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

Anouck Braggaar
University of Groningen
a.r.y.braggaar@student.rug.nl

Rob van der Goot
IT University of Copenhagen
robv@itu.dk

Abstract

While high performance has been obtained for dependency parsing of high-resource languages, performance for low-resource languages lags behind. In this paper we focus on the parsing of the low-resource language Frisian. We use a sample of code-switched, spontaneously spoken data, which proves to be a challenging setup. We propose to train a parser specifically tailored towards the target domain, by selecting instances from multiple treebanks. Specifically, we use Latent Dirichlet Allocation (LDA), with word and character N-gram features. The best single source treebank (NL_ALPINO) resulted in an LAS of 54.7 whereas our data selection outperformed the single best transfer treebank and led to 55.6 LAS on the test data. Additional experiments consisted of removing diacritics from our Frisian data, creating more similar training data by cropping sentences and running our best model using XLM-R. These experiments did not lead to a better performance.

1 Introduction

As parsers are improving (with currently scores higher than 96 (Mrini et al., 2020)), parsing scores for low-resource languages lag behind. In recent years there has been an increase in interest in ways to parse and annotate these languages. This paper will focus on parsing the low-resource language Frisian (Germanic language spoken in the north-western part of the Netherlands, approximately 612.000 speakers[1]). We will use spontaneous speech data containing code-switches from the FAME! corpus created by Yilmaz et al. (2016). We annotate a small portion of this data with UPOS tags and Universal Dependencies (Zeman et al., 2020) for evaluation purposes. An example of such an utterance can be seen in Figure 1.

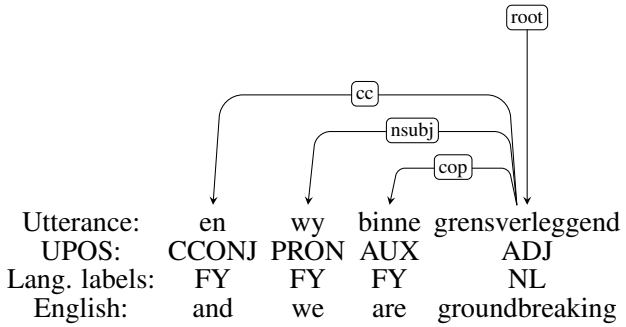

Figure 1: Example of annotated utterance.

Because it is unclear which (parts of) datasets are likely to be good candidates for training a parser for spoken Frisian, we propose to use instance-based selection from a variety of sources. This leads to our main 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? After finding our best model, we will try to improve our results by removing diacritics in our Frisian data and by making our train data more similar to spoken Frisian. All data and code are available on GitHub[2].

2 Related Work

Previous work on dependency parsing for spoken data created English annotations for conversational agents (Davidson et al., 2019) and Slovenian data (Dobrovoljc and Martinc, 2018). Partanen et al. (2018) create a treebank for spoken Komi-Zyrian with code-switching to Russian. Contemporary to our treebank, Çetinoğlu and Çöltekin (2019) created a treebank for spoken code-switched Turkish-German. They adapt the guidelines to deal with the spoken and code-switch nature of the data. Seddah et al. (2020) create a treebank for an Arabic dialect that contains a high amount of code-switching.

[1] https://fy.wikipedia.org/wiki/Fryske_talen

[2] https://github.com/Anouck96/ParsingFrisian

Meechan-Maddon and Nivre (2019) and Blodgett et al. (2018) show that annotating a small amount of target language and target domain data outperforms a cross-lingual setup, but we focus on a pure zero-shot scenario instead. Previous work on zero-shot parsing has mainly focused on annotation projection using parallel data (Barry et al., 2019) and selecting transfer treebanks (Meechan-Maddon and Nivre, 2019). Previous work exploring more fine-grained selection methods are either focusing only on 3 domains (and multiple topics within the news) and 3 corpora (Plank and Van Noord, 2011), or focus on parser selection during test time (Litschko et al., 2020).

