The DiDi Project:

Collecting, Annotating, and Analysing South Tyrolean CMC Data

ird-cmc-rennes: Social Media and CMC Corpora for the eHumanities

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

Institute for Specialised Communication and Multilingualism
EURAC, Bozen/Bolzano, Italy

EURAC research

October 23rd, 2015

Project Team

EURAC research

- Aivars Glaznieks
  Writing research, Language acquisition, Corpus linguistics, Sociolinguistics
- Jennifer-Carmen Frey
  Technology Enhanced Learning and Media Literacy, Language Acquisition, Corpus Linguistics, Sociolinguistics
- Myself
  Cognitive Science, Computational Linguistics, Artificial Intelligence
- Sabrina Galasso, Nicole Stuckey
  interns (for around 3 months, each)

Outline

EURAC research

1. The Project DiDi
2. Data Acquisition
3. Data Annotation, Processing and Access
4. Outlook

The Project DiDi I

- DiDi: Digital Natives and Digital Immigrants
  Writing on social network sites: a corpus-based observation of the current language use in South Tyrol, with particular consideration of the writers' age.
- Period: 01.06.2013 – 31.07.2015

AUTONOME PROVINZ BOZEN SÜDTIROL
PROVINCIA AUTONOMA DI BOLZANO ALTO ADIGE

The project was financed by *the Autonomous Province of Bozen/Bolzano, South Tyrol, Department for the Promotion of Education, Universities and Research, Provincial Law 13.Dec.2006, No. 14 'Research and Innovation'.*

The Project DiDi II

EURAC research

Research Questions

1. How do South Tyrolean users of SNS (with L1 German) use German in quasi-public and private communication?
2. How do they use other languages?
3. Are there differences in language use that can be explained by age?
   numerical age (young vs. old) versus digital age (novice vs. experienced user)?

Finding Answers

1. Construct a corpus of South Tyrolean language use (and sociolinguistic metadata)
2. Analyse the data

Data Acquisition – Forecast

EURAC research

Data Acquisition – Considerations

EURAC research

Data and more Data

What and more What

- SMS, e-mail, chat room, Face- book, Twitter, online forum
- meta and sociolinguistic data

Participants

Pick, Reach, Convince

- selection criteria
- print, phone, radio/TV, online
- kind of incentive

Legal and Ethical Issues

- legal situation with (private) data
- adhere to ethical standards

The DiDi Facebook App and Online Survey

Landing Page

didi

Digital Natives und Digital Immigrants

Wie schreibt Südtirol auf Facebook?

EURAC research

italiano deutsch

Im Rahmen unseres Projektes DiDi wollen wir Forscherinnen und Forscher des Eurac-Instituts für Fachkommunikation und Mehrsprachigkeit die Sprachgewohnheiten der Südtirolerinnen und Südtiroler auf Facebook untersuchen.

Besonders interessiert uns dabei...

- Verwendung von unterschiedlichen Dialekten Südtirols
- Verwendung von Deutsch, Italienisch und anderen Sprachen
- Schreibstil von jüngeren und älteren Usern
- Verwendung von Smileys, Abkürzungen, etc.
- Groß- und Kleinschreibung
- andere Stilmittel

Dafür benötigen wir Südtirolerinnen und Südtiroler, die uns ihre Facebook-Texte (natürlich vollständig anonym) zur Verfügung stellen und bereit sind, ein paar Fragen zu ihren Nutzungsgewohnheiten im Internet zu beantworten. Das Ausfüllen des Fragebogens wird max. 10 Minuten dauern.

Die Ergebnisse der Untersuchung werden ab Mai 2015 auf www.eurac.edu/didi veröffentlicht.

Weitere Informationen
Möchtest du mehr über das Projekt erfahren? Besuche unsere Webseite!

Um dich an der Studie zu beteiligen, wähle hier eine oder mehrere Optionen. Sobald du deine Auswahl auf Facebook bestätigt hast, wirst du automatisch zum Fragebogen weitergeleitet.

