sms4science.ch: A multi-lingual challenge for Part-of-Speech tagging

ird-cmc-rennes

Simone Ueberwasser M.A.

I like you –> saumässig –> and my little –> härzli pöpperlet –> toujour –> per te! –> You are –> mon ceur, –> tu sei –> min stärn, –> I have you –> eifach –> molto –> gärn! –> Hdslmf –> din –> Hase

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

The corpus

The project

- Funded by Swiss National Science Foundation: € ~1.5 Mio
- Seven doctoral students
- Zurich, Bern, Neuchâtel, Leipzig
- Lead: Elisabeth Stark, Zurich
- www.sms4science.ch

The corpus

The data collection

- Nov 2009 – Jan 2010
- Collected in co-operation with Swisscom
- ~26,000 SMS
- ~650,000 Tokens
- More than 50% with code-switching
- Demographic questionnaires: 1,316 covering 79% of SMS
- Freely available for research

The corpus

Languages

Language	**SMS**
Swiss German dialect	10’734
Standard German	7’254
French	4’622
Italian	1’475
Romansh	1’120
English	538
⋮	⋮
Dialetto	50
Spanish	43
Patois	28

The corpus

Processing

- Anonymisation
- Language tagging (manual processing):
  - Main language contributes most tokens
  - Borrowings (established, in dictionary)
  - Nonce-borrowings (spontaneous, not in dictionary)
- Normalization (manual processing)
- PoS tagging
- (Morphosyntactic tagging)

Outline

The corpus

Necessity of normalization

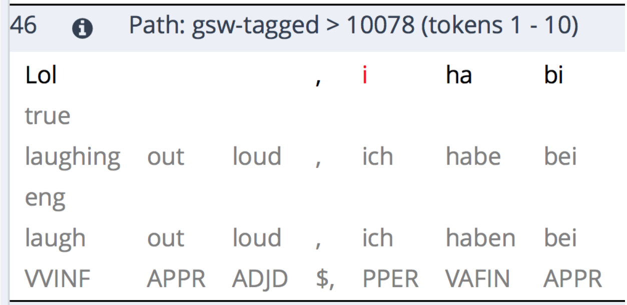
Aim of normalization

Approach

Conclusions

Outlook

Abbreviations, borrowings



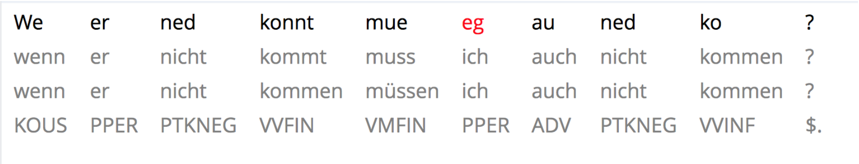
46 ℹ Path: gsw-tagged > 10078 (tokens 1 - 10)

Lol			,	i	ha	bi
true						
laughing	out	loud	,	ich	habe	bei
eng						
laugh	out	loud	,	ich	haben	bei
VVINF	APPR	ADJD	$,	PPER	VAFIN	APPR

Figure: *lol* as an abbreviation and borrowing

Dialect

Far from standard German



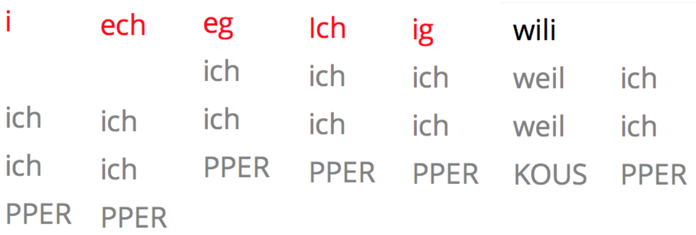
We	er	ned	konnt	mue	eg	au	ned	ko	?
wenn	er	nicht	kommt	muss	ich	auch	nicht	kommen	?
wenn	er	nicht	kommen	müssen	ich	auch	nicht	kommen	?
KOUS	PPER	PTKNEG	VVFIN	VMFIN	PPER	ADV	PTKNEG	VVINF	$.

