What's up, Switzerland? Challenges of a Large, Multilingual CMC Corpus

CLARIN-PLUS workshop "Creation and Use of Social Media Resources". Kaunas

Simone Ueberwasser, Zurich Center for Linguistics

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing

The SMS corpus

- Data collection: 2009
- 27,000 SMS
- 0.5 Mio Tokens
- Multilingual (German, French, Italian, Romansh, others)
- Manual annotations (student assistants):
  - Languages
  - Anonymization
  - Normalization
- Automated PoS

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing

WhatsApp

- Data collection: 2014
- 617 Chats
- 763,650 messages with text
- 5,543,692 tokens
- Multilingual

Terminology

Figure 1: Chats and messages

The Projekt

- Swiss National Science Foundation (1.48 Mio €)
- Lead: Elisabeth Stark (French)
- Six (post-)docs
- Universities: Zurich, Bern, Neuchâtel, Leipzig
- 1/1/2016 – 31/12/2018

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing
- Ready
- Problems
- Appraches

Processing

- Anonymization: *Paul -> Peter* etc.
- Without consent: -> redactedQ12tokens12characters
- 😹 -> emojiQcatFaceWithTearsOfJoy

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing

- Ready
- Problems
- Appraches

Problems

- CMC data: non-standard spelling, emojis, emoticons, elipsis etc.
- Different languages
- Code-Switching
- Swiss German dialect
- Short text units (Average tokens per message: 6.8 without emojis)

Code-Switching

(from SMS corpus)

- *Olla fratello!!! Come stai? Wie geht's dir so? Immer noch so lange am arbeiten wie früher? Ich hab endlich mein eigenes Restaurant und mucho travajo...;-) aber macht mir extrem spass...;-) allora amore, buona giornata und luegsch uf di, gäll peace*
- Spanish, Italian, Standard German, Swiss German Dialect, English

Swiss German Dialect

(My own example)

(1)	**Dialect:**	**gaschs**			**abe**	**ga**	**sueche**	**bitte**
	Dialect:	**gasch**	**ø**	**s**	**abe**	**ga**	**sueche**	**bitte**
	Standard:	gehst	du	es	unten	ø	suchen	bitte
	English:	go	you	it	down$_{directional}$	go	look for	please

'Can you go and look for it downstairs please'

Outline

Looking back: The SMS corpus

WhatsApp

Corpus processing
- Ready
- Problems
- Appraches

Languages

- Look at first 500 messages per chat
- Define:
  - Language dominating more than 100 messages
  - Language present in less than 100 messages
  - Create sub-corpora per main language
- Within German sub-corpus: n-gram analysis to differentiate between dialect and standard per message.

Normalization/PoS

French sub-corpus

- MaltParser (Denis and Sagôt, 2009)
- Normalization precision: 92%
- PoS precision: 85%

Normalization/PoS

German/Italian/Romansh

- Manually normalize as many tokens as we can pay for
- Use the normalized data from the SMS corpus
- -> Character-level machine translation
- -> German: TreeTagger/RFTagger
- -> Italian/Romansh: ???

Availability

- SMS corpus: freely available for academic research (non-commercial, cc-by-nc)
- WhatsApp: until 2019: only on request.
- WhatsApp: after 2019: freely available for academic research (non-commercial, cc-by-nc)
- Either: no raw data