3 Data & Annotations

3.1 Data

The creators of the FAME! corpus (Yilmaz et al., 2016) had already transcribed and segmented the data and annotated the code-switches. We copied their word-level language labels to our miscellaneous column. The data consists of broadcasts from Omrop Fryslân (Frisian radio broadcaster) and mainly contains spontaneous interviews. The manually annotated radio broadcasts contain approximately 18.5 hours of speech and contains 3837 word- and sentence-level code-switches (Yilmaz et al., 2016). The majority of these switches are from Frisian speakers who switch to Dutch, because as Yilmaz et al. (2016) also mention it is not common for Dutch speakers to switch to Frisian. 67.8% of the words are Frisian and 26.1% are Dutch. The remainder of the words are annotated as being Frisian-Dutch, are hesitations (like "eh") or are of a different language. From this corpus we randomly selected and annotated 400 utterances. Each utterance contains at least one switch from Frisian to Dutch or the other way around. In our selection there are 144 different speakers, 135 of them Frisian speakers (the other 9 Dutch). Most of the speakers only occur one, two or three times. In the corpus 3,067 tokens are Frisian, 625 Dutch and 37 are other languages or annotated as Frisian-Dutch. The distributions over the languages for POS tags can be found in Appendix A.

3.2 Annotations

As a starting point, 150 utterances where annotated by two annotators (the authors of the paper, who have backgrounds in NLP, parsing and linguistics), in three batches of 50 utterances. After each batch, disagreements where discussed and resolved. Afterwards, 250 more utterances where annotated by one of the annotators and checked by the other. Table 1 shows the inter-annotator agreement between the three initial batches of annotation. We report accuracy over Universal Part-of-speech tags, Unlabelled Attachment Score (UAS) and Labelled Attachment Score (LAS) (Zeman et al., 2018). Most of our disagreements were due to difficulties with non-standard constructions, sentence segmentation, ambiguities of utterances and the fact that the guidelines had to be applied consistently, and in some cases had to be adapted. Adaptations to the guidelines can be found in Appendix B. For a more detailed discussion regarding the annotations we refer to Braggaar and van der Goot (2021).

	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.

3.3 Analysis of disagreements

To find trends in the disagreements, we calculated confusions over all labels (POS tags and Universal Dependency relations). The most occurring confusion for the POS tags that is solved in the third round is between ADV-AUX. This is due to shortcuts in the annotation tool ConlluEditor (Heinecke, 2019). Another confusion that is mostly solved in the third round is the confusion between AUX and VERB. For the Universal Dependency relations, the most common disagreements included amod-advmod confusions and differences in selecting the root.

4 Experiments

4.1 Data selection

First, we selected 24 treebanks of languages close to Frisian, which contained code-switching or consisted of transcribed speech. The languages included are Dutch, English, German, French, Naija, Afrikaans, Danish and Hindi-English[3]. As a baseline, 24 parsers were trained, each on a single treebank. The eight best scoring treebanks (based on LAS score) were used for further experiments.

[3]The names of all treebanks can be found in Appendix C

The eight best scoring treebanks were: Dutch Alpino (Van der Beek et al., 2002), Dutch LassySmall (Van Noord et al., 2013), German GSD, German PUD, English GUM (Zeldes, 2017), English EWT(Silveira et al., 2014), English ParTUT (Sanguinetti and Bosco, 2015) and the Tweebank by Liu et al. (2018). We chose to use these eight because we observed a drop in performance beyond the lowest scoring treebank in this selection.

4.2 Latent Dirichlet Allocation

For the data selection, we use Latent Dirichlet Allocation (LDA) with character 1-5 grams and word 1-2 grams as features. We choose LDA, because it is unsupervised, efficient and returns the probability of instances belonging to each clusters. Traditionally LDA is commonly used for topic classification, but when given data from multiple languages, we hypothesize that it will cluster based on language similarity instead. In LDA, one has to define the number of classes; we started with eight topics (because we used eight treebanks) and multiplied the number by two for the successive runs (8, 16, 32, 64 and 128).

4.3 Experimental setup

We use the full training data from all eight treebanks. For the treebanks without train, we use the test data, as we do not perform any experiments on the source data. For the LDA target data, we sampled sentences from the train split of the FAME! (Yilmaz et al., 2016) corpus. Note that we did not use the samples that are in our annotated data set. After the clustering, we select similar sentences based on the euclidean distance to the mean of the Frisian input. We ran experiments with 8, 16, 32, 64 and 128 LDA components. From the results we selected 1000, 2000 and 4000 sentences for training to see if the number of sentences in training was also influential. We did a total of 75 runs (8, 16, 32, 64, 128 components with each 1000, 2000 and 4000 sentences, 5 seeds). Training was done using a deep biaffine parser as implemented by MaChAmp 0.2 (van der Goot et al., 2020) with default parameters and mBERT embeddings (Devlin et al., 2019). Five random seeds were used for training.

