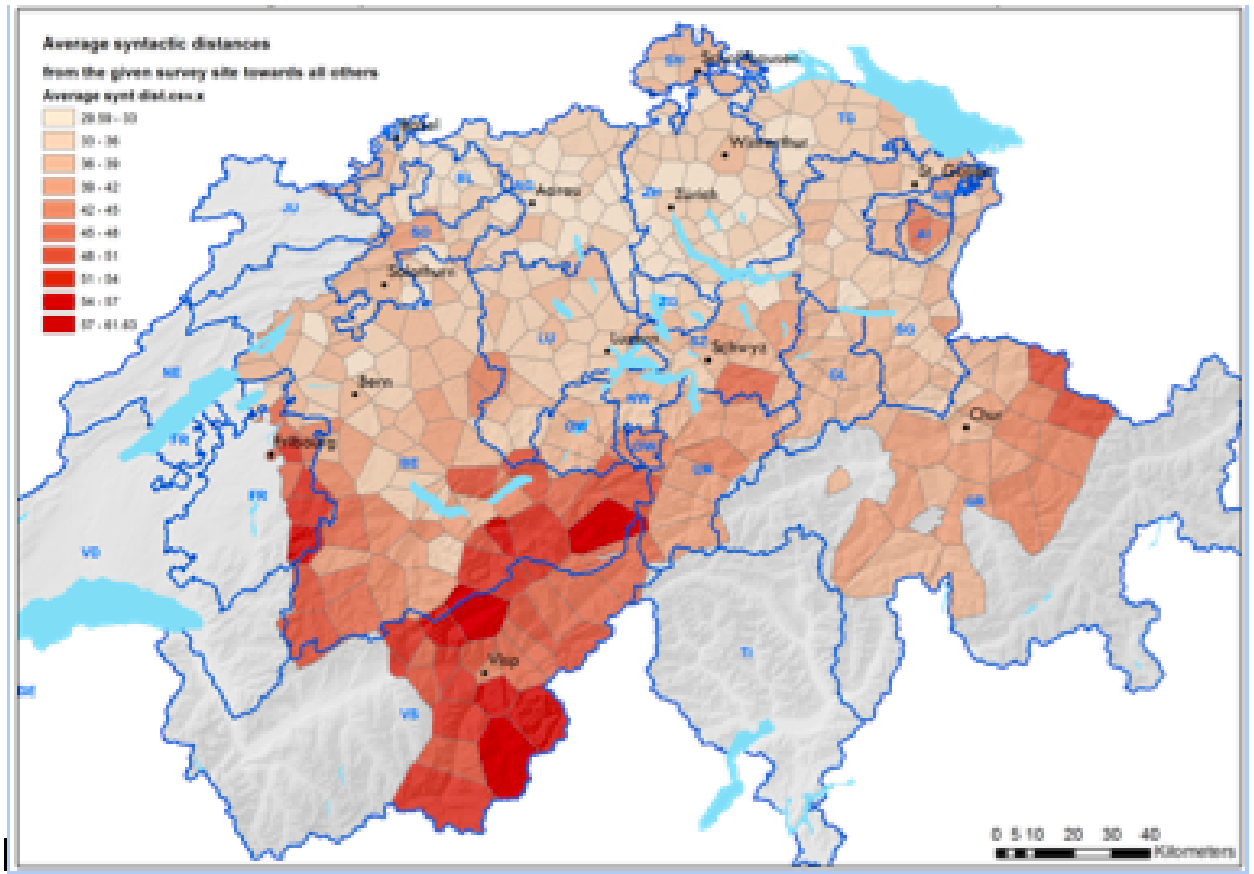
Normalising orthographic and dialectal variants for the automatic processing of Swiss German

Tanja Samardžić, Yves Scherrer, Elvira Glaser

Swiss German Dialects



Map by Péter Jeszenszky

The ArchiMob corpus

A corpus for everybody

Query mein 182 (1,611.86 per million)

