

# eurac research

## CLARIN C-Centre

Jennifer-Carmen Frey, PhD  
Institute for Applied Linguistics



# The Eurac Research CLARIN Centre

## **Autumn 2017:**

CLARIN-IT membership application

## **Spring 2018:**

CLARIN C-Centre

## **Summer 2021:**

CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media Corpora (CKCMC)

## **Ongoing:**

B-Centre certification



## **Responsible people**

**Dr. Andrea Abel**  
Head of Institute

**Egon W. Stemle**  
Researcher

# Deposited resources at the ERCC

non-standard language corpora (learner corpora, corpora of student writing, social media and computer-mediated communication corpora, dialectal corpora...)

## Learner and student writing

- [Beldeko Summary Corpus v1.0.0](#)
- [KoKo German L1 Learner Corpus](#) (v1, v2 & v3)
- [Kolipsi-1 Corpus v1.0](#)
- [Kolipsi-2 Corpus v1.0](#)
- [LEONIDE - Longitudinal Learner Corpus in Italiano, Deutsch and English 1.1](#)
- [MERLIN Written Learner Corpus for Czech, German, Italian](#) (1.0 & 1.1)

## Web and CMC

- [DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0](#)
- [KrdWrd CANOLA Corpus](#) (1.0 & 1.1)
- [PAISÀ Corpus of Italian Web Text](#)

## Other

- [ACTER \(Annotated Corpora for Term Extraction Research\)](#) (v1.3 & v1.4)
- [VinKo \(Varieties in Contact\) Corpus](#)

# Deposited resources at the ERCC

non-standard language corpora (learner corpora, corpora of student writing, social media and computer-mediated communication corpora, dialectal corpora...)

## Learner and student writing

- [Beldeko Summary Corpus v1.0.0](#) external
- [KoKo German L1 Learner Corpus](#) (v1, v2 & v3)
- [Kolipsi-1 Corpus v1.0](#)
- [Kolipsi-2 Corpus v1.0](#)
- [LEONIDE - Longitudinal Learner Corpus in Italiano, Deutsch and English 1.1](#)
- [MERLIN Written Learner Corpus for Czech, German, Italian](#) (1.0 & 1.1)

## Web and CMC

- [DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0](#)
- [KrdWrd CANOLA Corpus](#) (1.0 & 1.1)
- [PAISÀ Corpus of Italian Web Text](#)

## Other

- [ACTER \(Annotated Corpora for Term Extraction Research\)](#) (v1.3 & v1.4) external
- [VinKo \(Varieties in Contact\) Corpus](#) external

# The DiDi corpus of South Tyrolean CMC

Collecting and sharing user-generated, sensitive data from Facebook

# Project DiDi



**Digital Natives – 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**

Project duration: June 2013 - June 2015

Aims:

- Document language use in digital everyday communication in South Tyrol
- Analyse choice of language and variety



Document = Collect and share!

# Background – South Tyrol

## Local setting:

- Culture with 2 official minorities and one official majority group that is however the local minority
  - diverse and potentially delicate structures in social networks of people
  - importance of identity
  - culture and identity expressed in language and in non-institutional personal communication with own social network
  - language choice as social statement
- Facebook as main platform for digital everyday communication, in ST
  - > related to identity work

# Background

## The not so straightforward data types...

- personal data
- private data
- sensitive data
- potentially harmful/compromising data

before GDPR (2013, GDPR since 2018)

Nevertheless, legal restrictions and ethical considerations

- What do we owe our data donors? How would we want our data to be used?

# Data collection in DiDi

## Facebook App

Web application with  
integration to Facebook API to  
retrieve socio-demographic  
metadata and access tokens

## Querying Facebook API

Backend script to access and  
download Facebook data with  
access tokens from users



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 [ISCM](#) der EURAC 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
- u.Ä.

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](http://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

# Legal restrictions and beyond...

- Usage consent
- Terms of licence, privacy policy

**EURAC**  
research



INSTITUTE FOR SPECIALISED COMMUNICATION AND MULTILINGUALISM  
INSTITUT FÜR FACHKOMMUNIKATION UND MEHRSPRACHIGKEIT  
ISTITUTO DI COMUNICAZIONE SPECIALISTICA E PLURILINGUISMO

DRUSUSALLEE/ VIALE DRUSO 1  
39100 BOZEN/ BOLZANO (BZ)

TEL: +39 0471 055 111  
FAX: +39 0471 055 199

communication.multilingualism@eurac.edu  
www.eurac.edu

#### Aufklärung über die Verwendung Ihrer persönlichen Daten

Gemäß Art. 13 GvD Nr. 196/03 (Decreto Legislativo del 30 giugno 2003, n. 196) informieren wir Sie, dass Ihre Daten zum Zwecke der Durchführung eines Forschungsprojektes (Durchführung: Europäische Akademie Bozen-EURAC) verarbeitet werden.

