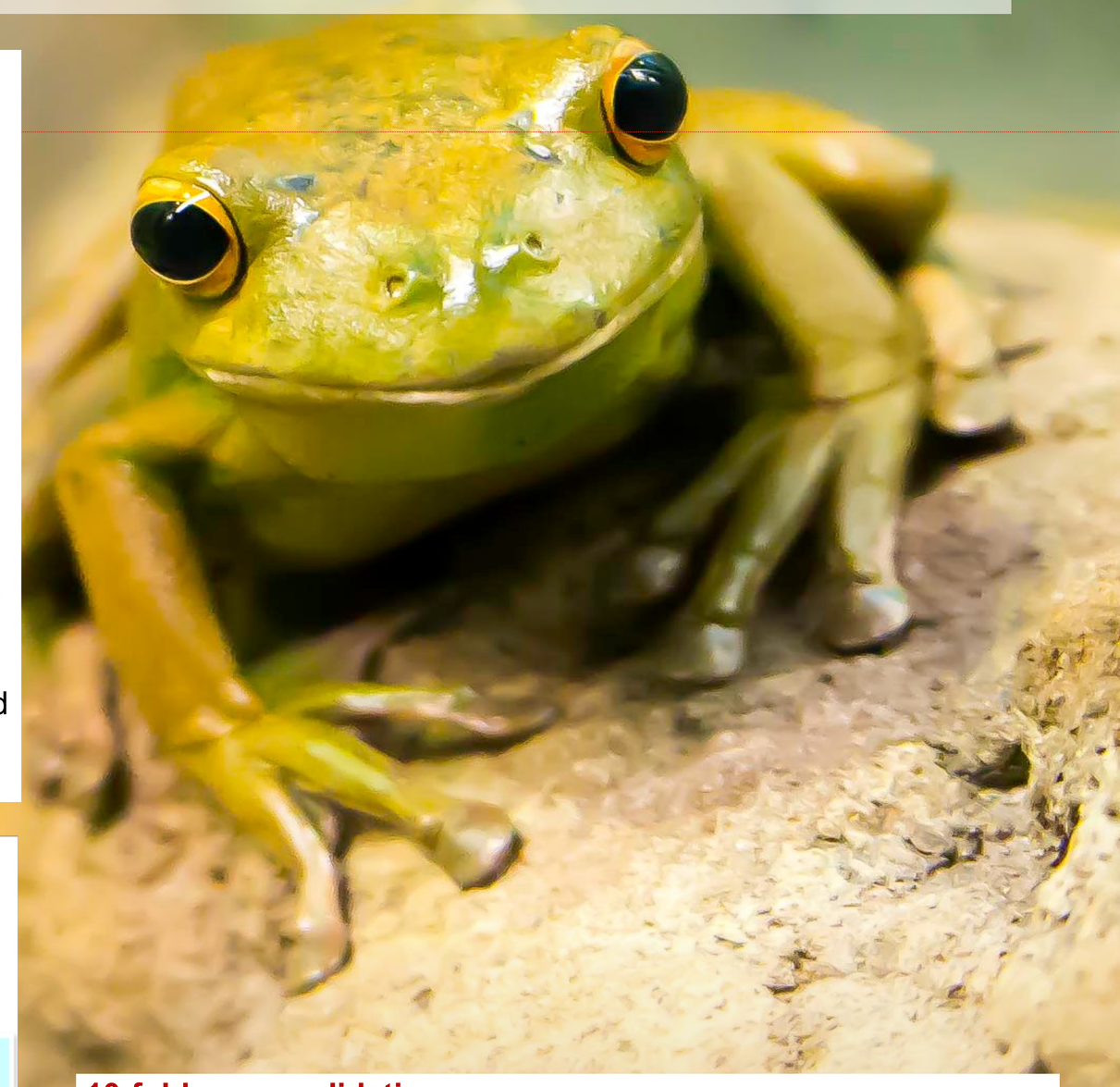
How high can Frog leap?

