TLT 2021

20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)

Proceedings

To be held as part of SyntaxFest 2021
21–25 March, 2022
Sofia, Bulgaria

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-955917-16-2

Preface

The 20th International Workshop on Treebanks and Linguistic Theories (TLT 2021) follows an annual series that started in 2002 in Sozopol, Bulgaria. TLT addresses all aspects of treebank design, development, and use. "Treebank" is taken in a broad sense, comprising any spoken, signed, or written data augmented with computationally processable annotations of linguistic structure at various levels. For the second time, TLT is part of SyntaxFest, which co-locates four related but independent events:

- The Sixth International Conference on Dependency Linguistics (Depling 2021)
- The Second Workshop on Quantitative Syntax (Quasy 2021)
- The 20th International Workshop on Treebanks and Linguistic Theories (TLT 2021)
- The Fifth Workshop on Universal Dependencies (UDW 2021)

The reasons that suggested bringing these four events together in 2019 still hold in 2021. There is a continuing, strong interest in corpora and dependency treebanks for empirically validating syntactic theories, studying syntax from quantitative and theoretical points of view, and for training machine learning models for natural language processing. Much of this research is increasingly multilingual and cross-lingual, made in no small part possible by the Universal Dependencies project, which continues to grow at currently nearly 200 treebanks in over 100 languages.

For these reasons and encouraged by the success of the first SyntaxFest, which was held in 2019 in Paris, we – the chairs of the four events – decided to bring them together again in 2021. Due to the vagaries of the COVID-19 pandemic, it was eventually decided to push the actual SyntaxFest 2021 back to March 2022. In order not to delay the publication of new research and not to conflict with other events, we decided however to publish the proceedings that you are now reading in advance, in December 2021.

As in 2019, we organized a single reviewing process for the whole SyntaxFest, with identical paper formats for all four events. Authors could indicate (multiple) venue preferences, but the assignment of papers to events for accepted papers was made by the program chairs.

38 long papers were submitted, 25 to Depling, 11 to Quasy, 17 to TLT, and 24 to UDW. The program chairs accepted 30 (79%) and assigned 8 to Depling, 5 to Quasy, 7 to TLT, and 10 to UDW. 22 short papers were submitted, 6 to Depling, 7 to Quasy, 9 to TLT, and 9 to UDW. The program chairs accepted 14 (64%) and assigned 3 to Depling, 3 to Quasy, 3 to TLT, and 5 to UDW.

At the time of this writing, we do not yet know whether SyntaxFest will be a hybrid or purely online event. We regret this uncertainty but are nevertheless looking forward to it very much. Our sincere thanks go to everyone who is making this event possible, including everybody who submitted their papers, and of course the reviewers for their time and their valuable comments and suggestions. We would like to thank Djamé Seddah, whose assistance and expertise in organizing SyntaxFests was invaluable. Finally, we would also like to thank ACL SIGPARSE for its endorsement and the ACL Anthology for publishing the proceedings.

Radek Čech, Xinying Chen, Daniel Dakota, Miryam de Lhoneux, Kilian Evang, Sandra Kübler, Nicolas Mazziotta, Simon Mille, Reut Tsarfaty (co-chairs)

Petya Osenova, Kiril Simov (local organizers and co-chairs)

December 2021

Program co-chairs

The chairs for each event (and co-chairs for the single SyntaxFest reviewing process) are:

- Depling:
  - Nicolas Mazziotta (Université de Liège)
  - Simon Mille (Universitat Pompeu Fabra)
- Quasy:
  - Radek Čech (University of Ostrava)
  - Xinying Chen (Xi'an Jiaotong University)
- TLT:
  - Daniel Dakota (Indiana University)
  - Kilian Evang (Heinrich Heine University Düsseldorf)
  - Sandra Kübler (Indiana University)
- UDW:
  - Miryam de Lhoneux (Uppsala University / KU Leuven / University of Copenhagen)
  - Reut Tsarfaty (Bar-Ilan University / AI2)

Local Organizing Committee of the SyntaxFest

- Petya Osenova (Bulgarian Academy of Sciences)
- Kiril Simov (Bulgarian Academy of Sciences)

Program Committee for the Whole SyntaxFest

Patricia Amaral (Indiana University Bloomington)
Valerio Basile (University of Turin)
David Beck (University of Alberta)
Ann Bies (Linguistic Data Consortium, University of Pennsylvania)
Xavier Blanco (UAB)
Igor Boguslavsky (Universidad Politécnica de Madrid)
Bernd Bohnet (Google)
Cristina Bosco (University of Turin)
Gosse Bouma (Rijksuniversiteit Groningen)
Miriam Butt (Universität Konstanz)
Marie Candito (Universtité Paris 7 / INRIA)
Radek Cech (University of Ostrava)
Giuseppe Giovanni Antonio Celano (University of Pavia)
Xinying Chen (Xi'an Jiaotong University)
Silvie Cinková (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics)
Cagri Coltekin (University of Tuebingen)
Benoit Crabbé (Université Paris 7 / Institut national de recherche en informatique et en automatique, Paris)
Daniel Dakota (Indiana University)
Eric De La Clergerie (Institut national de recherche en informatique et en automatique, Paris)
Felice Dell'Orletta (Institute for Computational Linguistics, National Research Council, Pisa)
Kaja Dobrovoljc (Jožef Stefan Institute)
Kilian Evang (Heinrich Heine University Düsseldorf)
Thiago Ferreira (University of São Paulo)
Ramon Ferrer-I-Cancho (Universitat Politècnica de Catalunya)
Kim Gerdes (Laboratoire Interdisciplinaire des Sciences du Numérique, Université Paris-Saclay)
Carlos Gómez-Rodríguez (Universidade da Coruña)
Jan Hajic (Institute of Formal and Applied Linguistics, Charles University, Prague)
Eva Hajicova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Dag Haug (University of Oslo)
Richard Hudson (University College London)
András Imrényi (Eszterházy Károly Egyetem)
Leonid Iomdin (Institute for Information Transmission Problems, Russian Academy of Sciences)
Jingyang Jiang (Zhejiang University)
Sylvain Kahane (Modyco, Université Paris Ouest Nanterre / CNRS)
Vaclava Kettnerova (Institute of Formal and Applied Linguistics)
Sandra Kübler (Indiana University Bloomington)
Guy Lapalme (University of Montreal)
François Lareau (Observatoire de linguistique Sens-Texte, Université de Montréal)
Alessandro Lenci (University of Pisa)
Nicholas Lester (University of Zurich)
Lori Levin (Carnegie Mellon University)
Haitao Liu (Zhejiang University)

Marketa Lopatkova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Olga Lyashevskaya (National Research University Higher School of Economics)
Teresa Lynn (Dublin City University)
Jan Macutek (Mathematical Institute of the Slovak Academy of Sciences / Constantine the Philosopher University in Nitra)
Robert Malouf (San Diego State University)
Alessandro Mazzei (Dipartimento di Informatica, Università di Torino)
Nicolas Mazziotta (Université de Liège)
Alexander Mehler (Text Technology Group, Goethe-University Frankfurt am Main)
Wolfgang Menzel (Department of Informatics, Hamburg University)
Jasmina Milicevic (Dalhousie University)
Simon Mille (Pompeu Fabra University)
Yusuke Miyao (The University of Tokyo)
Simonetta Montemagni (Institute for Computational Linguistics, National Research Council, Pisa)
Kaili Müürisep (University of Tartu)
Alexis Nasr (Laboratoire d'Informatique Fondamentale, Université de la Méditerranée, Aix-Marseille II)
Sven Naumann (University of Trier)
Anat Ninio (The Hebrew University of Jerusalem)
Joakim Nivre (Uppsala University)
Pierre Nugues (Lund University, Department of Computer Science Lund, Sweden)
Kemal Oflazer (Carnegie Mellon University-Qatar)
Timothy Osborne (Zhejiang University)
Petya Osenova (Sofia University / Institute of Information and Communication Technologies, Sofia)
Robert Östling (Department of Linguistics, Stockholm University)
Agnieszka Patejuk (Polish Academy of Sciences / University of Oxford)
Alain Polguère (Université de Lorraine)
Prokopis Prokopidis (Institute for Language and Speech Processing/Athena RC)
Laura Pérez Mayos (Pompeu Fabra University)
Owen Rambow (Stony Brook University)
Rudolf Rosa (Institute of Formal and Applied Linguistics, Charles University, Prague)
Tanja Samardzic (University of Zurich)
Giorgio Satta (University of Padua)
Nathan Schneider (Georgetown University)
Olga Scrivner (Indiana University Bloomington)
Djamé Seddah (Alpage, Université Paris la Sorbonne)
Alexander Shvets (Institute for Systems Analysis of Russian Academy of Sciences)
Maria Simi (Università di Pisa)
Achim Stein (University of Stuttgart)
Reut Tsarfaty (Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot)
Francis M. Tyers (Indiana University Bloomington)
Zdenka Uresova (Institute of Formal and Applied Linguistics, Charles University, Prague)
Gertjan Van Noord (University of Groningen)
Giulia Venturi (Institute for Computational Linguistics, National Research Council, Pisa)
Veronika Vincze (Hungarian Academy of Sciences, Research Group on Articial Intelligence)
Relja Vulanovic (Kent State University at Stark)