Für die Untersuchung möchte ich folgende Daten aus dem Jahr 2013 zur Verfügung stellen:

- ☐ Pinnwandeinträge
- ☐ eigene Mitteilungen aus meiner Inbox

Teilnehmen

Verwendung der Daten:
Mit der Teilnahme an der Untersuchung werden einmalig deine in Facebook gespeicherten Sprachdaten (je nach Auswahl Pinnwandeinträge und/oder Inbox-Messages) des Jahres 2013 abgerufen. Dabei lesen wir nur deine eigenen Beiträge aus sowie die Anzahl und sprachliche Verteilung der Folgekommentare. Beiträge deiner Freunde werden nicht ausgelesen. Die Daten werden nur für wissenschaftliche Zwecke verwendet. Es geht uns dabei ausschließlich darum, die Sprache und den Stil der Kommunikation zu untersuchen, der Inhalt der Gespräche wird hierbei außer Acht gelassen. Die Daten werden zudem großteils automatisch mit Hilfe spezieller computerlinguistischer Software ausgewertet.
(Mehr dazu in unseren Datenschutzbestimmungen.)

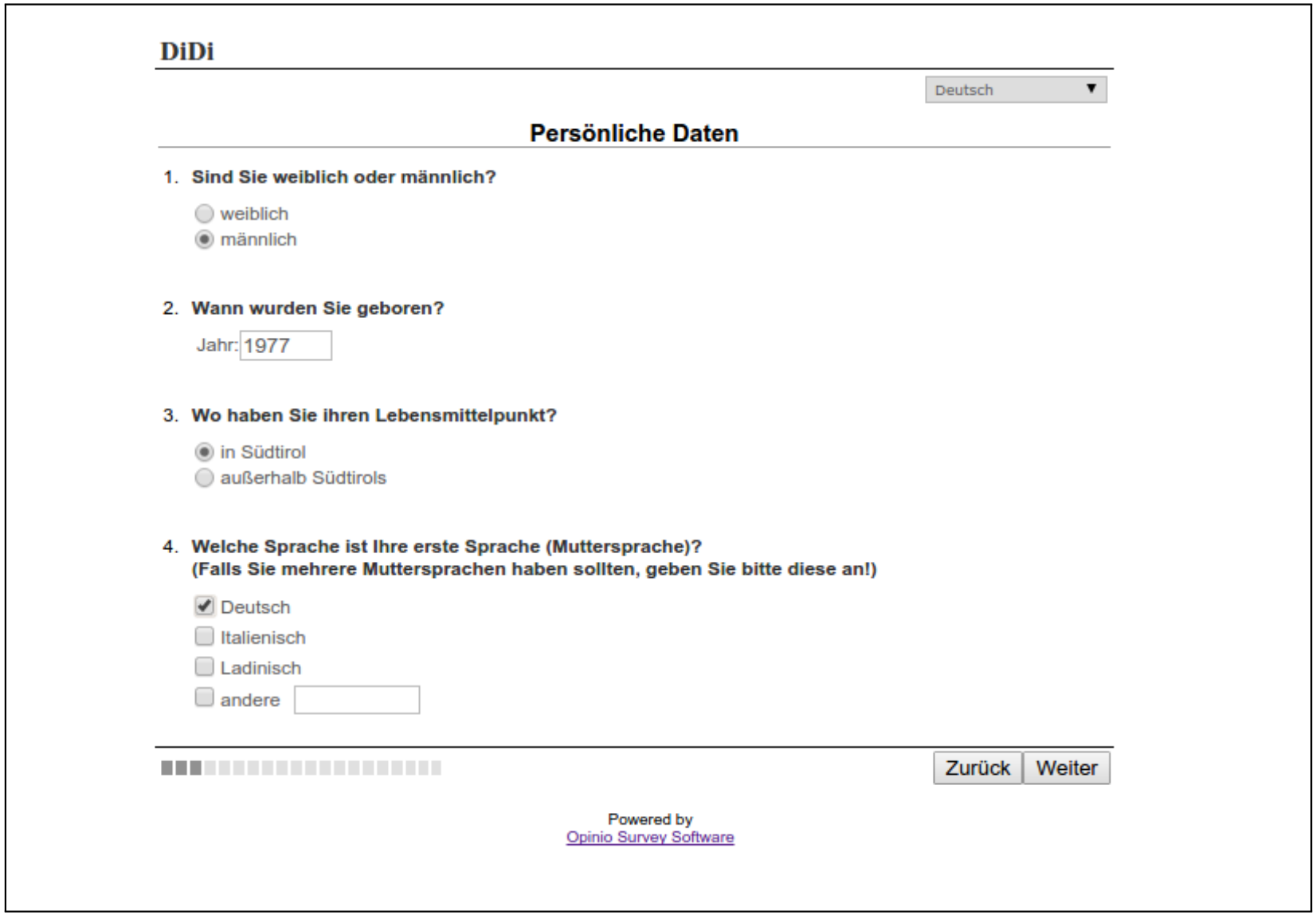
Anonymisierung der erhobenen Daten:
Die Datenerhebung erfolgt vollständig anonym. Allen Teilnehmenden am Forschungsprojekt wird eine ID-Nummer zugewiesen, die keinen Rückschluss auf die einzelnen Teilnehmenden erlaubt. Alle Daten werden mit Hilfe der ID-Nummer in anonymisierter Form gespeichert und für weitere Projektschritte zur Verfügung gestellt: In den Sprachdaten werden alle Hinweise auf die Verfasserin bzw. den Verfasser/ und andere Personen (d.h. Personen- und Ortsnamen, E-Mail-Adressen und andere Kontaktdaten) anonymisiert (d.h. durch willkürliche Namen ersetzt). Im Fragebogen wird weder der Name, das Geburtsdatum oder der Geburtsort noch der genaue Wohnort erfragt, eine eindeutige Identifizierung der teilnehmenden Person ist damit nicht möglich.