Iris Hendrickx, Ko van der Sloot, Maarten van Gompel and Antal van Den Bosch

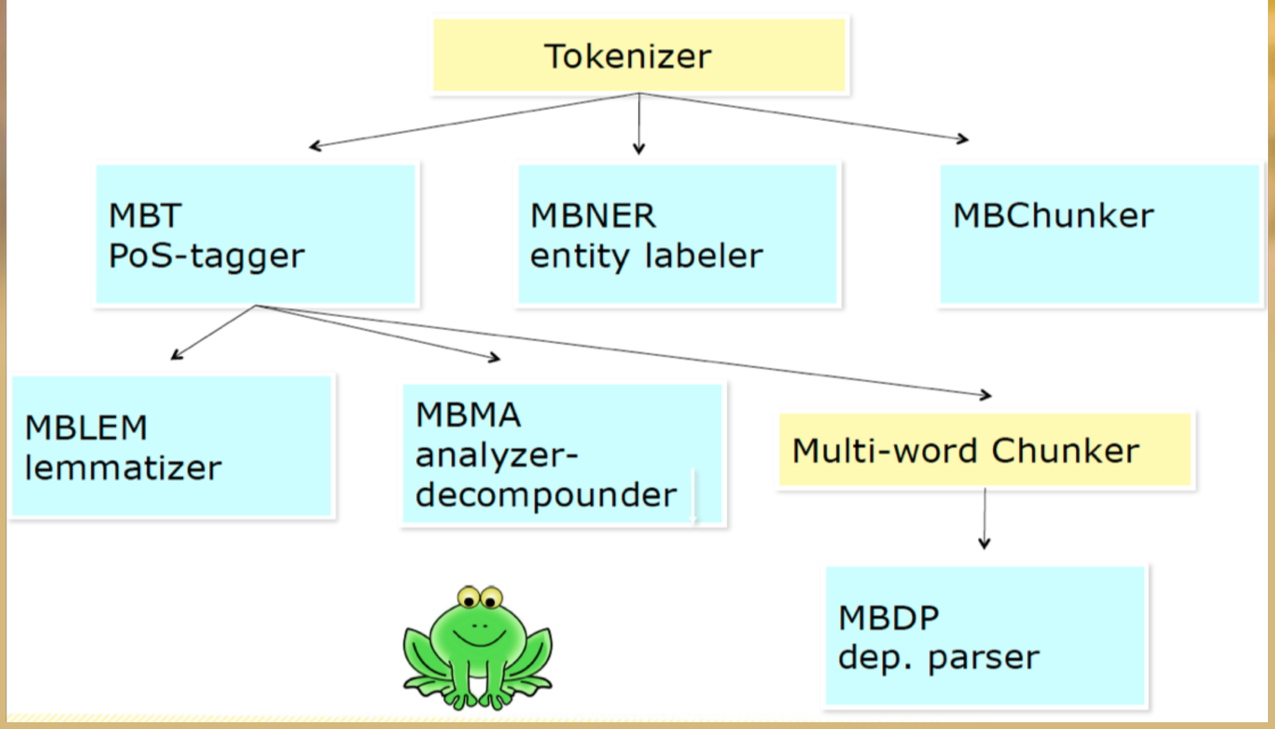
Frog is a natural language processing pipeline for Dutch that enriches a text with information on word and sentence boundaries, part-of-speech-tags, lemmas, morphological analysis, syntactic information and named entities.

Most of the NLP modules in Frog use a k-nearest neighbour approach and are trained using Timbl, the Tilburg memory-based learning software package. Many modules were created already in the 1990s by ILK Research Group (Tilburg University, the Netherlands) and the CLiPS Research Centre (University of Antwerp, Belgium) but have been updated and retrained over the years. Frog is thus the result of many years of work and still has an active support and continues to be improved.

Here we present an evlaution of these modules.The optimal evaluation of the modules is performed by using a new and unseen test set made. The developers of frog are not suited to do this annotation as theu might be biased by the implementation decisions in frog and will be inclined to follow those. However, finding independent test sets was not always possible, and there are used 10-fold cross validation on training material to get a performance indication.



FROG architecture



Tokenization

Evaluation on manually verified samples from DCOI and Meantime corpus.

Observations:

- Results are very high scores,
- remaining errors concern splits on dashes and uncommon names with unexpected characters in the strings

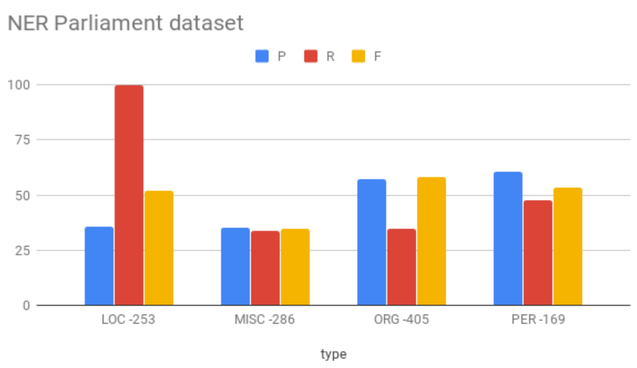
	data	sent	P	R	F
NL	Dcoi	2897	99.9	99.9	99.9
NL	meantime	61	99.1	99.5	99.3
EN	meantime	59	98.9	99.4	99.2
IT	meantine	59	99.9	99.0	99.5

NER

We evaluated on the NERC CLIN26 shared task data where FROG obtained F-scores between 49-57%, very similar to the tests on parliament data.

Observations:

- NER scores are rather low
- NER is domain dependent and FROG is trained on (old) SoNAR-1 data



10-fold cross validation

We evaluated modules on training material. POS-tagger and chunker are trained on 11.133K tokens, 998.796 sentences, Lemmatizer is trained on CELEX.

Observations: biased over-optimistic results as part of data was automatically labeled

Module	Tags	Accuracy
POS	12	97.97
POS	All	95.66
LEM	All	
CHUNK	22	92.50

Speed test

Evaluation on 2788 plain text files with 183 words on average

Observations:

- XML FoLiA input instead of plain text -> hardly influences speed
- Slower speed when all modules are used -> due to NER
- parsing is also a slow step but is executed in parallel to NER

Speed test	files per/second
all modules - textual output	2,30
all modules - FoLiA XML output	2.25
pos-lemmas only - textual output	12

References

Jonkers, R. Named Entity Recognition on Dutch Parliamentary Documents using Frog, BA thesis, University of Amsterdam, 2016

Minard, A. L., Speranza, M., Urizar, R., Altuna, B., van Erp, M. G. J., Schoen, A. M., & van Son, C. M. (2016). MEANTIME, the NewsReader multilingual event and time corpus. In Proceedings of LREC 2016

Oostdijk, N., Reynaert, M., Monachesi, P., van Noord, G. J. M., Ordelman, R., Schuurman, I., & Vandeghinste, . (2008). From D-Coi to SoNaR: A reference corpus for Dutch. In proceedings of LREC 2008

LAnguage MAchines

Frog documentation: https://frognlp.readthedocs.io/

Radboud University