Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data

Anouck Braggaar and Rob van der Goot

University of Groningen, IT University of Copenhagen

Take-aways

- Code-switched, spontaneously spoken data with low-resource language Frisian.
- Instance selection from multiple treebanks using Latent Dirichlet Allocation (LDA).
- Best single source treebank (NL_ALPINO) resulted in a LAS of 54.7, our instance selection resulted in a LAS of 55.6 on test data.
- Additional experiments (removing diacritics, creating more similar training data and using XLM-R) did not lead to better performance.

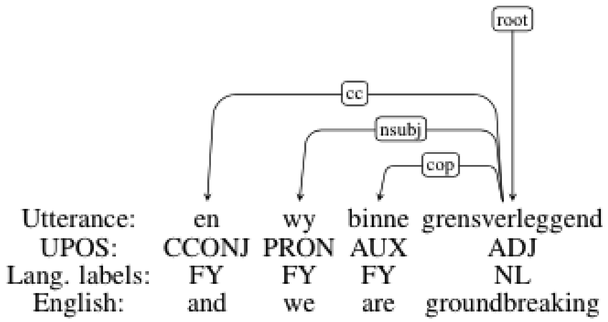
Research Question

- Can automatic data selection on the instance level outperform selection on the treebank level for training a parser for a new target domain/language?

Frisian

- Germanic language.
- Spoken in north-western part of the Netherlands.
- Approximately 612.000 speakers.

Data and Annotation



- Data of the FAME! corpus.
- Broadcasts of Omrop Fryslân, transcribed and annotated with code-switches by the creators of the corpus.
- Random selection of 400 utterances for annotation.
- Three batches of 50 utterances were annotated by both authors and disagreements were discussed and resolved.
- Next 250 utterances were annotated by one author and checked by the other.

	POS	UAS	LAS
Round 1	69.5	72.3	60.9
Round 2	87.1	76.1	64.6
Round 3	89.7	80.1	71.4

Experiments

- Selected 24 treebanks of languages close to Frisian, contained code-switching or consisted of transcribed speech.
- Single-treebank training: best eight scoring treebanks (based on LAS) were used in further experiments.
- For data selection: Latent Dirichlet Allocation (LDA) with character 1-5 grams and word 1-2 grams as features.
- Number of topics tested: 8, 16, 32, 64 and 128.
- 1000, 2000, and 4000 sentences for training.
- Training was done using MaChAmp 0.2.

Results

	NL_ALPINO		Eight		Best model	
	Dev	Test	Dev	Test	Dev	Test
POS	79.9	81.1	80.3	80.3	79.4	80.2
UAS	72.5	69.7	70.9	69.2	73.8	70.2
LAS	55.3	54.7	54.3	53.2	57.1	55.6

What did not work?

	Diacritics		Orphans		XLM-R	Best
POS	78.2		78.5		81.2	79.4
UAS	72.7		74.4		73.6	73.8
LAS	55.8		56.5		57.0	57.1