Das Projekt wird finanziert von der Autonomen Provinz Bozen - Südtirol, Abteilung Bildungsförderung, Universität und Forschung, Landesgesetz vom 13. Dezember 2006, Nr. 14 "Forschung und Innovation".

AUTONOME PROVINZ BOZEN SÜDTIROL
PROVINCIA AUTONOMA DI BOLZANO ALTO ADIGE

The DiDi Online Survey

Qs about Gender, Age, 'Residence', L1



DiDi

Deutsch

Persönliche Daten

1. **Sind Sie weiblich oder männlich?**
   - ○ weiblich
   - ◉ männlich

2. **Wann wurden Sie geboren?**
   - Jahr: 1977

3. **Wo haben Sie ihren Lebensmittelpunkt?**
   - ◉ in Südtirol
   - ○ außerhalb Südtirols

4. **Welche Sprache ist Ihre erste Sprache (Muttersprache)?**
   (Falls Sie mehrere Muttersprachen haben sollten, geben Sie bitte diese an!)
   - ☑ Deutsch
   - ☐ Italienisch
   - ☐ Ladinisch
   - ☐ andere

Zurück | Weiter

Powered by
Opinio Survey Software

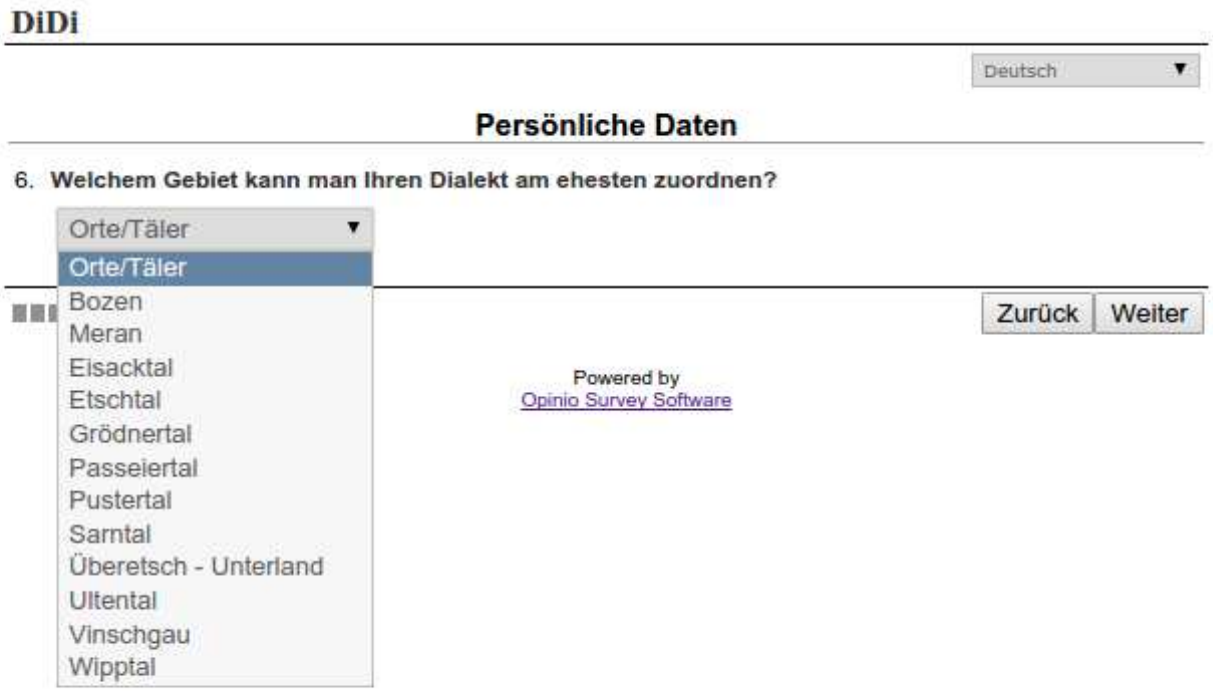
The DiDi Online Survey

Q about Language Variety

The DiDi Online Survey

Q about Localisation of Language Variety

EURAC research



The DiDi Facebook App and Online Survey

Leaving Page

EURAC research

Solbai, Ziofungo, und *lol*...

Vielen Dank für die Teilnahme!

Aktuelle Informationen zum Projektverlauf finden sich auf www.erurac.edu/didi. Dort werden ab Mai 2015 auch die Ergebnisse des Projekts veröffentlicht.

Südtirolerinnen und Südtiroler haben bereits teilgenommen!

Bitte hilf uns die Didi-App in Südtirol bekannt zu machen!

Das Gelingen unseres neuartigen Projekts die Sprache auf Facebook in Südtirol festzustellen hängt stark davon ab, dass wir die Daten die wir von dir bekommen haben auch mit denen von anderen vergleichen können. Dabei ist es wichtig, dass so viele Südtirolerinnen und Südtiroler wie möglich am Projekt teilnehmen. Du kannst uns dabei helfen und das Projekt deinen Bekannten empfehlen Z.B. indem du das App auf deiner Facebook-Pinnwand teilst!

Teile die Didi-App mit deinen Freunden!