5 Results

Overall, our best model consisted of 128 components and used 2000 sentences. Table 2 shows the effect of changing the number of components when using 2000 sentences for training. While 128 components is the best for LAS and UAS, 8 components gave the best result for POS. When using the best performing number of components (128) and varying the amount of training data (Table 3), we can also see that the best model only needs 2000 sentences while using more sentences seemed to improve only the POS score.

Components	8	16	32	64	128
POS	**81.2**	80.4	81.0	79.4	79.4
UAS	72.7	71.4	70.8	72.6	**73.8**
LAS	55.2	55.1	54.8	55.5	**57.1**

Table 2: POS, UAS and LAS scores on dev data for 2000 sentences, mean over 5 random seeds. Highest scores are marked in bold.

Sentences	1000	2000	4000
POS	78.5	79.4	**79.8**
UAS	71.5	**73.8**	73.3
LAS	54.6	**57.1**	56.0

Table 3: POS, UAS and LAS scores on dev data for 128 components, mean over 5 random seeds. Highest scores are marked in bold.

Table 4 shows the results of our best model (128 components, 2000 sentences) compared to the best single treebank parser (NL_ALPINO) and training on the eight treebanks simultaneously. The scores for test are for LAS and UAS slightly lower than in the development phase. Surprisingly the POS scores are a bit higher. Our model outperforms the baselines on the test scores for LAS and UAS (but not for POS). We tested significance with random Bootstrapping as done by Udapi (Popel et al., 2017) compared to both baselines. Unfortunately, none of the results have proven to be significant at an alpha of 0.05.

	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

Table 4: POS, UAS and LAS scores baselines versus best model (128 components/2000 sentences). Dev over five random seeds, test over the best random seed of dev.

NL_ALPINO	1514	NL_LASSYSMALL	447
DE_GSD	14	EN_EWT	14
EN_GUM	5	EN_TWEEBANK	4
EN_PUD	2		

Table 5: Sources of training data instances.

NL_ALPINO		Best Model	
INTJ-ADV	45	INTJ-ADV	48
ADV-ADJ	20	INTJ-PRON	20
DET-ADP	19	ADV-ADJ	17
AUX-VERB	14	DET-ADP	16
PRON-ADV	12	AUX-VERB	14

Table 6: Top five confusions on POS. Actual-predicted, number of confusions.

A closer look at the data selection of our model (Table 5) shows that the training set consists of mainly Dutch data. Only few sentences are selected that are not in the Dutch treebanks. This comes as no surprise, as we have seen that in the total corpus 26.1% of the words are Dutch, and the Frisian language is related to Dutch. This also agrees with the fact that the Dutch treebanks perform best in single treebank training.

6 Error Analysis

We perform an error analysis of our best model and the baseline (NL_ALPINO) on the development data. Table 6 shows the top five confusions on POS tags. The baseline is often unable to correctly predict interjections and our model does not seem to solve this confusion. Some of the difficulties that we have come across during annotation also arise here (e.g. ADV-ADJ and AUX-VERB).

The confusions on dependency labels are similar for both models (Table 7). Our model is slightly better on discourse-parataxis, but other than that it shows no big improvements. This is not surprising as our data consists mainly out of Alpino sentences.

NL_ALPINO		Best Model	
discourse-parataxis	39	discourse-parataxis	31
cc-mark	20	cc-mark	20
det-case	16	discourse-advmod	16
discourse-advmod	13	det-case	13
root-parataxis	12	advmod-amod	10

Table 7: Top five confusions on dependency labels. Actual-predicted, number of confusions.

The most occurring confusion (discourse-parataxis) was also a cause of disagreement in the annotation process.

7 What did not work?

In an attempt to obtain better results we have tried three different approaches to improve our best model. Muller et al. (2020) have shown that transliteration to a script of a related high source language on which the language model is trained leads to better results. Even though Frisian is not in a different script as the training data, it does contain a high amount of diacritics, which results in both many unseen wordpieces during testing, as well as differences in tokenization[4]. We attempt to overcome this mismatch by removing the diacritics from the characters in our Frisian data. In our development set, which consists of 150 utterances, there are 44 utterances with at least one diacritic and there are a total of 53 tokens with diacritics.