Chunshan Xu (anhui jianzhu university)
Xiang Yu (University of Stuttgart)
Zdenek Zabokrtsky (Institute of Formal and Applied Linguistics, Charles University, Prague)
Amir Zeldes (Georgetown University)
Daniel Zeman (Institute of Formal and Applied Linguistics, Charles University, Prague)
Hongxin Zhang (Zhejiang University)
Yiyi Zhao (Institute of Applied Linguistics, Communication University of China, Beijing)
Heike Zinsmeister (University of Hamburg)
Miryam de Lhoneux (University of Copenhagen)
Marie-Catherine de Marneffe (The Ohio State University)
Valeria de Paiva (Samsung Research America and University of Birmingham)

Additional Reviewers

Chiara Alzetta (National Research Council of Italy)
Aditya Bhargava (University of Toronto)
Lauren Cassidy (Dublin City University)
Simon Petitjean (Heinrich Heine University Düsseldorf)
Xenia Petukhova (National Research University Higher School of Economics)
Daniel Swanson (Indiana University)
He Zhou (Indiana University)
Yulia Zinova (Heinrich Heine University Düsseldorf)

Table of Contents

Typological Approach to Improve Dependency Parsing for Croatian Language

Diego Alves
FFZG, University of Zagreb
Zagreb, Croatia
dfvalio@ffzg.hr

Boke Bekavac
FFZG, University of Zagreb
Zagreb, Croatia
bbekavac@ffzg.hr

Marko Tadić
FFZG, University of Zagreb
Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

This article presents the results of the experiments concerning different typological approaches considering syntactic structures with the aim to identify similar languages which can be combined with Croatian to improve UAS and LAS metrics when using a deep learning tool. From the eight selected languages coming from different linguistic families and genera, we showed that Slovene and Irish are the best candidates which improved significantly dependency parsing results. Slovak is the only language presenting negative synergy when combined with Croatian. Both typological approaches presented in this study, using quantitative data concerning rules from context-free grammar extracted from corpora using Marsagram tool and using syntactic features from lang2vec language vectors, did not allow us to explain the observed synergy when the different languages were combined. The traditional genealogical classification does not explain either the improvement provided by Irish or the negative impact of the Slovak language in both considered metrics.

1 Introduction

Since the 1980s, NLP field has increasingly relied on statistics, probability, and machine learning methods which require a large amount of linguistic data. Furthermore, from 2015 onward, the usage of deep learning techniques has been dominant in this field (Otter et al., 2018). These approaches require a large amount of annotated data which can be problematic for some languages considered as low-resourced.

Linguistic manual annotation of texts can be very costly (Fort et al., 2014), therefore, other solutions for improving PoS-MSD (Part-of-Speech and Morphosyntactic descriptors) and Dependency Parsing tagging scores have been proposed. One way to overcome this issue is to combine data from similar languages according to established typological classifications (Smith et al., 2018)(Alzetta et al., 2020). Although some improvement can be observed, most of these studies, however, do not present a deep analysis of typological features which may play a significant role when corpora are combined. Furthermore, none has considered statistics concerning possible (or impossible) syntactic constructions inside the available training datasets as a possible typological classification.

Therefore, our aim in this paper is to propose an innovative way of considering typological aspects when combining datasets for dependency parsing improvement. The study is focused on the Croatian language and its association with several European languages from different linguistic families. Our hypothesis is that by comparing syntactic rules automatically extracted from Universal Dependencies datasets by inferring context-free grammars (together with its statistics), we are able to classify languages according to these syntactic criteria. Combining languages closer in terms of syntactic structure to train deep learning parsing models should improve final LAS and UAS metrics.

The paper is composed as follows: Section 2 presents related work to this topic. Section 3 describes the campaign design: datasets selection, typological classification strategies, and extrinsic evaluation using trained models; Section 4 present the obtained results which are discussed in Section 5. In Section 6 we provide conclusions and possible future directions for research.

2 Related Work

Combining data from multiple languages has the ultimate aim of creating Universal Morphological and Dependency Parsing systems by considering the relationship between different languages morphology and syntactic structure (Otter et al., 2018). The Universal Dependencies (UD) framework (Nivre et al., 2020) proposes a robust set of rules for annotating parts of speech, morphological features, and syntactic dependencies across different human languages, and is inserted in this strategy as it allows multi-lingual data to be annotated with the same set of tags.

Udify tool (Kondratyuk and Straka, 2019) proposes an architecture aimed for PoS-MSD and dependency parsing tagging integrating Multilingual BERT language model[1] (104 languages) (Pires et al., 2019). It can be fine-tuned using specific corpora (mono or multilingual) to enhance overall results. The authors showed that by using a corpus composed by the association of all Universal Dependencies training sets, there is a considerable improvement in the results of parsing for low-resourced languages. Nevertheless, the authors did not conduct an experiment based on typological features to test the potential of the model when only similar languages are combined.

An interesting example of the usage of typological features to improve results of NLP methods was presented by (Üstün et al., 2020). They proposed UDapter, a tool that uses a mix of automatically curated and predicted typological features obtained via URIEL language typology database (Littell et al., 2017). These features were used as direct input to a neural parser as language-typology vectors and results showed that they were crucial for the improvement of the dependency parsing accuracy for low-resourced languages. A similar study, using different deep learning architecture had been performed by (Ammar et al., 2016), however, in both cases, there is no detailed analysis on which features were the most relevant.

The above-mentioned language typology database offers the lang2vec tool (Littell et al., 2017) which provides uniform, consistent and standardized information about languages drawn from typological, geographical and phylogenetic databases. Its sources include WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran and McCloy, 2019), Ethnologue (Lewis, 2009), and Glottolog (Hammarström et al., 2020). While (Üstün et al., 2020) used lang2vec in an automatized way to cluster languages, (Naseem et al., 2012) selected specific typological features to fine-tune effective automatic annotation of data from languages with no available training sets.

The strategy proposed by (de Lhoneux et al., 2018) concerns sharing 27 parameters using Uppsala parser using pairs of languages from the same linguistic family, showing that general typological classifications can already contribute to enhancing final results on low-resourced languages. They also observed that by combining features even from unrelated languages overall scores can be improved in some specific cases. Nevertheless, as it is the case for most of the similar studies, no specific linguistic analysis was presented in order to explain why languages coming from different families can improve overall results.

An interesting and detailed experiment was conducted by (Lynn et al., 2014) concerning the Irish language. The authors performed a series of cross-lingual direct transfer parsing for the Irish language and the best results were achieved when using Indonesian, a language from the Austronesian language family. They also propose some analysis considering similarities between the treebanks of both languages in terms of dependency parsing labels, however, detailed statistical analysis of corpora and complete comparison of specific typological features were not carried out.