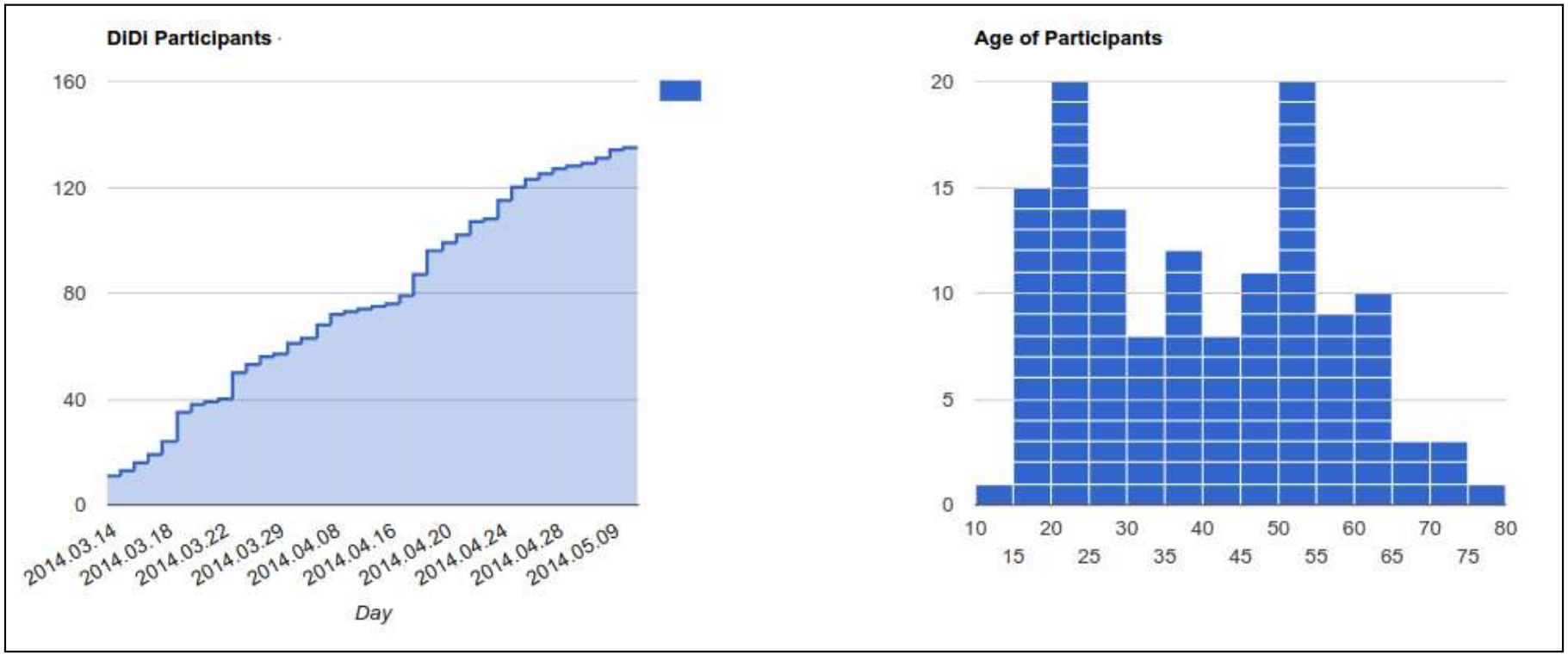
The DiDi Facebook App

Facebook Wall Post and Comments

Monitoring of Data Acquisition Progress

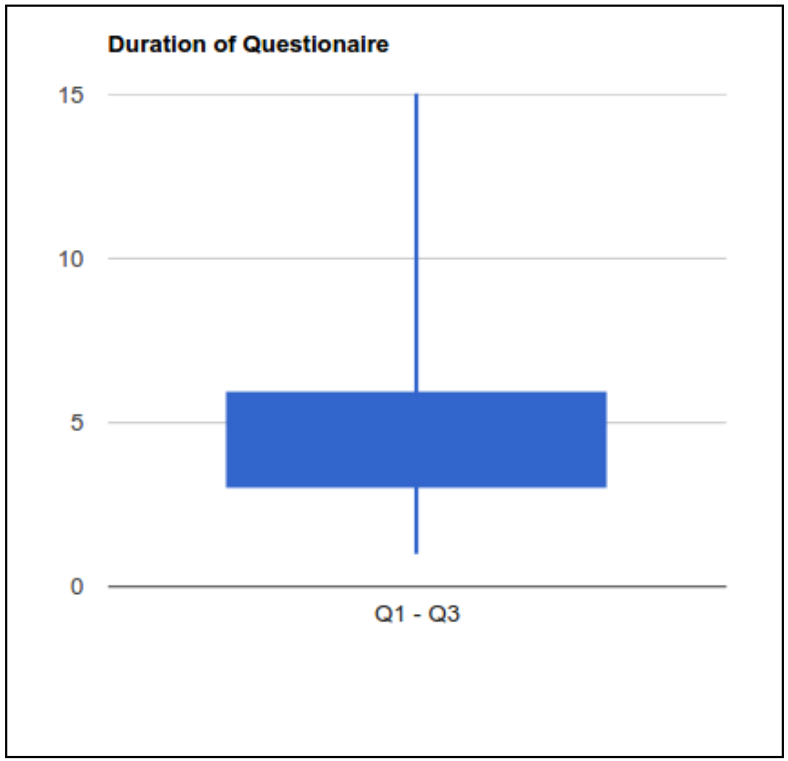
Number of Participants and their Age

EURAC research



Monitoring of Data Acquisition Progress

Sanity Check: Duration of the Questionnaire



Data Annotation and Processing I

EURAC research

Automatic Processing

Tokenisation *twokenizer*
POS Tagging, Lemmatisation *IMS TreeTagger*

Manual Annotation

Normalisation

WHY: corpus search, POS tagging (pretest: 0.5 -> 0.9), dialect identification

Anonymisation

indicate and classify issues in text

POS Corrections

(most frequent) OOVs, South Tyrolean lexemes

Semi-Automatic Post-Processing

Anonymisation

remove metadata (ids, attachments) and other users' data

CMC Phenomena

emoticons, @mentions, iterations, . . .

Language and Dialect ID

langid.py+manual post-processing, heuristic-driven dialect annotation

POS Corrections

CMC, anonymised tokens

DiDi Annotation ‘Interface’

EURAC research

It seems that perfection is reached not when there is nothing left to add, but when there is nothing left to take away.

A. de Saint Exupéry

original	anonym	token_corr	comments	norm1	norm2
<comment id=1234_4405601>					
3		3,5			
,		__			
5m		__			
		m			
sind					
die					
Betonsäulen					
Rudi	p				
</comment>					
<comment id=2345_4407568>					
...					
sel				_stir	
muasch				musst	
grod				gerade	
du					
sogn				sagen	
Issi	p				
!!!!					
</comment>					

DiDi Meets ANNIS

Document View

EURAC research