Page 1 of 10 Go Next | Last

hät er dä ganz möse höore / aber / eh jez	mine	wunsch isch dän dùe noch gsii / das ich
zwaar bin ìich / i / in wolfeschiesse /	mii	vatter und mi mueter sind wolfeschiesser
chenn ich det ässen und schlaaffe daa / sait	mii	/ vatter / das isch doch härrlich de mues
dich nid aamälde schüsch märkid si as dui	mis	/ as dui / eh / äbe z jung bisch / und
(de) han ich / scheen verdienet / und	mi	vatter isch glüklich gsii und gsait pliib
het er sibeze jaar gschaffet / han ich den	min	maa isch driìssgi gsii / und ìich äbe zwaijezwänzgi
/ nuur hemmer enand halt jaare und jaare	mi	jüngscht brieder isch do achtjäärig gsii
han ich no tänkt dasch ez doch äaine wi	mii	brieder mi brieder het ganz / wiissblondi
tänkt dasch ez doch äaine wi mii brieder	mi	brieder het ganz / wiissblondi haar ghaa
chilen uis bin / dasch / glòüb ich / eh /	mi	brieder de sait er ja / het ja no vil so
und de het mer sich doch müe ggää / und	mi	vatter isch natierli gliklich gsi / i dère
jeregot mer hend gad / allernöötigisch ghaa /	min	maa het esoo nach em firaabig / eppis gschaffet
die sind den al no i de lèèr gsii / und	min	maa isch fascht echli ire maischter gsii
/ de het me gwüsst ass e chrieg chunt /	min	maa het immer gsaait / muesch de chasch
mobiilmachchung bekant gegeeben / ja / und de isch	mi	maa / am zwölfi choo dee het doo vo de
iiruke de isch i de / gliiche ainhait (wo)	min	maa / bin ich zu dèren und ha gsait duu
wäärde / isch dä kiosk offe / de isch immer	mi	maa det gsii und de (?) de bek bringt
irgend es vorigs bet oder eppis haig aso	mii	maa isch doo im dienscht / plibe / biss
und pro maa / e franken über / und daa het	mi	maa mir zurgg-gschribe sig alles schöön
e franke ghaa han ich drii franke ghaa /	mi	maa het achzg rappe sold ghaa / und de

Page 1 of 10 Go Next | Last

Goal: 700 000 words (44 docs) Currently : 500 000 words (34 docs)

The problem of variation

Other docs	mine mini määin mäin main	man	hed hèd hèt heet haa händ hüt	ime imer emmer iiimer	gsäit gsääit
Same doc	mi mii miin mis miis	ma	hat hät		gsait
Transcription	*min*	*maa*	*het*	*immer*	*gsaait*
Normalised	**mein**	**mann**	**hat**	**immer**	**gesagt**
Translation	My husband has always said...				

Table : A segment of a transcribed and normalised text with corresponding variants found in the same document and in other documents.

Word-level annotation

jaa	ja	ITJ
de	dann	ADV
het	hat	VAFIN
me	man	PIS
no	noch	ADV
gluegt	gelugt	VVPP
tänkt	gedacht	VVPP
dasch	das_ist	PDS+
ez	jetzt	ADV
de	der	ART
genneraal	general	NN
jaa	ja	ITJ
das	das	PDS
ischsch	ist	VAFIN
en	ihn	PPER
ez	jetzt	ADV

The approach

- Manual normalisation of 6 documents with VARD 2 and IGT
- Developing an automatic approach using the manually annotated sample for training and testing
- Semi-automatic processing of the remaining documents:
  1. Train the system on all the manually annotated data
  2. Process a new document
  3. Evaluate and correct the output
  4. Add the new document to the training set
  5. Repeat 1-4 for all the remaining documents

Automatic normalisation as translation

jaa	ja
de	dann
het	hat
me	man
no	noch
gluegt	gelugt
tänkt	gedacht
dasch	das_ist
ez	jetzt
de	der
genneraal	general

jaa	ja
das	das
ischsch	ist
en	ihn
ez	jetzt

The approach

The problem: little data, a lot of variation

Dividing the task:

- **Unique**: no ambiguity, only one normalisation possible
- **Ambig1**: multiple possible normalisations, one predominant
- **Ambig2**: multiple possible normalisations, none predominant
- **New**: unknown words (no information about possible normalisations)

Automatic methods

W-b-w: Assign to each test word the most probable normalisation based on word frequency

CSMT: Generate for each test word the most probable normalisation based on Character-Based Statistical Machine Translation

ALM: Rerank the candidates using the ArchiMob language model

ELM: Rerank the candidates using an extended language model (TüBa-D/S + ArchiMob)

Results

	Proportion (%)	Accuracy (%)					
		No	W-b-w	CSMT	ALM	ELM	Best
Unique	44.53	23.06	**87.83**	87.83	87.83	87.83	87.83
Ambig1	43.68	21.92	**80.63**	80.63	70.11	67.12	80.63
Ambig 2	0.52	4.17	**38.43**	**39.81**	**46.06**	**49.54**	49.54
New	11.27	3.64	3.64	**23.88**	23.88	23.88	23.88
All	100	20.28	74.94	77.23	72.67	71.38	77.28

Table : Accuracies of the automatic normalisation methods for the different word classes, averaged over the 5 folds.

Conclusions

- Combined methods are better than any individual method
- Advanced methods help with hard cases (ambiguous and new words), but more data needed
- Data from standard German do not help much