Concerning syntax more specifically, (Alzetta et al., 2020) presented a study whose main objective was to identify cross-lingual quantitative trends in the distribution of dependency relations in annotated corpora from distinct languages by using an algorithm (LISCA - LInguiStically– driven Selection of Correct Arcs) (Dell'Orletta et al., 2013) capable of detecting patterns of syntactic constructions in large datasets. Only four Indo-European languages were scrutinised but some interesting insights concerning languages peculiarities were observed.

Another approach to extract and to compare syntactic information from treebanks is proposed by (Blache et al., 2016) by inferring context-free grammars (together with its statistics) from syntactic struc-

[1] https://github.com/google-research/bert/blob/master/multilingual.md

Language	Linguistic Family	Genus	UD Dataset	Corpus size (K tokens)
Bulgarian	Indo-European	Slavic	BTB	156
Croatian	Indo-European	Slavic	SET	199
Greek	Indo-European	Greek	GDT	63
Hungarian	Uralic	Ugric	Szeged	42
Irish	Indo-European	Celtic	IDT	115
Latvian	Indo-European	Baltic	LVTB	220
Maltese	Afro-Asian	Semitic	MUDT	44
Slovak	Indo-European	Slavic	SNK	106
Slovene	Indo-European	Slavic	SSJ	140

Table 1: Selected Languages, corresponding Linguistic Families and Genus, and corresponding UD datasets information (v.2.7).

tures inside annotated corpora. The analysis comparing 10 different languages showed the potential of the proposed tool (MarsaGram), however, like (Alzetta et al., 2020), the authors do not explore how this information can be used to improve existing NLP tools, which is the main objective of this paper.

3 Campaign Design

In this section, we describe the corpora that have been selected, the typological classification methods that were considered, and the experimental design used to evaluate the effects on dependency parsing metrics of the combination of different training datasets.

3.1 Languages and Datasets selection

As mentioned before, the focus of this study is the Croatian language. The main idea is to combine its training dataset with other European languages to improve UAS and LAS scores. From all 24 European Union official languages, we have chosen the following ones for our experiments: Bulgarian, Greek, Hungarian, Irish, Latvian, Maltese, Slovak, and Slovene. We have decided to work with European languages as this ensemble already provides languages from diverse linguistic families and allows us to test our hypothesis.

All the selected languages have Universal Dependencies datasets (version 2.7) and were chosen as they have only one UD corpus. Slovene is the exception, it has two different UD datasets but one is composed of spoken language, therefore, the other available corpus (written language) was used. The choice of including Slovene is also due to its genealogical proximity to Croatian.

Table 1 presents the languages involved in the experiment, with the respective linguistic family and genus (from Worl Atlas of Language Structure Online[2]) and the size of their UD corpora (Version 2.7).

3.2 Typological Analysis

In this study, we propose to compare the chosen languages using two different typological approaches. One considers the statistical analysis of context-free grammar rules extracted from dependency parsing trees using the software Marsagram, while the other strategy uses information from lang2vec tool language vectors.

3.2.1 Statistical comparison of Dependency Parsing Trees

Marsagram is a tool for exploring treebanks, it extracts context-free grammars (CFG) from annotated datasets that allow statistical comparison between languages as proposed by (Blache et al., 2016). We have used the latest release of this software[3] developed by ORTOLANG. This software has been chosen as it allows easy extraction and analysis of surface word order patterns which have never been used before as a way to interpret results of language combination for training deep learning models.

[2]https://wals.info/languoid/genealogy

[3]Available at: https://www.ortolang.fr/market/tools/ortolang-000917

Approach	Number of Rules
All rules, all properties	714 399
All rules, only linear properties	96 789
Common rules, all properties	1 912
Common rules, only linear properties	247

Table 2: Different approaches for the statistical typological approach and the respective number of the considered syntactic rules.

For this analysis, we have combined train, development, and test sets, and extracted quantitative information about its syntactic properties for each language. Distance matrices were, then, generated using R.

This software identifies four types of properties: precede, require, exclude, and unicity. The extracted syntactic rules contain information concerning part-of-speech and dependency parsing label as well as the associated property type. For example: *NOUN-conj_precede_CCONJ-cc_DET-det* which means that a *CCONJ* which has the dependency relation *cc* precedes a *DET* with *det* as dependency label in the context of a node having *NOUN* as head. Marsagram also indicates the frequency of each rule inside the corpus.

In the previous work (Blache et al., 2016), the authors proposed two different analyses: considering all possible properties or taking into account only the linear property (precede). They have shown that the linear approach was better for classifying languages typologically as results were closer to classic genealogical lists. Nevertheless, in our study, we still consider both scenarios in order to analyse which one is better when the aim is to combine languages for improving dependency parsing metrics.

For each language, Marsagram generates a specific set of rules and the percentage corresponding to its frequency inside the corpus. Some rules are common to all languages and some of them appear only in one or a few corpora. Therefore, the typological classification can be done by considering all possible identified rules (frequency equal to zero for languages in which the rule does not appear) or, considering only the rules present in all corpora (common rules).

Thus, we have 4 different possible comparisons which are presented in Table 2 together with the number of syntactic rules and considered properties used in each one.

3.2.2 Comparison using Language Vectors

Lang2vec is a library[4] that allows simple queries of the URIEL database which are presented as language vectors (Littell et al., 2017). For this study, we have considered syntactic information (*syntax_average* option). For example: *S_NEGATIVE_SUFFIX* which gives a value of 1 if the language has a negative suffix and 0 if it does not have, and *S_SUBJECT_AFTER_VERB*, 1 for languages in which the subject appears after the verb and 0 otherwise.

One disadvantage of this tool is that for some languages, not all information is available. If all official European Union are considered, the number of existing syntactic properties in lang2vec is 103. However, Croatian has values for only 12 of them. As our focus is this language, we have considered the syntactic features for which Croatian has associated values [5]. The distance between languages was calculated using cosine similarity. Among the other selected languages, only Maltese and Slovak do not have values for all these features and, therefore, were discarded for this specific analysis.

3.3 Training Models

We have selected Udify tool to train dependency parsing modules using the combined corpora as it allows fine-tuning of Multilingual BERT language model and for which the authors showed that multilingual

[4]https://pypi.org/project/lang2vec/

[5]Selected syntactic features: *S_SVO, S_SOV, S_VSO, S_VOS, S_OVS, S_OSV, S_SUBJECT_BEFORE_VERB, S_SUBJECT_AFTER_VERB, S_OBJECT_AFTER_VERB, S_OBJECT_BEFORE_VERB, S_SUBJECT_BEFORE_OBJECT, and S_SUBJECT_AFTER_OBJECT*.

Combination	Number of added sentences	Ratio
Smaller	450	94% Croatian, 6% other language
Medium	909	88% Croatian, 12% other language
Larger	1 662	81% Croatian, 19% other language

Table 3: Information concerning the different combinations of the Croatian training set and other languages.

corpus can potentially enhance overall results (specially for under-resourced languages) (Kondratyuk and Straka, 2019). Training parameters were defined as:

- Number of epochs: 80
- Warmup: 500
- Baseline training set: Croatian SET
- Development and test sets: Croatian SET

Our baseline is the result obtained by training Udify using the Croatian Universal Dependencies training set (SET) which contains 6 914 sentences. To obtain statistical significance, for each test using a specific dataset we have conducted 6 experiments varying the Random Seed value in the configuration file of Udify: standard value, 13370 (proposed by the developers), 10, 100, 1000, and 100000. For each test, we have calculated the standard deviation and the p-value when compared to the baseline.

As explained before, the objective is to combine the Croatian dataset with annotated data of the other selected languages. We have combined its training set with three different sizes of the other languages annotated data as detailed in table 3.

One problem is that each training set has a different size, thus, to have homogeneity in terms of size to allow results to be compared, we have decided to add the first 909 sentences of the second language training corpus to the Croatian one. This value corresponds to the size of the Hungarian training set (the smallest one among the chosen languages and, therefore, being totally used), this limitation concerning the Hungarian language is what determined the ratio of all language combinations.