1 / 135 Displaying Results 1 - 10 of 1350 Result for: stir=/.*/

4 Path: didi > 54625_10201437841060438 (normtok 3 - 17) left context: right context: 7

grid

p	0					2									
token	Südtirol	Online	(	@stol_it	):	"	wenn	lei	die	Hälfte	von	die	Versprechen	wird	wohr
normtok	Südtirol	Online	(	@stol_it	):	"	wenn	lei	die	Hälfte	von	den	Versprechen	wird	wahr
lemma	Südtirol	_unknown_	(	_unknown_	_unknown_	"	wenn	lei	die	Hälfte	von	die	Versprechen	werden	wahr
pos	NE	NE	$(	NN	XY	$(	KOUS	FM	ART	NN	APPR	ART	NN	VAFIN	ADJD
stir								True							
cmc				at_organisation	emoticon										

norm

[0] Retweeted Stol Südtirol Online (@stol_it):
[2] " wenn lei die Hälfte von den Versprechen wird wahr , geht es uns gut in den nächsten 5 Jahren " - Messner-Windschnur #Wahlkampfblue ...

orig

[0] Retweeted Stol Südtirol Online (@stol_it):
[2] " wenn lei die Hälfte von die Versprechen wird wohr , geats ins guat in die nächsten 5 Johr " - Messner-Windschnur #Wahlkampfblue ...

DiDi Meets ANNIS

Metadata

EURAC research

nfo for salt:/didi/54625_10201437841060438 + ×

Metadata

document: 54625_10201437841060438

Name	Value
PA_Alter	36
PA_Ausbildungsabschluss	Oberschule mit Matura
PA_Beruf	ArbeitnehmerIn
PA_Dialektsprecher_ITA	no data
PA_Dialektsprecher_STIR	Bozen
PA_Geburtsjahr	1977
PA_Geschlecht	m
PA_L1_Deutsch	1
PA_L1_Italienisch	0
PA_L1_Ladinisch	0
PA_L1_andere	0
PA_Lebensmittelpunkt_STIR	1
PA_Schule	no data
RespondentId	54625
StartDate	2014.03.14 12:24
ZA_Frequenz_Blog	min. 1x pro Monat

DiDi Meets ANNIS

Analysis

New Analysis

Showing historgram of top 100 results, see table below for complete dataset.

