# Herausforderungen bei der automatischen Verarbeitung von dialektalen IBK-Daten

Aivars Glaznieks & Egon W. Stemle

<{aivars.glaznieks, egon.stemle}@eurac.edu>

Institute for Specialised Communication and Multilingualism

European Academy of Bozen/Bolzano
(EURAC)

EURAC research

September 23th, 2013

## Digital Natives - Digital Immigrants

Schreiben auf Social Network Sites: Eine korpusunterstützte Sprachbeobachtung des aktuellen Sprachgebrauchs in Südtirol unter besonderer Berücksichtigung des Alters.

Wir anaysieren und untersuchen strategien der Nutzer von Online-Netzwerken aus einer auf die Provinz Südtirol eingeschränkten Perspektive. Anhand von Texten auf Social Network Sites (SNS) soll untersucht werden, wie die deutsche Sprache zu kommunikativen Zwecken in geschriebener Form verwendet wird. Das Hauptaugenmerk der Untersuchung liegt auf der Frage, inwiefern das Alter einen Einfluss auf die Verwendung des Deutschen in geschriebener Form hat.

# heut&etz: DiDi

EURAC research

## Problem

Abweichungen von der Standardschreibung und genrespezifische Elemente führen mit vorhandenen Verarbeitungswerkzeugen häufig zu unbefriedigenden Ergebnissen, weshalb die Werkzeuge eine Anpassung oder Überarbeitung, letztlich vielleicht sogar eine Neuentwicklung benötigen.

## Fragestellnug

*Wie* unbefriedigend sind denn die Ergebnisse in unserem Fall (IBK-Daten aus Südtirol), und wie wirken sich (einfache) Anpassungen aus?

## Auf zu einer Antwort

Daten von der Facebook-Seite *Spotted: Südtirol* passen wir an

&

die Fehlerrate beim POS Tagging des *IMS TreeTaggers* evaluieren wir.

## Spotted: Südtirol

About:
Es is ganz einfach! Wenn du jemanden süsses irgendwo entdeckst (schule, disko usw...), von dem du deine Augen nicht lassen kannst, dann sende dieser Fanpage eine Nachricht, und wir posten sie ganz anonym. +++ SHARE THE LOVE! ♡
Beschreibung:
Bei uns geht es um Liebe, Romantik, Spaß & Anonymität. Deswegen,

- keine herabwürdigenden, oder verletzenden Kommentare
- diese werden von uns, oder auf Hinweis gelöscht
- ebenso ganze Posts, indem ihr uns eine Nachricht sendet.

Dane für euer Verständnis. In diesem Sinne: SHARE THE LOVE ♡

# Daten


Spotted: Südtirol
September 4
Hey Leute
Meine freindin und i sain grot 14:37 af dor mebo unterweigs und mir tattn gearn wissn wear dei 5 oddor 4 typpm in den kluanen weißen VW sein.. xD
Schaugn gonz schnuckelig aus
Danke im voraus :))
See Translation
Like · Comment · Share

# Daten

**Spotted: Südtirol**
September 2

I hon gestern 1.09 pa hockey in sterzing a gruppe gitschn gsegn de sein do trainingsloger glab i und oane fa de hot an grauen pulli unkop, i hatse gerne gsechn konnmer jemand helfn?
See Translation

Like · Comment · Share

Hannes Rainer likes this. Top Comments

Write a comment...

**Lena Haselrieder** Pass sellm besser afn luki au bitte:D
See Translation
Like · Reply · 3 · September 2 at 11:20pm via mobile

**Luki Messner** Nimm wose kregn knsh Lukas Tötsch
See Translation
Like · Reply · 2 · September 2 at 11:20pm via mobile

**Luki Messner** Jo mindigshtns 3
Like · Reply · 2 · September 2 at 11:17pm via mobile

**Valentina Rier** Na enk foln sochn in
Like · Reply · 2 · September 2 at 11:14pm via mobile

**Lena Haselrieder** Wirtschoftskriese:D
See Translation
Like · Reply · 1 · September 2 at 11:18pm via mobile

**Lena Haselrieder** Zu sein geburtstog werter zuastechn;)

# Daten

**Spotted: Südtirol**
September 3

I suachat an typ der hot volle a netts bierbeichl

See Translation

Like · Comment · Share

14 people like this. Top Comments

Write a comment...