The final size of the combined training sets is 7 823 sentences (88% Croatian and 12% from the other language).

4 Results

In this section, we present the typological classification of the languages obtained using the methods presented previously followed by the results of the combination of the different datasets.

4.1 Typological classification using statistics from syntactic trees

Tables 5 shows the distance between each language and Croatian concerning the different choices of rules and properties selection using Marsagram.

In the scenario considered in the second column of table 4 (considering all rules and properties), we observe that Slovene and Slovak are closest to Croatian (all Slavic languages), however, Bulgarian, which is also Slavic, comes after Greek, Maltese, Hungarian and Latvian which are from different genealogical families.

The third column of table 4 shows the results of the analysis of all rules but considering only the linear properties (*precede*). Again, Slovene and Slovak are the most similar to Croatian, followed by Greek. When only linear properties are considered, Latvian and Irish are classified as closer to Croatian compared to the previous scenario. Bulgarian, again beside being Slavic, occupies the second to last position.

When only common rules are considered (fourth column of table 4), Slovene is still the closest one to Croatian, however, in this case, Bulgarian is classified as much closer. Slovak loses the second position to Latvian. Maltese, Greek, and Hungarian are the most distant languages.

Language	d(All/All)	d(All/Linear)	d(Common/All)	d(Common/Linear)
Slovene	68.0	21.4	4.0	1.1
Slovak	69.4	24.0	4.9	1.2
Greek	70.0	24.5	5.9	1.6
Maltese	73.7	24.8	5.8	1.1
Hungarian	77.2	26.4	6.2	1.5
Latvian	78.5	24.5	4.3	1.0
Bulgarian	80.0	25.5	4.4	1.1
Irish	80.6	25.2	5.3	1.7

Table 4: Distance from Croatian using Marsagram results, first word correspond to the type of rules considered and the second word to the type of properties.

Language	Distance
Slovene	0.01
Bulgarian	0.03
Latvian	0.11
Greek	0.11
Hungarian	0.12
Irish	0.51

Table 5: Cosine distances calculated between Croatian language vector and other languages considering syntactic features from lang2vec.

Finally, when only common rules and linear properties are taken into account (fifth column of Table 4), we observe important changes in the classification. Slovene is no longer classified as the closest to Croatian. Maltese, and Bulgarian are the closest ones (second and third position) behind Latvian only.

Typological classification differs when different sets of rules and properties are considered. Slovene and Slovak are most of the time the closest languages to Croatian which was expected considering that they are all Slavic languages. These results show that it is difficult to determine which type of choice concerning rules and properties is the most adapted for syntactical language classification. Results may be biased by size, genre, and also the type of sentences composing the corpora (for example: length of sentences and syntactic complexity).

4.2 Typological classification using similarity between language vectors

By using cosine distance between the language vectors built with syntactic features from lang2vec, we obtain the classification present in Table 5.

Both Slavic languages (Slovenian and Bulgarian) are the most similar to Croatian, therefore more coherent to the typical genealogical classification of languages. As mentioned before, Slovak, also Slavic, does not have values for the analysed features and was therefore excluded from this comparison. Latvian, Greek, and Hungarian have similar distances, but much higher than the ones concerning Slavic languages and Irish is the most distant one.

4.3 Dependency parsing results with combined corpora

In tables 6, 7, and 8 we present the UAS and LAS values obtained when Udify was trained using the Croatian training set alone (baseline) and with the combined datasets (Croatian associated with another language) with three different ratios, as well as the delta when compared to the baseline. Each result corresponds to the mean value calculated with the six different trials using different Random Seed initial values. Highlighted results concern the experiments for which p-value is inferior to 0.05. Development and test sets were purely Croatian.

When the smaller ratio is used to train Udify (94% Croatian, 6% other language), we observe that only Bulgarian, Greek and Irish contribute positively in increasing both UAS and LAS metrics. Association

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Smaller)	**92.38**	**0.06**	**89.05**	**0.06**
Croatian + Greek (Smaller)	**92.40**	**0.09**	**89.07**	**0.08**
Croatian + Hungarian (Smaller)	92.33	0.02	88.98	-0.01
Croatian + Irish (Smaller)	**92.42**	**0.11**	**89.09**	**0.10**
Croatian + Latvian (Smaller)	92.39	0.07	88.98	0.00
Croatian + Maltese (Smaller)	92.32	0.01	88.97	-0.01
Croatian + Slovak (Smaller)	92.24	-0.07	**88.89**	**-0,09**
Croatian + Slovene (Smaller)	92.36	0.05	89.02	0.04

Table 6: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (94% Croatian, 6% other language).

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Medium)	92.35	0.03	89.02	0.03
Croatian + Greek (Medium)	92.35	0.03	89.98	-0.01
Croatian + Hungarian (Medium)	92.33	0.02	89.01	0.02
Croatian + Irish (Medium)	**92.43**	**0.12**	**89.07**	**0.08**
Croatian + Latvian (Medium)	92.26	-0.06	88.92	-0.06
Croatian + Maltese (Medium)	92.36	0.04	88.97	-0.01
Croatian + Slovak (Medium)	**92.21**	**-0.11**	**88.89**	**-0,09**
Croatian + Slovene (Medium)	**92.42**	**0.10**	**89.09**	**0.10**

Table 7: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (88% Croatian, 12% other language).

of Croatian and Irish being the one providing the highest increase. Negative synergy is only observed for LAS metric when Croatian is combined with Slovak.

For the medium ratio (88% Croatian, 12% other language), combinations of Croatian with Irish and with Slovene provide a positive synergy. As for the smaller ratio, when Croatian is combined with Slovak, there is a negative synergy which is, this time, observed for both UAS and LAS.

Concerning the larger ratio (81% Croatian, 19% other language), again the combination of Croatian and Slovak decrease significantly both UAS and LAS metrics. The corpus composed by both Croatian and Irish no longer provides a positive synergy. The only significant increase is obtained for LAS metric when Croatian is combined with Slovene.

5 Discussion

By analysing the UAS and LAS results presented in the previous section, it is possible to observe that Bulgarian, Greek, Irish, and Slovene training corpora have the potential to improve UAS and LAS metrics when combined with the Croatian training dataset. However, results strongly depend on the ratio between Croatian sentences and the other combined language. Bulgarian and Greek languages provided a positive synergy only for the smaller ratio, while the combination with Irish was positive for both smaller and medium ratios. Slovene did not improve the metrics for the smaller ratio but had a positive impact for both medium and larger ones. What is clear for all three ratios is the strong negative impact of Slovak when this language is associated with Croatian.

In their article, (Kondratyuk and Straka, 2019) presented results for Croatian from a model which was trained combining 124 languages. The obtained UAS and LAS values are respectively 91.10 and 86.78. It is possible to see that all the models presented in this study are higher than these, even for our baseline and for the combination with Slovak. Thus, it seems that finding typological ways to combine languages wisely and on the smaller scale is more effective.

Training Corpus	UAS	delta UAS	LAS	delta LAS
Croatian (baseline)	92.32	-	88.99	-
Croatian + Bulgarian (Larger)	92.33	0.01	88.97	-0.02
Croatian + Greek (Larger)	92.34	0.03	89.99	-0.01
Croatian + Hungarian (Larger)	-	-	-	-
Croatian + Irish (Larger)	92.37	0.05	89.03	0.04
Croatian + Latvian (Larger)	92.33	0.01	88.97	-0.02
Croatian + Maltese (Larger)	-	-	-	-
Croatian + Slovak (Larger)	**92.20**	**-0.11**	**88.83**	**-0,16**
Croatian + Slovene (Larger)	92.36	0.04	**89.06**	**0.07**

Table 8: UAS and LAS metrics obtained by training Udify with different training datasets: Croatian alone and associated with other languages (81% Croatian, 19% other language). Hungarian and Maltese training corpora do not have enough annotated sentences to be combined with Croatian in this specific ratio.