1. Ziel des Forschungsprojekts:  
Dieses Forschungsprojekt mit der Bezeichnung *Digital Natives und Digital Immigrants (DIDI)* sieht vor, die Sprachverwendung von Südtiroler Internetnutzern auf *Social Network Sites (SNS)* zu beobachten, zu dokumentieren und zu analysieren. Das Hauptaugenmerk der Untersuchung liegt auf der Frage, inwiefern das Alter einen Einfluss auf die Verwendung des Deutschen (in seiner Standard- sowie dialektalen Variante) in geschriebener Form hat. Die Ergebnisse der Studie sollen helfen, die schriftliche Sprachverwendung abseits institutionalisierter und redaktionell bearbeiteter Texte zu dokumentieren und Besonderheiten der alltäglichen normungebundenen Verwendung des Deutschen (in seiner Standard- sowie dialektalen Variante) in Südtirol herauszuarbeiten.

Um das Forschungsprojekt durchführen zu können, werden a) Sprachdaten und b) personenbezogene Daten von freiwilligen TeilnehmerInnen am Forschungsprojekt gesammelt.

- Die Sprachdaten (d.h. auf SNS von der/dem TeilnehmerIn selbst verfasste Pinnwandeinträge und eigene Mitteilungen in der Inbox) werden von den TeilnehmerInnen freiwillig zu Verfügung gestellt. Die Sprachdaten stammen aus dem Zeitraum vom 01.01.2013 - 31.12.2013.
- Die personenbezogenen Daten (Angaben zum Alter, Geschlecht und zu Nutzungsgewohnheiten von SNS) werden über einen Online-Fragebogen erhoben.

#### 2. Datenerhebung:

- Die Sprachdaten werden über eine Applikation gesammelt, die die Sprachdaten von der Facebook-Seite der/des TeilnehmerIn/Teilnehmers ausliest, das heißt, es wird eine Kopie der Sprachdaten auf einem Server der EURAC gespeichert. Die/der TeilnehmerIn muss ihre/seine Einwilligung dazu geben, dass die Applikation auf ihre/seine Sprachdaten zugreifen darf. Die/der TeilnehmerIn definiert selbst, auf welche Sprachdaten die Applikation zugreifen darf (Pinnwandeinträge und/oder Konversationen aus der Inbox). Es werden nur die Sprachdaten aus dem Jahr 2013 ausgelesen. Die/der TeilnehmerIn nimmt am Forschungsprojekt teil, sobald sie/er der Facebook-Applikation den Zugriff auf ihre/seine Daten erlaubt.
- Die personenbezogenen Daten werden über einen Online-Fragebogen erhoben. Nachdem die/der TeilnehmerIn die Applikation zugelassen hat und dadurch am Forschungsprojekt teilnimmt, wird sie/er automatisch zum Online-Fragebogen umgeleitet. Die/der TeilnehmerIn muss den Online-Fragebogen vollständig ausfüllen, bevor sie/er die personenbezogenen Daten übermitteln kann.

#### 3. Anonymisierung der erhobenen Daten:

Die Datenspeicherung erfolgt vollständig anonym. Jeder/jedem TeilnehmerIn am Forschungsprojekt wird eine ID-Nummer zugewiesen, die keinen Rückschluss auf die/den TeilnehmerIn erlaubt. Alle Daten werden mit Hilfe der ID-Nummer in anonymisierter Form gespeichert und für weitere Projektschritte zu Verfügung gestellt: In den

# Legal restrictions and beyond...

- Voluntary data donors



## 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 [JSCM](#) der EURAC 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
- u.Ä.

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](http://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:

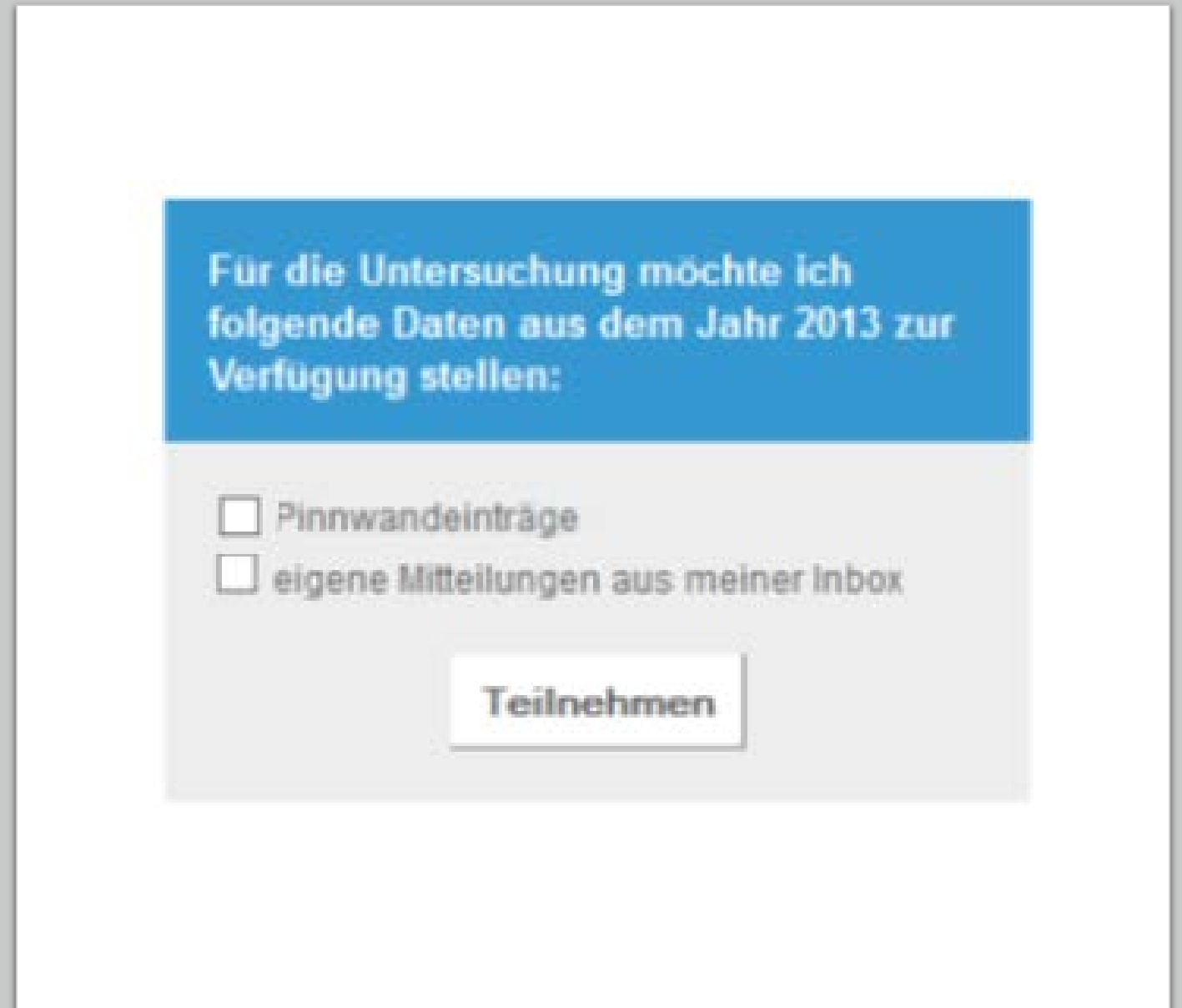
- Pinnwandeinträge
- eigene Mitteilungen aus meiner Inbox

Teilnehmen



# Legal restrictions and beyond...

- Voluntary data donors
- No pre-filled web forms



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

- Pinnwandeinträge
- eigene Mitteilungen aus meiner Inbox

**Teilnehmen**

# Legal restrictions and beyond...

- Voluntary data donors
- No pre-filled web forms
- Information about data treatment and scope of future data use

## **Verwendung der Daten:**

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 größtenteils 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.

# Legal restrictions and beyond...

- Voluntary data donors
- No pre-filled web forms
- Information about data treatment and scope of future data use
- Possibility to stop data sharing at any time of the donation workflow
- Possibility to withdraw right for data usage retrospectively



# Legal restrictions and beyond...

- Voluntary data donors
- No pre-filled web forms
- Information about data treatment and scope of future data use
- Possibility to stop data sharing at any time of the donation workflow
- Possibility to withdraw right for data usage retrospectively
- Full manual anonymisation

Auf Besuch im <InstNE>-Zeltlager <GeoNE>

vor allem für die, die in <GeoNE> studieren -  
<PersNE> <PersNE> >PersNE> =)))

<PersNE> <PersNE>: Singen, beim <InstNE>-  
Chef

A viertl Seite kostet 100 Euro + MwSt.  
Miaßasch di mit meindo Kollegin <PersNE>  
<PersNE> in Verbindung setzn, Tel. <tel>, E-  
Mail <mail>

....

# The DiDi corpus at ERCC

Eurac Research CLARIN Centre Repository Home / Search

[Advanced Search](#)

Showing 1 through 1 out of 1 results

1



Corpus

CMC & WaC

**DIDI - The DiDi Corpus of South Tyrolean CMC 1.0.0**

(Institute for Applied Linguistics, Eurac Research / 2019-03-07)

**Author(s):**

Frey, Jennifer-Carmen ; Glaznieks, Aivars ; Stemle, Egon W.

This item contains 6 files (67.06 MB).



Academic Use



Browse

> All of the Repository

My Account

Logout

Profile

Submissions

Administrative

Control Panel

Access Control

> Content Administration

**THANK YOU!**