**Sabrina Ferro** fa die sellmen weards do eppr genua gebm in Südtirol!! hahahahha
See Translation
Like · Reply · 12 · September 3 at 10:45pm

**Matthäus Tratter** Wisoen a sixpack wenn a gonzes fassl hm konnsch xD xD
See Translation
Like · Reply · 7 · September 3 at 11:08pm via mobile

**Manuel Eder** Julian Tschigg
Like · Reply · 5 · September 3 at 11:00pm via mobile

**Luca De Ginger Bob** frior wors amol es sixpack jetzt isch afoamol es bierbeichl mode ^^ das man lei ollm zu spat kimp ^^
See Translation
Like · Reply · 4 · September 3 at 10:48pm

**Hannes Lamprecht** Christoph Baron Zu Kirchler
See Translation
Like · Reply · 3 · September 3 at 10:47pm via mobile

# Daten

**Spotted: Südtirol**
September 4

Mohlzeit...
I tat lei gern frogen wer olles afn Pc Call of Duty, Battlefield.... zockt. Por kollegen und i mechaten gern ba CoD an Clan mochen. Jeder Like wert ungschrieben

See Translation

Like · Comment · Share

32 people like this. Top Comments

Write a comment...

**Alexander Mair** Schreib lai aml un, war echt gnz a guate idee aml an südtiroler pc clan zu mochen!
See Translation
Like · Reply · 3 · September 5 at 12:35pm

**Patrick Trafoier** good old mw2 esl times, missing them
Like · Reply · 3 · September 4 at 11:30pm

**Philipp Fischer** Norman Lettieri Patrick Unterberger Thomas Ortler
See Translation
Like · Reply · 2 · September 5 at 1:14pm

**Philipp Fischer** bf 3+4, cod mw3
Like · Reply · 1 · September 5 at 1:13pm

**Manuel Gasser** GTA vorbestellt, griags schun in 16tn
See Translation

# Gesammelt: Daten, am 6.September

72 Messages, 231 Comments
∅ 3.2083 Cmnts/Msg, Median: 1
5520 Tokens

abzüglich 7 Comments, da komplett in ENG (2) bzw. ITA (5)

## Datenset

72 Messages, 224 Comments
∅ 3.1111 Cmnts/Msg, Median: 1
5452 Tokens

# Anpassungen

EURAC research

| | |
|---|---|
| `tok` | Tokenisierungsfehler |
| `cmpnd` | Zusammen-/Getrenntschreibung nach DUDEN |
| `pnct.` | Satzzeichen – nur Satzgrenzen |
| `pnct,` | Satzzeichen im Satz |
| `cap` | Groß-/Kleinschreibung nach DUDEN |
| `abbr` | Angabe der ausgeschriebenen Form abgekürzter Wörter |
| `trns_p` | Phonologische Entsprechung im Standarddeutschen |
| `trns_m` | Morphologische Entsprechung im Standarddeutschen |
| `trns_l` | Lexikalische Entsprechung im Standarddeutschen |

# Anpassungen: Beispiele

EURAC research

| | |
|---|---|
| cmpnd | mitn ischs griags afn |
| abbr | u v vv vlt bz |
| trns_pml | i mi mir hot isch a dor |

| tok | cmpnd | abbr | trns_pml |
|---|---|---|---|
| 37 :) | 30 n | 7 und | 107 ich |
| 33 .. | 22 s | 5 volle | 76 hat |
| 31 ;) | 13 af | 5 vielleicht | 60 ist |
| 12 :D | 9 i | 3 Bozen | 53 ein |
| 7 . | 7 mit | 2 schreiben | 43 die |
| 6 " | 7 a | 2 haha | 42 der |
| 5 ;D | 5 se | 1 Zugbahnhof | 41 eine |
| 5 ^^ | 5 r | 1 morgen | 37 einen |
| 4 :)) | 5 jo | 1 lieben | 35 das |
| 4 ;)) | 5 isch | 1 jemand | 34 war |

EURAC research

## Tokenisierung

Tokens: 5452, Messages: 72

Genauigkeit: 0.946 (bzgl. Korrektur in `tok`)

| | IMS Tok | tok | cmpnd | pnct., | cap | abbr | transl_pml |
|---|---|---|---|---|---|---|---|
| IMS Tok | 0.482 | | 0.481 | 0.515 | 0.524 | 0.486 | 0.877 |
| tok | | 0.508 | 0.507 | 0.542 | 0.553 | 0.520 | 0.922 |
| $\rightarrow$ cumulative | | | | 0.540 | 0.586 | 0.596 | 1 $\sim$ 0.94 |

## POS-Tagging

Tokens: 551, Messages:9

Genauigkeit: 0.935 (`transl_pml` bzgl. ‘manuellem’ Goldstandard)

AUTONOME PROVINZ BOZEN SÜDTIROL

PROVINCIA AUTONOMA DI BOLZANO ALTO ADIGE

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