Moreover, (de Lhoneux et al., 2018) included the Croatian language in their study and the LAS obtained was 77.9, also inferior to the values in our experiments. However, the combined languages were not the same.

In terms of typology, if we consider the traditional genealogical classification of languages, we can state that being part of the same linguistic family and genus do not guarantee a positive synergy when corpora are combined. Even though Bulgarian and, especially, Slovene can improve the final results when combined with Croatian, Slovak, which is also from the same genus, is the only language with a negative influence in all tested scenarios. Moreover, Irish, which is from a different genus is a good candidate for improving UAS and LAS results when combined with Croatian.

If we consider the classifications established using Marsagram, it is not possible to find any correlation between the classification lists considering the syntactic criteria with the observed results from Udify. Slovene is the closest language to Croatian when all rules are considered (with all properties considered and only linear ones too) and also when only common rules are compared. However, the calculated distances between Irish and Croatian do not explain the improvement obtained by associating both languages. Also, Slovak does not appear as being the most distant language when compared to Croatian, a result that would explain the negative synergy observed when its corpus is combined with the Croatian dataset.

One possible explanation for this lack of correlation may come from the fact that the distances were calculated using the results obtained by Marsagram which were composed of rules coming from the whole Universal Dependency datasets for each language. However, when Udify experiments were conducted, only a small part of the respective corpora have been used. Therefore, a more precise correlation may be possible if distances are calculated using only the sentences that have been added to the combined training corpus. Another aspect that may need further research concern the homogeneity of extracted rules using Marsagram from subcorpora of a dataset from a single language. It may be possible that the variation inside a corpus may be higher than when two different languages are compared. This case could be accommodated with the usage of controlled content, i.e. parallel corpora of languages investigated. However, this is not always available, particularly for under-resourced languages.

Furthermore, the selected corpora have different sizes and different contents. It may impact heavily the type of syntactic patterns that were extracted using Marsagram. The number of patterns obtained seems to be correlated with the size of the corpus. A comparison using parallel corpora could avoid this bias.

Moreover, positive synergies may not be caused by the whole ensemble of extracted rules but maybe by specific syntactic relations which are shared by the associated languages. Further qualitative analysis of similarities between Irish and Croatian Marsagram results should be conducted.

When analysing the typological classification using lang2Vec, Slovene and Bulgarian are the closest

to Croatian, which we can relate to the positive synergy observed in Udify results. However, Irish is the most distant one which is contradictory with the improvement obtained for both UAS and LAS in two different scenarios. Also, as Slovak does not have values for the selected syntactic features, it was impossible to check whether the combination with Croatian has any negative impact. Thus, even though this tool is a powerful instrument to compare languages, in the approach described here, it seems limited. The idea of combining corpora to improve parsing is most useful for under-resourced languages, and, unfortunately, some of these languages are also under-resourced in terms of language vector information in lang2vec. For example, from the 103 possible syntactic features, the Croatian language only has values for 12 of them.

Considering all the aspects presented above, we can affirm that none of the genealogical and typological approaches were able to explain precisely what was observed when different languages were combined to Croatian.

6 Conclusions and Perspectives

In this article, we presented different approaches to identify languages that can be combined with Croatian to improve dependency parsing evaluation metrics (UAS and LAS) when using Udify deep learning tool.

The possible typological classifications were compared to the results obtained when combining the Croatian training dataset to other European languages from different linguistic families to train Udify models. Three different association ratios were used.

We showed that the association of Croatian with Irish and Slovene languages showed the best positive synergy, increasing UAS and LAS for at least two different combination ratios. Moreover, from all selected languages, the only one which decreased significantly in both metrics is Slovak.

These results show that the classical genealogical classification of languages is not enough to explain the observed phenomena. Slovak and Slovene are from the same linguistic family and genus as Croatian but with totally different impacts on the final results. Also, the Irish language does not belong to the same genus as Croatian, nevertheless, it helped improve UAS and LAS significantly.

The two typological approaches proposed in this paper, using rules from a context-free grammar with Marsagram and comparing lang2vec syntactic features of language vectors, also did not allow us to predict the results obtained when languages were combined. Slovene is identified as the closest language to Croatian in three out of four different analysed Marsagram scenarios. However, the classification of Irish and Slovak does not correspond to the influence these languages have when combined with Croatian. Moreover, the lang2vec classification shows Irish as being the least similar to Croatian, and, unfortunately, Slovak was not analysed due to the lack of syntactic information of this language in this tool.

The study presented in this article was conducted only for Croatian, therefore, we intend to test this approach with other under-resourced languages, also enlarging the selection of languages to be combined to understand better the existing synergies and, also, possible exceptions as the one that has been identified in this article concerning the association between Croatian and Irish.

For future research we will check the quality of Slovak data because it consistently differ from other Slavic languages although genealogically and culturally Slovak is closely connected to Croatian.

Furthermore, our aim is to conduct a more detailed analysis concerning Marsagram results, first, checking the homogeneity of rules extracted from different subcorpora of the same language, and, secondly, using only the sentences that were appended to the combined training corpora to calculate the distances.

Acknowledgements

The work presented in this paper has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 812997 and under the name CLEOPATRA (Cross-lingual Event-centric Open Analytics Research Academy).

References

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Petya Osenova, Kiril Simov, and Giulia Venturi. 2020. Quantitative linguistic investigations across universal dependencies treebanks. In Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Philippe Blache, Stéphane Rauzy, and Grégoire Montcheuil. 2016. MarsaGram: an excursion in the forests of parsing trees. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2336–2342, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Miryam de Lhoneux, Johannes Bjerva, Isabelle Augenstein, and Anders Søgaard. 2018. Parameter sharing between dependency parsers for related languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4992–4997, Brussels, Belgium, October-November. Association for Computational Linguistics.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistically–driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 17.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Karen Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Gamification for Information Retrieval (GamifIR'14) Workshop*.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2020. *Glottolog 4.3*. Jena.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April. Association for Computational Linguistics.

Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *Proceedings of the First Celtic Language Technology Workshop*, pages 41–49, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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, May. European Language Resources Association.

Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2018. A survey of the usages of deep learning in natural language processing. *CoRR*, abs/1807.10854.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online, November. Association for Computational Linguistics.

The *RigVeda* goes "universal": annotation and analysis of equative constructions in Vedic and beyond

Erica Biagetti
University of Pavia
erica.biagetti01@universitadipavia.it

Abstract

By presenting a case study on Rigvedic equative and similative constructions, this paper demonstrates that treebanks constitute an important support for research in historical linguistics for two main reasons. First, by providing quantitative evidence on linguistic phenomena, they can confirm or dismiss hypotheses formulated on the base of qualitative data. Second, by capturing correlations among linguistic phenomena which could hardly be grasped by linguists' naked eye, treebank-based analyses allow scholars to formulate new hypotheses. Since an analysis of Rigvedic equative constructions calls for a granular and informative annotation scheme, the Vedic Treebank implements the UD scheme for equative constructions with sub-relations; while some such extensions were specifically designed for a study on Rigvedic similes, others might be adopted by every treebank developer interested in representing equative strategies.

1 Introduction

Historical linguistics has always relied on collections of written texts, i.e., corpora, which constitute the only source of evidence available for ancient languages. Annotated corpora revolutionized historical linguistics because they allow scholars to automatically retrieve large quantitative evidence on linguistic phenomena whose account has been previously based on qualitative evidence and to capture correlations among them which could hardly be grasped by linguists' naked eye (Eckhoff et al., 2018: 303; Biber, 2009; Anthony, 2013). Furthermore, morphosyntactically annotated corpora require automatic data selection through explicit query expressions, crucially making historical linguistic research replicable (Haug, 2015).