linear scale

$\log_{10}$ scale

Download as CSV

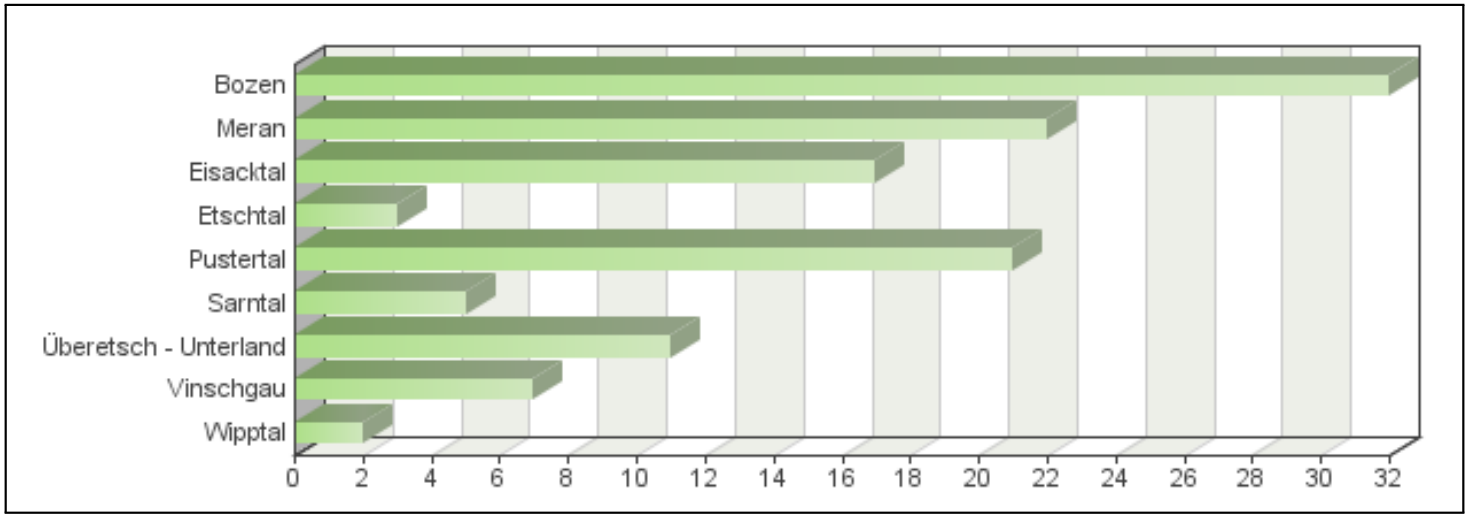
1505 items with a total sum of 75724 (query on didi)

rank	#1 | normtok	count
1	(y)	56,218
2	;)	1,885
3	:-)	1,668
4	....	1,652
5	<3	1,064
6	;-)	1,056
7	_link_	1,001
8	:D	820
9	;)	762
10	^^	574
11	.....	522

Outlook I

EURAC research

Can we find known features of spoken South Tyrolean German in our CMC data?



- Lexis
  nicht (not) [net] or [nit] or [it, et] (W to E)
- Vocalism
  -er as [r], [ər] or [o] (W to E)
  ei as [ua] or [oa] (W to E)
  ge- as [gə], [gi] or [ga] (N to S)

Outlook II

EURAC research

- Follow-up research project
  “Factors of Language Choice and Language Change”
- Use revised STTS POS tag set
- Export corpus in TEI format
- Further data analysis

Project Homepage

http://www.eurac.edu/didi/

ANNIS Corpus Interface

http://commul.eurac.edu/annis/didi/

(selected) Publications I

EURAC research

Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle.
The DiDi Corpus of South Tyrolean CMC Data.
In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media at GSCL2015 (NLP4CMC2015)*, Essen, September 2015. German Society for Computational Linguistics & Language Technology.
https://sites.google.com/site/nlp4cmc2015/NLP4CMC-2015.pdf.

Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks.
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 2014. Universitatsverlag Hildesheim, Germany.
http://www.uni-hildesheim.de/konvens2014/data/konvens2014-workshop-proceedings.pdf.

Aivars Glaznieks and Egon Stemle.
Challenges of building a CMC corpus for analyzing writer's style by age: The DiDi project.
Journal for Language Technology and Computational Linguistics (JLCL), 29(2):31–57, December 2014.
http://www.jlcl.org/2014_Heft2/2GlazniecksStemle.pdf.