Figure: Bern dialect

‘If he does not come, I don’t have to come either?’

Dialect without orthography

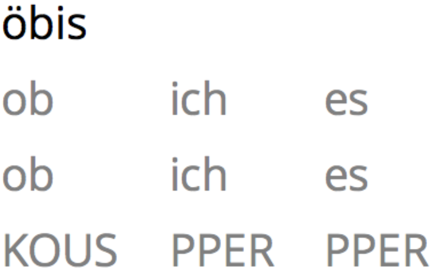
Some spelling variants of *ich*



i	ech	eg	Ich	ig	wili	
		ich	ich	ich	weil	ich
ich	ich	ich	ich	ich	weil	ich
ich	ich	PPER	PPER	PPER	KOUS	PPER
PPER	PPER					

Figure: Some spelling variants of *ich* (‘I’)

Clitics

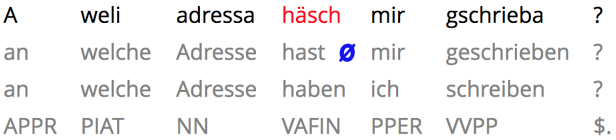


öbis		
ob	ich	es
ob	ich	es
KOUS	PPER	PPER

Figure: Double clitics

'. . . whether I it . . .'

Compulsory ellipses



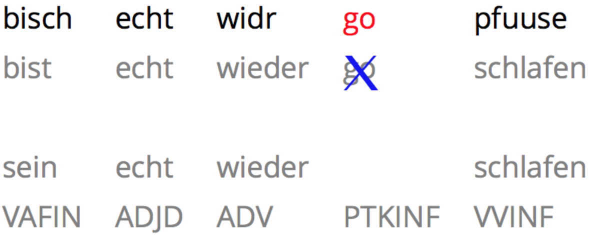
A	weli	adressa	häsch	mir	gschrieba	?
an	welche	Adresse	hast ∅	mir	geschrieben	?
an	welche	Adresse	haben	ich	schreiben	?
APPR	PIAT	NN	VAFIN	PPER	VVPP	$.

Figure: Standard German: *An welche Adresse hast Du mir geschrieben?*

‘To which one of my addresses did you write’

Compulsory particle

Does not exist in the Standard



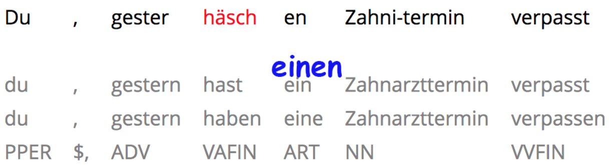
bisch	echt	widr	go	pfuuse
bist	echt	wieder	~~go~~	schlafen
sein	echt	wieder		schlafen
VAFIN	ADJD	ADV	PTKINF	VVINF

Figure: Particle *go*

‘Did you seriously go back to sleep?’

Case

No accusative in the NP in the Swiss German dialect



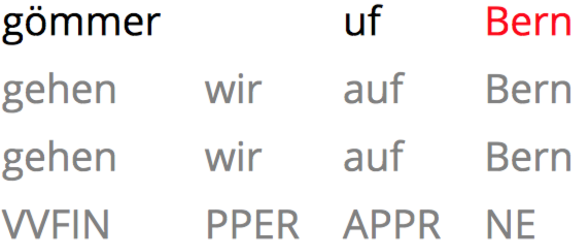
Du	,	gester	häsch	en	Zahni-termin	verpasst
				einen		
du	,	gestern	hast	ein	Zahnarzttermin	verpasst
du	,	gestern	haben	eine	Zahnarzttermin	verpassen
PPER	$,	ADV	VAFIN	ART	NN	VVFIN

Figure: Standard German: *. . . gestern hast [Du]* ***einen*** *Zahnarzttermin verpasst*

‘Hey, you skipped a dentist appointment yesterday’