By presenting a case study on Rigvedic equative and similative constructions, in this paper I provide further evidence for the relevance of treebanks for the study of ancient languages. The *Rigveda* (RV) is a collection of 1028 hymns, dating back to the second half of the second millennium BCE (Witzel, 1995), which constitutes the oldest layer of Vedic literature and whose language is strongly conditioned by the poetic and ritual character of the text. The division of the collection into ten books reflects the internal chronology of the work. The core of the collection and its oldest part are books II to VII (the so-called "Family Books"), whereas book X is the most recent. Books I, VIII, and IX are generally younger than the Family Books.

The Rigvedic treebank was created as part of the larger Vedic Treebank (VTB; Hellwig et al., 2020; Biagetti et al., 2021), a corpus of selected passages from Vedic Sanskrit literature syntactically annotated according to the Universal Dependency (UD) standard.[1] The VTB is maintained within the Digital Corpus of Sanskrit,[2] which provides a web-based interface for collaborative dependency annotation. A first version of the treebank was published in occasion of the release of UD version 2.6

[1] Although the UD standard covers most of the syntactic phenomena found in Vedic texts, some constructions require special attention during annotation and their annotation scheme within the VTB may deviate slightly from the official UD scheme (see the annotation guidelies available at: https://github.com/OliverHellwig/sanskrit/tree/master/papers/2020lrec/paper). While some such deviations were removed in occasion of the treebank release within the UD platform, others remain and are fully documented in Hellwig et al. (2020).

[2] http://www.sanskrit-linguistics.org/dcs/index.php?contents=texte

(15 May 2020); a new version, revised and considerably expanded, is currently under development (Hellwig and Sellmer, forthc.).

The case study presented in this paper is part of a project devoted to the study of Rigvedic similes. Similes, which are the most frequent trope found in the RV, are explicit comparative constructions that owe their figurative meaning to the fact that the compared entities are felt as being fundamentally unlike each other, and therefore unlikely to be compared (Israel et al., 2004). While the language of the RV disposes of different strategies for the encoding of comparison, equative and similative constructions introduced by the particles *ná* 'as, like', *iva* 'as, like' and *yáthā* (/*yathā*) 'as, like' have specialized for the encoding of figurative comparison. The aim of this paper is to demonstrate that a treebank-backed study on the syntax of these constructions allows us not only to understand their synchronic distribution, but also to confirm previous hypotheses on their origin and development, as well as to formulate new ones. Such a study calls for a granular and informative annotation scheme, which is able to capture the different strategies employed in the RV for the expression of comparison of equality; therefore, a second, major purpose of this paper is to present a new annotation scheme based on the UD standard for comparative constructions implemented with sub-relations.

The paper is organized as follows: Section 2 introduces the main strategies employed in the RV for the encoding of comparison of equality, among which we find similes introduced by *ná*, *iva*, and *yáthā*. After summarizing UD guidelines for the annotation of equative constructions (3.1), Section 3.2 introduces the implemented annotation scheme adopted by the VTB for the analysis of such constructions. Section 4.1 shows that quantitative data extracted from the treebank can provide interesting insights about the syntax and origin of Rigvedic similes. Section 5 suggests extending part of the enhanced scheme to other languages and constructions. Section 6 contains the conclusions.

2 Comparison of equality in the RV

Equative and similative constructions encode similarity between a comparee (CPREE) and a standard (STAND) with respect to some action or property, called parameter (PAR), and by means of a standard marker (STM; Haspelmath and Buchholz, 1998; Treis, 2017). While equative constructions encode quantitative comparison of equality (e.g. *Peter is as tall as Susan*), similative constructions encode qualitative comparison, or comparison of manner (e.g. *Peter runs like a hare*.)

In the RV, constructions introduced by the STMs *ná*, *iva*, and *yáthā* constitute the main strategy for the encoding of comparison of equality. They are characterized by systematic ellipsis of the verb in the STAND and by case transparency (Haspelmath and Buchholz, 1998: 307), i.e., identity of case and function between CPREE and STAND (Bergaigne 1887; Jamison 1982; Pinault 1997). Quantitative and qualitative comparison are encoded by the same constructions and are therefore nearly impossible to distinguish (henceforth: equatives). Rigvedic equatives occur in three main configurations of CPREE(s) and STAND(s). Single equatives can take an adjectival predicate as PAR or a verbal one, as in (1).[3]

(1)	*ví*	*ślóka*	*etu*	***pathyà***	***iva***	*sūréḥ*
	LP	signal_call.NOM	go.IMPV.3SG	pathway.NOM	like	patron.GEN
	PAR-	CPREE-	-PAR	STAND	STM	-CPREE

'Let the signal-call of the patron go forth afar like a pathway.'[4] (RV 10.13.1)

Double equatives are characterized by the presence of two parallel elements in the CPREE and in the STAND, and thus have a gapping structure (2). Less often, equatives may be triple, with CPREE and STAND consisting of three elements each.

(2)	*matáyaḥ*	*rihánti …*	*índraṁ*	***vatsáṁ***	***ná***	***mātáraḥ***
	thought.NOM.PL	lick.PRS.3PL	Indra.ACC	calf.ACC	like	mother.NOM.PL
	$CPREE_i$-	PAR	-$CPREE_j$	$STAND_j$-	STM	$STAND_i$

[3] In glosses, the nominal number is specified only if it is plural or dual while gender is specified only if it is feminine or neuter (singular and masculine are not indicated). Among verbal categories, indicative mood and active voice are not indicated.
[4] Translations of Rigvedic passages are taken from Jamison and Brereton (2014).

‘Thoughts lick … Indra like mothers a calf.’ (RV 3.41.5)

Besides being employed in syntagmatic comparison, the accented particle *yáthā* also introduces comparative clauses, whose main clause often contains a correlative adverb such as *evá* ‘so, in this way’ in (3). Note that the difference between clausal and syntagmatic comparison is not limited to the presence vs. absence of a verb: while in the former *yáthā* functions as a subordinator and occurs in clause-initial position, in the latter *yáthā* (with its unaccented variant *yathā*), *ná*, and *iva* have a clitic behavior and follow the STAND.

(3)	***yáthā***	***jaghántha***	*dhr̥ṣatấ́*	*purấ́*	*cid*
	like	smite.PF.2SG	boldly	before	PTCL
	evấ́	***jahi***	*śátrum*	*asmấ́kam*	*indra*
	so	smite.IMPV.2SG	rival.ACC	1PL.GEN	Indra.VOC

‘Just as you also smote boldly before, so smite our rival, o Indra.’ (RV 2.30.4cd)

Finally, comparison of equality can be expressed in the RV by a number of other constructions, including comparative compounds as in (4), adjectives meaning ‘same’ (*samá-*), or less grammaticalized strategies involving a verb whose meaning is ‘reach’ (“reach equatives” in Haspelmath et al., 2017), as in (5). For comparison and gradation in Vedic, see Kulikov (2021).

(4)	***agní-bhrājaso***	*vidyúto*	*gábhastiyoḥ*
	fire-flash.NOM.PL	lightning_bolt.NOM.PL	fist(M/F).LOC.DU
	STAND-PAR	CPREE	

‘Lightning bolts flashing like fire (are) in your fists.’ (RV 5.54.11c)

(5)	*nákiṣ*	*ṭáṁ*	*kármaṇā*	***naśan***
	no_one	3SG.ACC	ritual_work.INST	reach.SUBJ.AOR.3SG
	CPREE	STAND-STM	PAR	PM

‘No one can equal [lit. reach] him (Agni) in his ritual work.’ (RV 8.31.17)

3 Annotating Rigvedic similes

3.1 UD annotation scheme for equative constructions

UD guidelines provide annotation schemes for both basic and clausal equatives. In the former, the standard is linked to the parameter via the relation `obl`, while the standard marker depends on the standard via `case` (Figure 1). In clausal equatives, the verb of the comparative clause is attached to the main verb through `advcl`, the standard marker depending on it via `mark` (Figure 2).

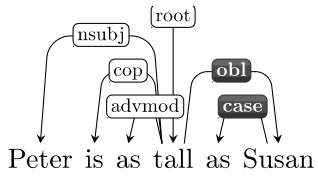


Figure 1. Basic equatives.

