Creating a Universal Dependencies Treebank of Spoken Frisian-Dutch Code-switched Data

Anouck Braggaar & Rob van der Goot

UD for spoken code-switched data

- Previous work:
  - Annotating code-switches (Çetinoglu and Çöltekin (2019), Seddah et al. (2020), Partanen et al. (2018))
  - Annotating spoken data (Davidson et al. (2019), Dobrovljc and Martinc (2018), Partanen et al. (2018), Çetinoglu and Çöltekin (2019))

UD for spoken code-switched data

- Common problems: disfluencies and sentence segmentation
- Solutions: adapting guidelines versus extending guidelines

- Our data:
  - FAME!-project dataset (broadcasts from Omrop Fryslân) by Yilmaz et al. (2016) (around 18 hours)
  - Spontaneous speech/low resource/Frisian-Dutch code switching

Annotations

- Universal Dependency guidelines
- 250 sentences for test and 150 for development
- 2 annotators
- 150 sentences for inter-annotator agreement: batches of 50

Fierljeppen

	POS	UAS	LAS
Round 1	69.5	72.3	60.9

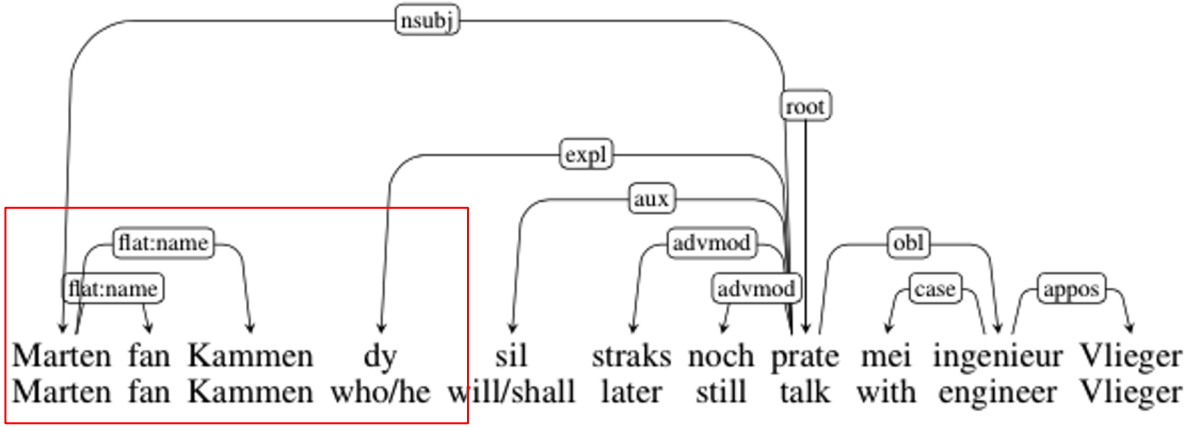
Inter-annotator agreement

- 150 sentences for inter-annotator agreement: batches of 50
- 2 annotators
- Disagreements due to:
  - difficulties with non-standard constructions
  - sentence segmentation
  - interpretation of utterances (ambiguity)
  - annotators had to learn the guidelines

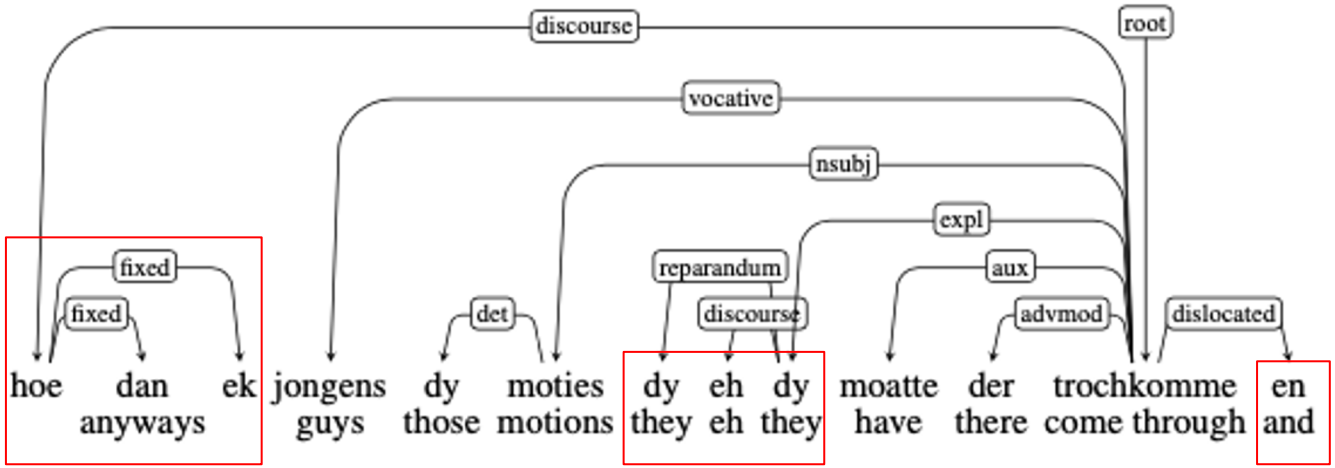
	POS	UAS	LAS
Round 1	69.5	72.3	60.9
Round 2	87.1	76.1	64.6
Round 3	89.7	80.1	71.4

Table 1: POS, UAS and LAS scores between the two annotators.

Annotation issues:



Annotation issues:



	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

Table 1: POS, UAS and LAS scores between the two annotators.

First experiments

- MaChAmp
- Single treebank training (selection of 25 treebanks)
  - Dutch Alpino: 72.10 UAS and 55.28 LAS
  - Dutch LassySmall: 71.01 UAS and 54.48 LAS
- Data selection
  - LDA/GMM
  - Number of treebanks
  - Number of clusters
  - Features
  - Can we exploit small Frisian Universal Dependency treebanks?

Leonoor Van der Beek, Gosse Bouma, Rob Malouf, and Gertjan Van Noord. 2002. The Alpino dependency treebank. In *Computational linguistics in the netherlands 2001*, pages 8–22. Brill Rodopi.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2017. Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 324–330, Valencia, Spain. Association for Computational Linguistics.

Özlem Çetinoglu and Çagri Çöltekin. 2019. Challenges of annotating a code-switching treebank. In: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019),* pages 82-90, Paris, France. Association for Computational Linguistics.

Sam Davidson, Dian Yu, and Zhou Yu. 2019. Dependency parsing for spoken dialog systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1513– 1519, Hong Kong, China. Association for Computational Linguistics.

Kaja Dobrovoljc and Matej Martinc. 2018. Er... well, it matters, right? On the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 37–46.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajicc, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132, Brussels, Belgium. Association for Computational Linguistics.

Djameé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content north-african arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150.

Gertjan Van Noord, Gosse Bouma, Frank Van Eynde, Daniel De Kok, Jelmer Van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In *Essential speech and language technology for Dutch*, pages 147–164. Springer, Berlin, Heidelberg.

Henk Wolf. 1996. Structural neutrality in Frisian- Dutch interaction. *Us Wurk*, 45(3-4):125–138

Emre Yilmaz, Maaike Andringa, Sigrid Kingma, Jelske Dijkstra, F Kuip, H Velde, Frederik Kampstra, Jouke Algra, H Heuvel, and David A van Leeuwen. 2016. A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research.

Daniel Zeman, Jan Hajicˇ, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

MaChAmp:
van der Goot, R., Üstün, A., Ramponi, A., & Plank, B. (2020). Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP. *arXiv preprint arXiv:2005.14672*.