Variation in prepositions

Auf in the dialect, *nach* in the Standard

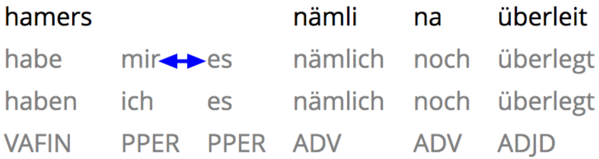


gömmer		uf	Bern
gehen	wir	auf	Bern
gehen	wir	auf	Bern
VVFIN	PPER	APPR	NE

Figure: Variation in the us of prepositions

‘let’s go to Bern’

Word order



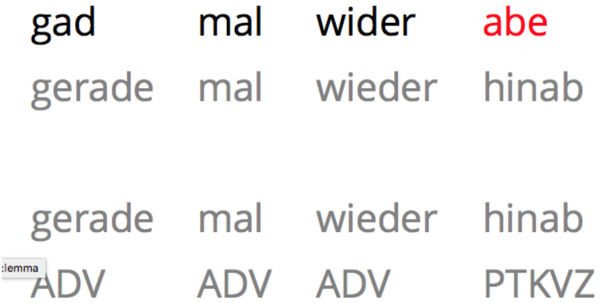
hamers			nämli	na	überleit
habe	mir ↔	es	nämlich	noch	überlegt
haben	ich	es	nämlich	noch	überlegt
VAFIN	PPER	PPER	ADV	ADV	ADJD

Figure: Standard German: *[ich] habe es mir nämlich noch überlegt*

‘I was actually thinking about that’

No adequate/distinct equivalent in the Standard

Example: *abe* ('downwards')



gad	mal	wider	abe
gerade	mal	wieder	hinab
gerade	mal	wieder	hinab
ADV	ADV	ADV	PTKVZ

Figure: Standard German equivalents: *hinunter, herunter, hinab, ø*

Necessity for normalization

Summary

- Unorthodox spelling (in all languages)
- Clitics
- Abbreviations, borrowings
- Dialect without spelling norms (German and Italian)
- Compulsory ellipses
- Compulsory particles
- Case
- Word order
- No adequate/distinct equivalents
- Five variants of Romansh

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

Main aims of normalization

- Research into the syntax (of the dialect)
- Research into lexical variation
- Prepare PoS

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

- Linguistic Approach
- Technical Approach

Conclusions

Outlook

Main aims of normalization

- Research into the syntax (of the dialect)
- Research into lexical variation
- Prepare PoS

Main aims of normalization

- Research into the syntax (of the dialect)
  - No change to word order
  - No compensation of ellipses
  - Leave required particles but standardize (*go, goge, ga, gage* –> *go*)
  - Do not adjust case
  - Separate clitics
  - No ‘replacement’ of prepositions
- Research into lexical variation
  - Use the Standard German variant wherever possible
  - Find a lemma that is similar in meaning and form where there is no equivalent, but be consistent
  - Mark abbreviations, emoticons and borrowings
  - Leave unrecognized elements as they are (e.g. *tkdn, iLSi*)
- Prepare PoS
  - Capitalize nouns (in German)
  - Expand abbreviations (e.g. *lg* –> *liebe Grüsse*)

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

- Linguistic Approach
- Technical Approach

Conclusions

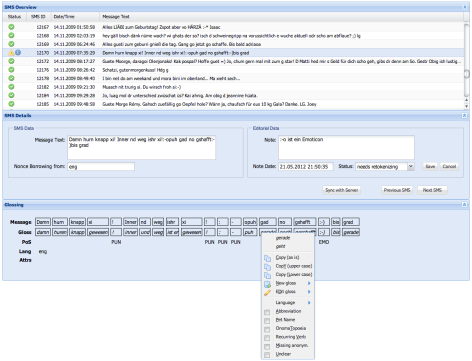
Outlook

Requirements

- Server based (–> co-operation)
- Common vocabulary for annotators
- Suggestions
- One-to-many and many-to-one
- Feedback (e.g. errors in tokenization)