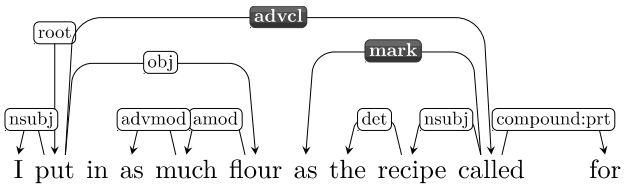


Figure 2. Clausal equatives.[5]

Gapping occurring in comparative constructions is treated in the same way as coordinate gapping. Thus, in the Swedish equative in (6), the promoted element *Joakim* takes the relation that the elided verb would otherwise bear (`advcl`), *tennis* takes the `orphan` relation, and the standard marker *än*, being a functional element, retains its relation `mark` (Figure 3).

(6)	*Dan*	*spelar*	*badminton*	*bättre*	*än*	*Joakim*	*tennis*
	Dan	play.PRS	badminton	better	than	Joakim	tennis

[5] https://universaldependencies.org/u/overview/specific-syntax.html#comparatives

'Dan plays badminton better than Joakim (does) tennis.'[6]

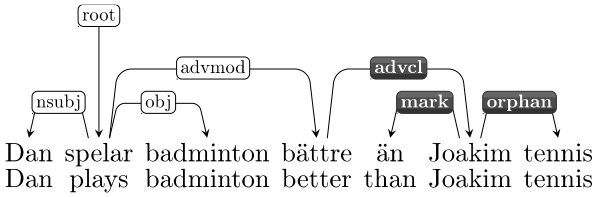


Figure 3. Annotation scheme for gapping in comparison.

3.2 Extending the scheme: language-specific relations

In UD, there are no relations designed specifically to mark equative constructions. First, UD adopts the same scheme for equality and inequality comparison. Furthermore, basic comparatives are simply assimilated to other obliques (`obl`), whereas clausal equatives are treated in the same way as other adverbial clauses (`advcl`). Similarly, standard markers take the same *deprel* as other function words such as adpositions (`case`) and subordinating conjunctions (`mark`). Take for instance the two trees in Figure 4, where the clausal comparative contained in the first sentence takes the same labels as the temporal clause contained in the second.

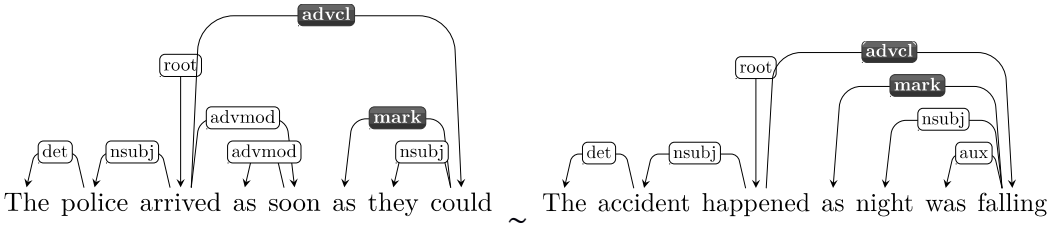


Figure 4. UD scheme for adverbial clause modifiers. L: comparative clause; R: temporal clause.[7]

In Early Vedic, the particles *ná*, *iva*, and *yáthā*/*yathā* have other functions beside that of standard marker of equative constructions: for instance, when employed as a subordinator, Vedic *yáthā* also introduces temporal, final, causal, and content clauses with verbs of knowing and saying (Delbrück 1888: 592-596). Furthermore, as we have seen in Section 2, Vedic has at its disposal several strategies for the encoding of comparison of equality.

Following the UD scheme, it would be possible to extract, e.g., all basic equatives featuring a gapping structure by retrieving all nodes a) that are not a finite verb, b) whose *deprel* is `advcl`, c) that have a child whose *deprel* is `mark` and d) that have at least another child whose *deprel* is `orphan`. In order to exclude other types of subordinate clauses characterized by gapping structure, it would also be necessary to specify e) the lemma of the former child. Even so, one would obtain all basic equatives introduced by *ná*, *iva*, and *yáthā* (and not subordinates introduced, e.g., by *yád* 'that'), but also other subordinates introduced by *yáthā* that present an elided verb. Cf. Figure 5:

```
cat rv.conllu | udapy -TM util.Mark node='a) node.feats["VerbForm"] == ""
and b) node.deprel == "advcl" and c) len([x for x in node.children if x.deprel
== "orphan"]) == 1 and d) len([x for x in node.children if x.deprel == "mark"
and e) x.lemma in ("na", "iva", "yathā")]) == 1' | less -R
```

Figure 5. Udapi[8] query: 'display all basic equatives with gapping structure'.

[6] https://universaldependencies.org/workgroups/comparatives.html
[7] https://universaldependencies.org/docs/u/dep/advcl.html
[8] https://udapi.github.io

Such query would also prevent one from detecting and isolating hybrid constructions such as the one in (7), whose standard has no verb, as in syntagmatic comparison, but in which *yáthā* precedes the standard, as in clausal comparison.

(7) ***yáthā*** *naḥ* *pitáraḥ* *párāsaḥ* *pratnā́so …*
like 1PL.GEN father.NOM.PL further.NOM.PL ancient.NOM.PL
śúcī́d *ayan* *dī́dhitim* *ukthaśā́saḥ*
blazing.ACC.N come.SUBJ.3PL vision(F).ACC reciting_praise.NOM.PL
'Like our further forefathers of old […], those reciting solemn speech (now) will come to the blazing (udder of sacrifice [=Vala]), to visionary power.' (RV 4.2.16)

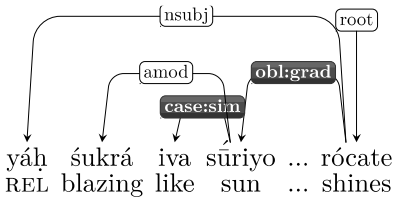
In order to represent the syntax of equatives in detail and to be able to make granular and targeted queries on different types of constructions, the VTB makes use of language-specific extensions that enrich the universal dependency taxonomy. Like language-specific extensions found in UD, extensions employed within the VTB are regarded as subtypes of existing UD relations and have the format `universal:extension`: for instance, `obl:manner` stands for `manner` extension of the UD relation `obl`. As in UD, extensions employed within the VTB are neither recursive nor multi-dimensional, which means that one node can instantiate at most one subtype of a universal relation. However, the VTB allows the user to employ a considerably high number of sub-relations for research-related purposes, provided that such extensions are fully documented in the guidelines.

Table 1 summarizes the scheme employed by the VTB for equative constructions.

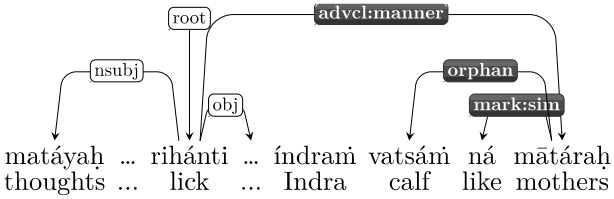
Table 1. Equative constructions with their respective annotation.

CONSTRUCTION	EXAMPLE	ANNOTATION (dependent → `relation` → head)
PREDICATIVE SIMILE	'Agni is like the sun.'	*sun* → `root` *sun* → `nsubj` → *Agni* *sun* → `case:sim` → *like*
SIMILE WITH ELLIPSIS	'Agni shines like the sun.'	*shines* → `obl:grad` → *sun* → `case:sim` → *like*
	'The lightning bellows like a cow.'	*bellow* → `obl:manner` → *cow* → `case:sim` → *like*
SIMILE WITH GAPPING	'Thoughts lick Indra like mothers a calf.'	*lick* → `advcl:manner` → *mothers* → `mark:sim` → *like;* *mothers* → `orphan` → *calf*
CLAUSAL SIMILE	'Just as you drank the previous soma drinks, so take a drink today.'	*drink* → `advcl:manner` → *drank* → `mark` → *as*; *drank* → `obj` → *previous drinks*; *drink* → `advmod` → *so*

As shown by Table 1, the VTB formally distinguishes simple, basic equatives (annotated with `obl` and `case`) from double equatives characterized by gapping structure (annotated with `advcl` and `mark`). As we have seen in Section 2, Vedic employs the same standard marker for equative and similatives; in order to be able to observe any syntactic difference in the expression of quantitative and qualitative comparison, for example in the order of constituents, the sublabels `:grad` and `:manner` are given on a lexical basis to dependents of gradable and non-gradable adjectives respectively.