The second approach tries to make our training data more similar to our Frisian spoken data. A similar approach is taken by Blodgett et al. (2018), they create synthetic data following Internet-specific conventions and syntactic features of African-American English. This synthetic data proved to be helpful for performance.Vania et al. (2019) also try similar methods of data augmentation and found that when no source treebank is available, data augmentation can be very helpful. In our case, the utterances are not always "full sentences". An example from our development set is *"dêr kinnen we in hele hoop oer sizze mar"* ("we can say a lot about that but"). Normally the sentence would continue after "mar" (but) and "mar" would connect to this next part, but in this case the utterance stops at this point. We chose to connect this "mar" with an orphan relation to the root. These kind of "non-standard" constructions occur relatively often in our target data compared to the non-spoken treebanks in our training data. We tried to make our training data more similar by cropping sentences and adding an orphan relation. We did this for approximately the same percentage of utterances that have this construction in our development data (approximately 9 %).

As a final experiment we replaced mBERT with XLM-R. The results of all three experiments can

[4]We saw empirically that the mBERT tokenizer was splitting words containing diacritics more, presumably because diacritics were underrepresented during training of the tokenizer.

	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

Table 8: POS, UAS and LAS scores for the additional experiments on dev set (128 components, 2000 sentences).

be found in Table 8. As can be seen, our diacritics and orphan experiments do not increase our scores. Only the UAS increases slightly when cropping some sentences, but unfortunately the LAS drops in both cases. The results with XLM-R are highly similar to the mBERT scores. Only the POS scores increase slightly but the difference in scores for UAS and LAS is very small.

8 Conclusion

In this paper we explored parsing and annotating the low-resource language Frisian and we have shown that selecting more similar training data (by using LDA) can lead to improvements in scores. These improvements were not significantly better compared to our baselines, but it did show that selecting fewer instances can outperform single treebank training. Besides slightly higher scores, using fewer instances sped up the training process. Additional experiments of removing diacritics and adjusting the training data to our development set did not show improvements. Future research can focus on improving the data selection method (using a different selection approach will probably result in different outcomes), weighing the importance of the selected instances and combining data selection with orthogonal approaches.

Acknowledgements

We would like to thank Gosse Bouma, Joakim Nivre, Lars Borin, Sif Dam Sonniks and Dick van der Goot for their help with the annotations. We thank Henk van den Heuvel and his collegues for allowing us to use and share a part of the FAME! dataset. We also thank members of NLP-North, Özlem Çetinoğlu and the anonymous reviewers for feedback on earlier versions of this paper. This research was supported by an Amazon Research Award.

References

James Barry, Joachim Wagner, and Jennifer Foster. 2019. Cross-lingual parsing with polyglot training and multi-treebank learning: A Faroese case study. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 163–174, Hong Kong, China. Association for Computational Linguistics.

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.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter universal dependency parsing for african-american and mainstream american english. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Anouck Braggaar and Rob van der Goot. 2021. Creating a universal dependencies treebank of spoken frisian-dutch code-switched data.

Özlem Çetinoğlu and Çağrı 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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, Brussels, Belgium. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, and Barbara Plank. 2020. Massive choice, ample tasks (machamp):a toolkit for multi-task learning in nlp.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris.

Robert Litschko, Ivan Vulić, Željko Agić, and Goran Glavaš. 2020. Towards instance-level parser selection for cross-lingual transfer of dependency parsers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3886–3898, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Ailsa Meechan-Maddon and Joakim Nivre. 2019. How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. *arXiv preprint arXiv:2010.12858*.

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.

Barbara Plank and Gertjan Van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Manuela Sanguinetti and Cristina Bosco. 2015. Parttut: The turin university parallel treebank. In *Harmonization and development of resources and tools for italian natural language processing within the parli project*, pages 51–69. Springer.

Djamé 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, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

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.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. 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 1105–1116.

Emre Yilmaz, Maaike Andringa, Sigrid Kingma, Jelske Dijkstra, Frits van der Kuip, Hans Van de Velde, Frederik Kampstra, Jouke Algra, Henk van den Heuvel, and David van Leeuwen. 2016. A longitudinal bilingual Frisian-Dutch radio broadcast database designed for code-switching research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4666–4669, Portorož, Slovenia. European Language Resources Association (ELRA).