The tool

SMS Glossing Tool



SMS Overview

Status	SMS ID	Date/Time	Message Text
✓	12167	14.11.2009 01:50:58	Alles LIÄBI zum Geburtstag! Zspot aber vo HÄRZÄ :-* Isaac
✓	12168	14.11.2009 02:03:19	hey gäll bisch dänk nüme wach? wi ghats der so? isch d schweinegripp na vorussichtlich e wuche aktuell odr scho am abflaue? ;) lg
✓	12169	14.11.2009 06:24:46	Alles gueti zum geburri gnieß die tag. Gang go jetzt go schaffe. Bis bald adriaoa
⚠ ⓘ	12170	14.11.2009 07:35:29	Damn hurn knapp xi! Inner nd weg ishr xi!:-opuh gad no gshafft:-)bis grad
✓	12172	14.11.2009 08:17:27	Guete Moorge, daragoi Olenjonaks! Kak pospal? Hoffe guet =) Jo, chum gern mal mit zum g star! D Matti hed mir s Geld für dich scho geh, gibs dr denn am So. Gestr Obig ish lustig...
✓	12176	14.11.2009 08:26:42	Schatzi, gutenmorgenkuss! Hdg g
✓	12178	14.11.2009 08:49:40	I bin net do am weekand und mora bini im oberland... Ma xieht sech...
✓	12182	14.11.2009 09:21:30	Muasch nit trurig si. Du wirsch froh si:-)
✓	12184	14.11.2009 09:29:28	Jo, luag mol dr unterschied zwüschat üs? Kai ahnig. Am obig d jeannine hüata.
✓	12185	14.11.2009 09:48:58	Guete Morge Rémy. Gahsch zuefällig go Oepfel hole? Wänn ja, chaufsch für eus 10 kg Gala? Danke. LG. Joey

SMS Details

SMS Data

Message Text: Damn hurn knapp xi! Inner nd weg ishr xi!:-opuh gad no gshafft:-)bis grad

