PoS Tagging, Lemmatization and Dependency Parsing of West Frisian

Wilbert Heeringa[1], Gosse Bouma[2], Martha Hofman[1], Jelle Brouwer[2], Eduard Drenth[1], Jan Wijffels[3], Hans Van de Velde[1,4]

[1]Fryske Akademy, [2]University of Groningen, [3]BNOSAC, [4]Utrecht University

Introduction

- West Frisian is a low-resource language spoken in the Dutch province of Fryslân having about 400,000 native speakers in 2018.
- We wish to lemmatize and PoS tag Frisian corpora in order to improve searchability.
- We also would like to study Frisian morphology and syntax using the corpora.
- For smaller languages like Frisian the possibilities for developing tools are limited, so we have to work efficiently.
- For a larger and related language like Dutch annotated corpora are already available (LassySmall, Alpino).
- Can we use these corpora in order to speed up the tagging of a Frisian corpus?

Aim

- To build a Frisian corpus
  - with lemmas, PoS-tags, morphological features, and syntactic dependencies
  - and using the widely used tagging and annotation standard of Universal Dependencies (version 2).
- To develop software that can be used to lemmatize/tag/annotate Frisian text (web app, web service).

Methods: PoS tagging via Dutch

1. Tag the Frisian text directly with a Dutch PoS tagger.
2. Tag via Dutch word-for-word translation and project the tags back onto the Frisian words. Translate with:
   - Google Translate,
   - Oersetter 1.0, (*Oersetter* is Frisian for 'translator')
   - Oersetter 2.0.
3. For a parallel corpus, align the Dutch sentences with respect to the Frisian sentences with `fast_align`. Tag the Dutch sentences and project the tags back onto the Frisian words.

Tested ...

- on a corpus: 20,926 tokens, 1,152 sentences, PoS tags were manually verified;
- using Dutch tagger trained on LassySmall 2.8 corpus.

Results

- Percentage of correct PoS tags per procedure:

	% correct
Tagging directly on Frisian	51.5%
Via Google Translate	76.0%
Via Oersetter 1	89.8%
Via Oersetter 2	87.1%
Via alignment	74.2%

- In all cases correction afterwards proved necessary.

Morphological and syntactic features

- Annotation also via Dutch.
- Initially via Dutch sentences that were aligned with the Frisian sentences. Manual correction where errors were in the alignment.
- Later via word-for-word translation with the Oersetter 1.0. The translation was manually corrected before the annotator was applied.

Composition of corpus

source	#tokens	%tokens	#words	%words	#sentences	%sentences
news	8,737	17	7,998	17	582	19
science	2,293	4	2,069	5	107	3
Wikipedia	13,780	27	12,040	27	505	16
museum	9,275	18	8,335	19	486	16
novels	17,176	34	14,272	32	1,446	46
total	51,261		44,714		3,126	

Training the lemmatizer/tagger/annotator

- We used UDPipe via the R package `udpipe` developed by Jan Wijffels.
- Using the newly trained Frisian PoS tagger 98.4% of the PoS tags in the test corpus (see above) are predicted correctly.

k-fold cross-validation

- k=10
- Training on Frisian with hyperparameter settings used for training based on LassySmall 2.5.
- For comparison, we also trained a model based on the Dutch LassySmall 2.8 corpus with the same hyperparameter settings.

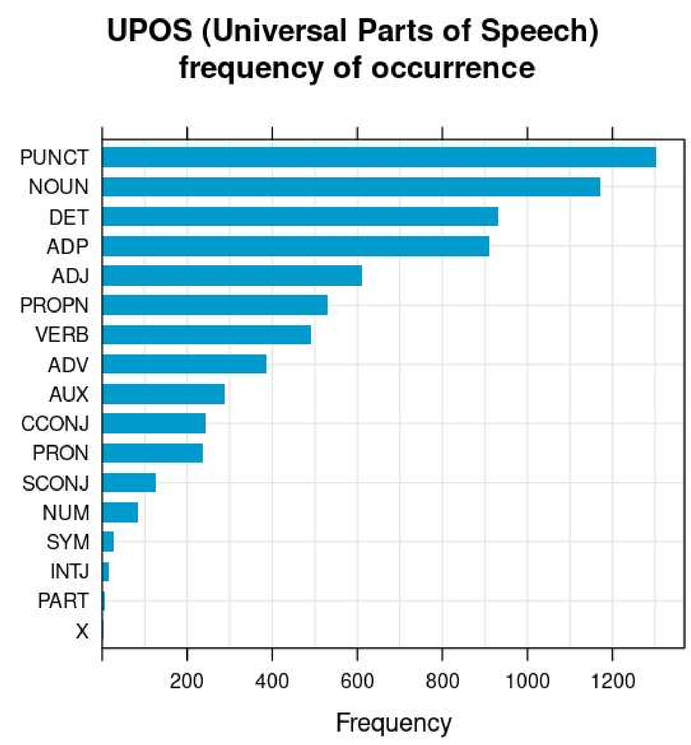
Results

	Raw text		Gold tok		Gold tok + mor	
	Frisian	Dutch	Frisian	Dutch	Frisian	Dutch
f1 words	100.0	99.1				
f1 sents	89.7	81.3				
UPOS	94.6	95.6	94.6	95.9		
XPOS	88.1	93.7	88.1	94.1		
UFeats	89.8	95.1	89.8	95.6		
Lemma	96.0	94.2	96.0	94.4		
UAS	72.5	81.3	73.1	83.4	78.5	87.1
LAS	66.4	77.5	67.0	79.4	73.9	84.0

The results for 'Dutch' (LassySmall) are usually significantly better, but for 'f1 words', 'f1 sents' and 'Lemma' they are significantly worse.

Web app and web service

- Web app: `frisian.eu/udpipeapp`
- Web service: `frisian.eu/udpipeservice`



Frequencies of the PoS tags in the Frisian Wikipedia text 'Ingelsk'

Future work

- Deposit current corpus in UD repository (UFAL).
- Extend corpus to 100,000 words, make corpus more balanced.
- Newly added text is now best initially lemmatized and annotated with UDPipe Frysk (so no longer via Dutch) followed by manual correction.
- Perhaps that the morphological and syntactic features can still be generated best via a Dutch word-for-word translation.

Contact

E-mail: `WHeeringa@fryske-akademy.nl`, `G.Bouma@rug.nl`, `HVandeVelde@fryske-akademy.nl`.