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Daniel Zeman, Jan Hajič, 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.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara,

Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurosson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna

Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Distribution languages over POS tags

	Frisian	Dutch	Other
ADV	365	89	0
PRON	388	53	1
NUM	58	2	0
NOUN	314	158	11
ADP	396	44	1
DET	337	21	0
VERB	343	50	6
CCONJ	152	7	0
INTJ	181	8	0
ADJ	158	54	3
PROPN	158	133	14
AUX	158	4	1
SCONJ	59	2	0

Table 9: Number of tokens for the combination language and POS tag.

B Annotation guidelines

We tried to follow the universal dependency guidelines as much as possible. In cases we could not follow them precisely we discussed and made our own guidelines:

- To determine if something is an ADJ or ADV we use a dictionary. If the word is not in the dictionary we use the frequencies of the Alpino treebank. For the decision between advmod and amod, we used amod only for nominals.
- If “toen/want/tot/dan” are at the beginning of utterances we tag them as mark and not as conj. “en” is cconj or cc.
- Orphan always attaches to the root. Orphan is used in cases where utterances seem to have a “non-standard” ending. Normally the word would be attached to the right part of the utterance but this part is missing. Both Figures 2 and 3 show utterances with orphans.

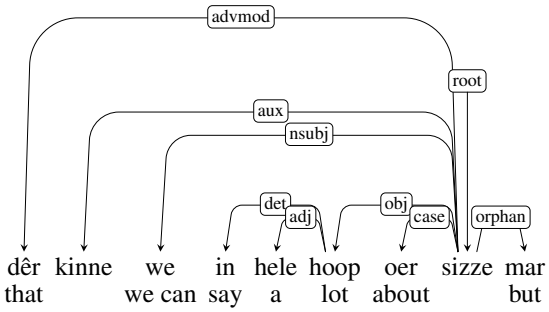

Figure 2: Utterance with orphan relation.

- Discourse is always attached to the highest node without any projectivity. Figure 3 shows an utterance with a discourse relation.

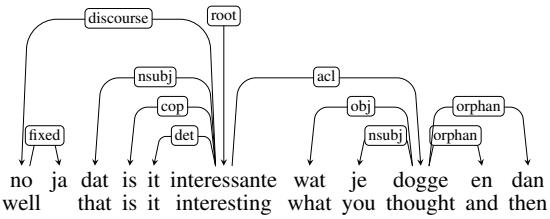

Figure 3: Utterance showing both orphan and discourse relations.

C List of treebanks

1. Treebanks in UD 2.7:
    - English ESL
    - Hindi English HIENCS
    - English GUM Reddit
    - English EWT
    - English GUM
    - English LinES
    - English ParTUT
    - English Pronouns
    - English PUD
    - French FQB
    - French ParTUT
    - French PUD
    - French Sequoia
    - French Spoken
    - German GSD
    - German LIT
    - German PUD
    - Dutch LassySmall
    - Dutch Alpino
    - Danish DDT
    - Naija NSC
    - Afrikaans AfriBooms
2. Treebanks not in UD:
    - ConvBank (English)
    - Tweebank (English)

D Standard deviations development data

Components	8	16	32	64	128
POS	**81.2**	80.4	81.0	79.4	79.4
	0.41	0.65	0.28	0.60	0.46
UAS	72.7	71.4	70.8	72.6	**73.8**
	0.96	1.0	1.16	0.95	0.49
LAS	55.2	55.1	54.8	55.5	**57.1**
	0.47	0.89	0.76	0.85	0.42

Table 10: POS, UAS and LAS scores (mean over 5 random seeds) and standard deviations on dev data for 2000 sentences. Highest scores are marked in bold.

Sentences	1000	2000	4000
POS	78.5	79.4	**79.8**
	0.26	0.46	0.87
UAS	71.5	**73.8**	73.3
	0.55	0.49	0.96
LAS	54.6	**57.1**	56.0
	1.12	0.42	0.92

Table 11: POS, UAS and LAS scores (mean over 5 random seeds) and standard deviation on dev data for 128 components. Highest scores are marked in bold.