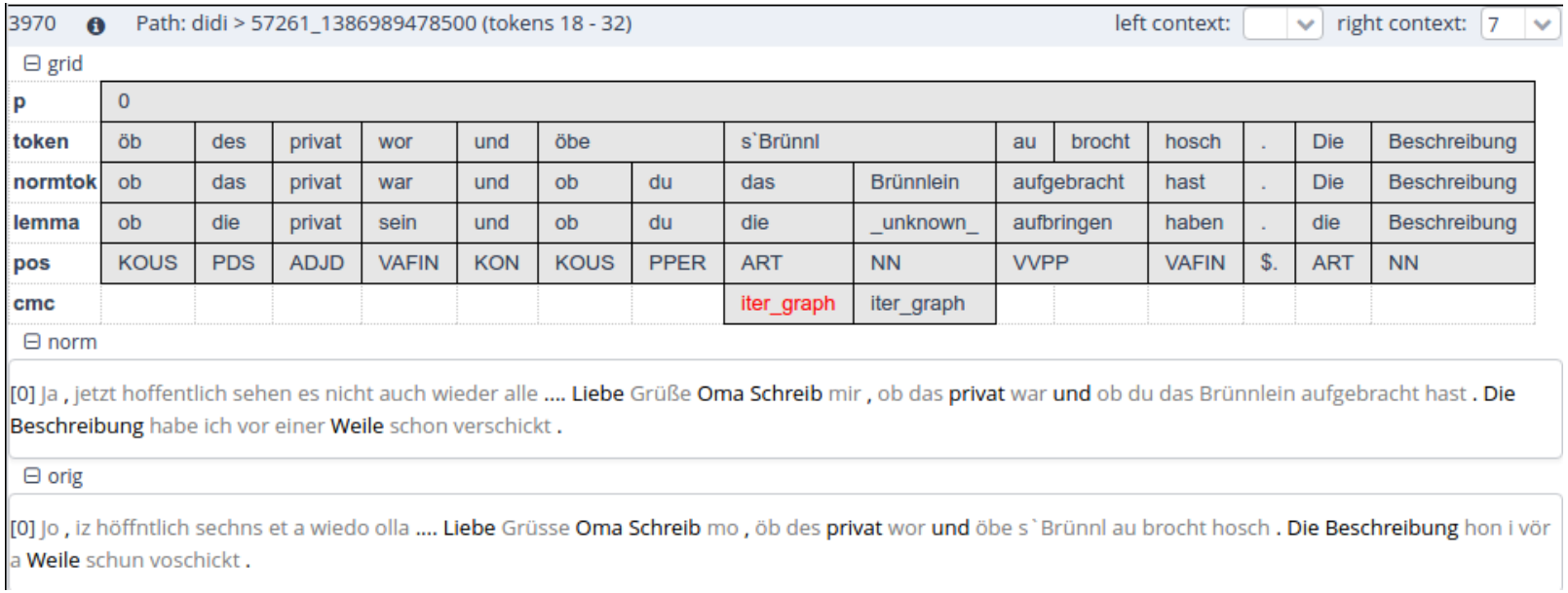
Language technology / Linguistic annotations II

- Which tagsets/annotation schemas are used?
  - Adapted version of Ruef and Ueberwasser (2013) for normalisation (of German)
  - STTS (for German pos), UD (for Italian pos)
  - Project specific (but available) guidelines for anonymisation
- How are the annotations represented in the resource?
  - python data-structures with API, json dump, and XML intermediate format (for transformation into relANNIS)

Language technology / Linguistic annotations III

```
<text><text_meta user_id="57261" CompletedDate="2014.05.08␣2:21" PA_Geschlecht="w" ... />
  <p p="0">
    ...
    <norm><nword lemma="und" normtok="und" pos="KON">
        <orig><oword token="und">und</oword></orig>
    </nword></norm>
    <norm><norm_meta contains_normalisation="true" norm_toks="ob␣du" orig_toks="&#xF6;be">
        <orig><oword token="&#xF6;be">
            <nword lemma="ob" normtok="ob" pos="KOUS"></nword>
            <nword lemma="du" normtok="du" pos="PPER"></nword>
        </oword></orig>
    </norm_meta></norm>
    <norm><norm_meta contains_normalisation="true" norm_toks="das␣Br&#xFC;nnlein" orig_toks="s'
        Br&#xFC;nnl">
        <orig><oword token="s'Br&#xFC;nnl">
            <nword cmc="iter_graph" lemma="die" normtok="das" pos="ART"></nword>
            <nword cmc="iter_graph" lemma="_unknown_" normtok="Br&#xFC;nnlein" pos="NN">Br&#xFC
                ;nnlein</nword>
        </oword></orig>
    </norm_meta></norm>
    <norm><norm_meta contains_normalisation="true" norm_toks="aufgebracht" orig_toks="au␣brocht
        ">
        <nword lemma="aufbringen" normtok="aufgebracht" pos="VVPP">
          <orig><oword token="au"></oword></orig>
          <orig><oword token="brocht"></oword></orig>
        </nword>
    </norm_meta></norm>
    ...
```

Language technology / Linguistic annotations IV



3970 Path: didi > 57261_1386989478500 (tokens 18 - 32) left context: right context: 7

⊟ grid

p	0													
token	öb	des	privat	wor	und	öbe		s`Brünnl		au	brocht	hosch	.	Die	Beschreibung
normtok	ob	das	privat	war	und	ob	du	das	Brünnlein	aufgebracht		hast	.	Die	Beschreibung
lemma	ob	die	privat	sein	und	ob	du	die	_unknown_	aufbringen		haben	.	die	Beschreibung
pos	KOUS	PDS	ADJD	VAFIN	KON	KOUS	PPER	ART	NN	VVPP		VAFIN	$.	ART	NN
cmc								iter_graph	iter_graph						

⊟ norm

[0] Ja , jetzt hoffentlich sehen es nicht auch wieder alle .... **Liebe** Grüße **Oma Schreib** mir , ob das **privat** war **und** ob du das Brünnlein aufgebracht hast . **Die Beschreibung** habe ich vor einer **Weile** schon verschickt .

⊟ orig

[0] Jo , iz höffntlich sechns et a wiedo olla .... **Liebe** Grüsse **Oma Schreib** mo , öb des **privat** wor **und** öbe s `Brünnl au brocht hosch . **Die Beschreibung** hon i vör a **Weile** schun voschickt .

Anonymization

In the DiDi Corpus, all words and sequences of word were anonymised that could reveal the identity of the authors of the text or other people that the text was related to. Public figures (politicians, sportsperson, artists, etc.), and neutral mentions of places and institutions were not included in the process of anonymisation.

Description of the anonymisation tags.

Metadata

Metadata in the DiDi Corpus is divided in three categories:

1. Sociolinguistic metadata in the DiDi Corpus, subdivided into two groups:
   (1) Personal details of the authors of the texts
   (2) Information about habits and activities of the authors with respect to new media
2. Metadata for users: from Facebook or Opinio (survey platform)
3. Metadata for texts: from Facebook or project annotations

Description of the metadata.

References I

Aivars Glaznieks, Jennifer-Carmen Frey, and Egon Stemle. *DiDi Project Homepage*. URL: http://www.eurac.edu/didi (visited on 07/01/2017).