Nonce Borrowing from: eng

Editorial Data

Note: :-o ist ein Emoticon

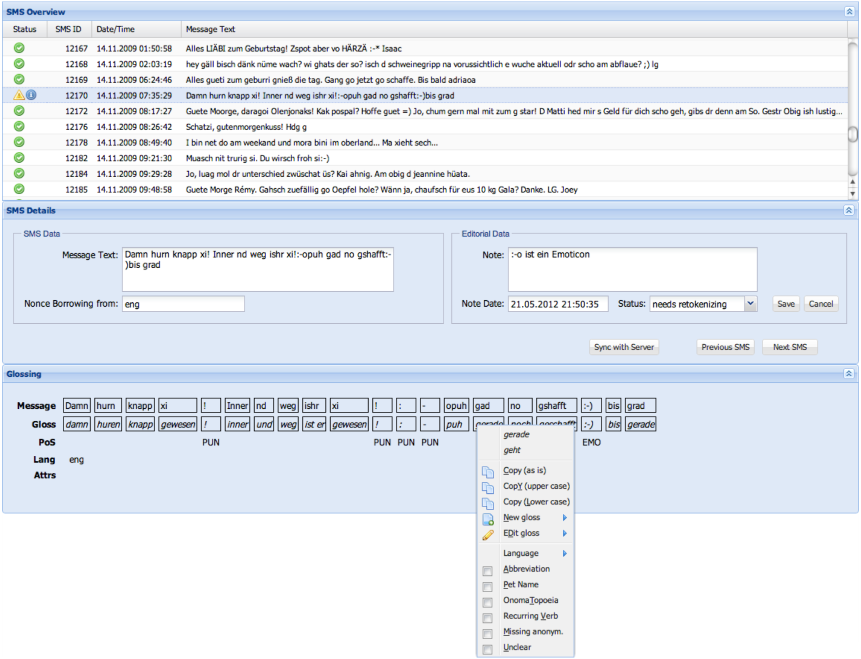
Note Date: 21.05.2012 21:50:35

Status: needs retokenizing

Glossing

Message	Damn	hurn	knapp	xi	!	Inner	nd	weg	ishr	xi	!	:	-	opuh	gad	no	gshafft	:-)	bis	grad
Gloss	*damn*	*huren*	*knapp*	*gewesen*	*!*	*inner*	*und*	*weg*	*ist er*	*gewesen*	*!*	*:*	*-*	*puh*	*gerade*	*noch*	*geschafft*	*:-)*	*bis*	*gerade*
PoS					PUN						PUN	PUN	PUN					EMO		
Lang	eng																			
Attrs																				

Figure: SMS Glossing Tool



SMS Overview

Status	SMS ID	Date/Time	Message Text
✓	12167	14.11.2009 01:50:58	Alles LIÄBI zum Geburtstag! Zspot aber vo HÄRZÄ :-* Isaac
✓	12168	14.11.2009 02:03:19	hey gäll bisch dänk nüme wach? wi ghats der so? isch d schweinegripp na vorussichtlich e wuche aktuell odr scho am abflaue? ;) lg
✓	12169	14.11.2009 06:24:46	Alles gueti zum geburri gnieß die tag. Gang go jetzt go schaffe. Bis bald adriaoa
⚠ ⓘ	12170	14.11.2009 07:35:29	Damn hurn knapp xi! Inner nd weg ishr xi!:-opuh gad no gshafft:-)bis grad
✓	12172	14.11.2009 08:17:27	Guete Moorge, daragoi Olenjonaks! Kak pospal? Hoffe guet =) Jo, chum gern mal mit zum g star! D Matti hed mir s Geld für dich scho geh, gibs dr denn am So. Gestr Obig ish lustig...
✓	12176	14.11.2009 08:26:42	Schatzi, gutenmorgenkuss! Hdg g
✓	12178	14.11.2009 08:49:40	I bin net do am weekand und mora bini im oberland... Ma xieht sech...
✓	12182	14.11.2009 09:21:30	Muasch nit trurig si. Du wirsch froh si:-)
✓	12184	14.11.2009 09:29:28	Jo, luag mol dr unterschied zwüschat üs? Kai ahnig. Am obig d jeannine hüata.
✓	12185	14.11.2009 09:48:58	Guete Morge Rémy. Gahsch zuefällig go Oepfel hole? Wänn ja, chaufsch für eus 10 kg Gala? Danke. LG. Joey

SMS Details

SMS Data

Message Text: Damn hurn knapp xi! Inner nd weg ishr xi!:-opuh gad no gshafft:-)bis grad

Nonce Borrowing from: eng

Editorial Data

Note: :-o ist ein Emoticon

Note Date: 21.05.2012 21:50:35

Status: needs retokenizing

Glossing

Message	Damn	hurn	knapp	xi	!	Inner	nd	weg	ishr	xi	!	:	-	opuh	gad	no	gshafft	:-)	bis	grad
Gloss	*damn*	*huren*	*knapp*	*gewesen*	*!*	*inner*	*und*	*weg*	*ist er*	*gewesen*	*!*	*:*	*-*	*puh*	*gerade*	*noch*	*geschafft*	*:-)*	*bis*	*gerade*
PoS					PUN						PUN	PUN	PUN					EMO		
Lang	eng																			
Attrs																				

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

Conclusions

- A small data set allows for manual treatment
- Linguistic rules: change as little as possible and be consistent when you have to change things
- Technical setup: Work with a self-growing dictionary that can be shared between annotators
- Resulting accuracy: ~95%

Outline

The corpus

Necessity of normalization

Aim of normalization

Approach

Conclusions

Outlook

Outlook

What's up, Switzerland

- Start: Jan 1st 2016
- 500 Mio tokens
- No manual processing possible
- SMS data will be used for automated annotation
- Accuracy: ???