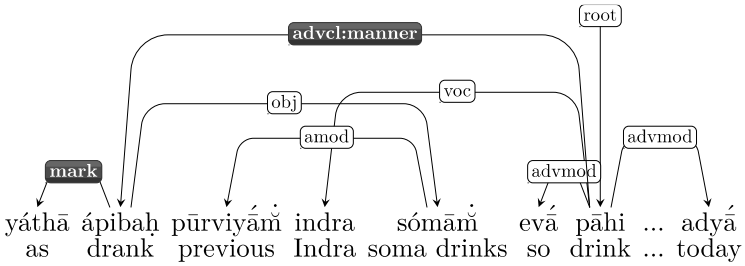


‘He (Agni) who shines like the blazing sun.’ (RV 1.43.5)
Figure 6. Extended scheme for simple equatives.



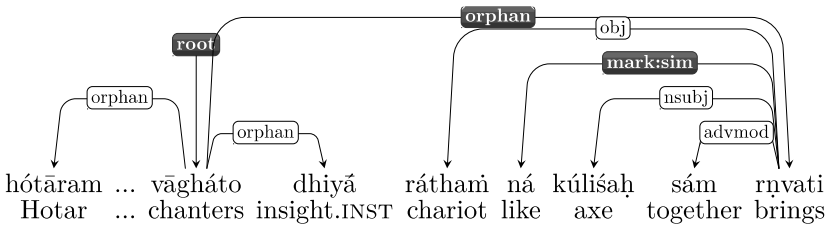
‘Thoughts lick Indra like mothers a calf.’ (RV 3.41.5)
Figure 7. Extended scheme for equatives with gapping structure.

The sublabel `:sim`[9] attached to the relations `case` and `mark` allows the user to easily retrieve all particles that introduce basic equatives and to distinguish them from those that introduce clausal similes (which take `mark` alone). Compare for instance the annotation of basic equatives like those in Figure 6 and Figure 7 with that of a clausal equative like the one in Figure 8:



‘Just as you drank the previous soma drinks, Indra, so take a drink today.’ (RV 3.36.3cd)
Figure 8. Extended scheme for clausal equatives.

In some cases, the verb is exceptionally constructed with the standard rather than with the comparee. As shown by Figure 9, such cases are also captured by the annotation scheme.[10]



‘As an axe brings together a chariot, the chanters **Ø** the Hotar with their insight.’[11] (RV 3.2.1)
Figure 9. Annotation of equatives whose verb is constructed with STAND.

[9] `:sim` stands for “simile”.
[10] In this example, we would expect a plural verb *sám ṛṇvati* in agreement with the comparee *vāghátas*.NOM.PL ‘chanters’; the verb *sám ṛṇvati*.PRS.3SG ‘brings’ agrees instead with the nominative singular *kúliśaḥ* ‘axe’ which constitutes the standard of the simile. As a whole, the sentence is treated similarly to a case of leftward gapping in coordination

4 Treebank-based analysis of Rigvedic similes

Despite employing different standard markers, Rigvedic comparisons introduced by *ná*, *iva*, and *yáthā*, constitute a coherent construction from the point of view of both syntax and semantics. Syntactically, they have a syntagmatic nature and present clitic standard markers; semantically, they are specialized for figurative comparison and can be defined as similes in all respects.

With the support of extant literature on the origin of Rigvedic similes, quantitative evidence provided by the treebank can help understanding how different particles came to be employed in the kind of constructions attested in the RV. In particular, four groups of queries run on a corpus of 857 similes[12] yielded interesting results in this regard. Queries employed in this study are reported in Appendix A.
Before presenting the results, two premises are in order. First, due to his complex internal chronology, the RV constitutes a diachronic corpus, thus lending itself to the study of language change. Second, in presenting word-order patterns attested in similes, I will only take similes introduced by *ná* and *iva* into account: basic equatives introduced by *yáthā* occur only 76 times in the RV and thus do not lend themselves to quantitative studies on word order (Levshina et al., fortch.).[13]

1. Query: Factors determining the relative order of STAND - PAR in Rigvedic similes

Typological studies on equative and similative constructions have shown that the STAND - PAR order correlates with the OV order (Andersen, 1983; Haspelmath et al., 2017: 26). Rigvedic similes feature STAND - PAR order in 60% of cases, a result which is in line with the fact that Vedic shows a preference for OV while also allowing the opposite order.[14]

Besides a language word-order preferences, heaviness is also responsible for the relative order of standard and parameter. As show by Table 2, similes with gapping, whose standard consists of at least two arguments of the verb, have PAR - STAND order more frequently than simple similes (62% vs. 52%). In turn, Table 3 shows that the percentage of STAND - PAR order is especially high (68%) in those similes whose standard consists of a single element (e.g., *pitā́ iva* 'like a father', *putrám ná* 'like a son'), and it decreases to 57% in those similes whose standard has adjectival, participial, or genitive modifiers (e.g., *nítyaṁ ná sūnúm* 'like a dear son').

Table 2. Order of STAND and PAR in similes a) with ellipsis and b) with gapping.

ORDER	SIMILES WITH ELLIPSIS		SIMILES WITH GAPPING	
STAND-PAR	360	**62%**	151	**52%**
PAR-STAND	212	37%	134	47%
TOTAL	572		285	
p-value (χ^2 test)	**0.0064**			

Table 3. Order of STAND and PAR in a) similes with ellipsis and simple STAND, and b) similes with ellipsis and complex STAND.

ORDER	ELLIPSIS AND SIMPLE STAND		ELLIPSIS AND COMPLEX STAND	
STAND-PAR	197	**68%**	163	**57%**
PAR-STAND	91	31%	121	42%
TOTAL	288		284	
p-value (χ^2 test)	**0.0083**			

Finally, the percentage of PAR - STAND order is increased by the high frequency of thetic sentences (e.g. *The telephone's ringing*), which in Vedic have verb-initial order (Lambrecht, 1994: 143; Viti, 2008). Cf. example (8):

[12] The annotated portion of the RV is available at: https://github.com/EricaBiagetti/VTB_Rigveda.
[13] Differently from *ná* and *iva* similes, whose origin is disputed, we do not need quantitative evidence in order to confirm the emergence of *yáthā* similes from comparative clauses and the consequent cliticization of the subordinator.
[14] In the annotated portion of the RV in the VTB (24109 tokens in 3092 sentences) OV occurs in 63% of cases. However, Ryan and Gunkel (2015) have shown that, in metrically neutral contexts, non-imperative finite verbs display OV order in 78% of cases (37 in total) and imperative forms in 77% of cases (22 in total).

(8) ***próthad*** *áśvo* *ná* *yávase* *aviṣyán*
snort.INJ.PRS.3SG horse.NOM like pasture(N).LOC eager.NOM
'He has snorted like a hungry horse in his pasture.' (RV 7.3.2a)

Knowing which factors determine the order of standard and parameter helps envisaging diachronic tendencies in the development of equative constructions as attested in the RV, presented in points 2 to 4 below.

2. Query: Frequency of STAND - PAR and PAR - STAND orders in *iva* e *ná* similes

Two main hypotheses have been proposed in the literature on the development of *ná* similes: a) according to Vine (1978), they derive from coordinate negative constructions with ellipsis of the verb in the second conjunct, (9); b) according to Pinault (1985), they stem from the so-called negative parallelism, i.e., a rhetorical device typical of Baltic and Slavic folk literature, consisting of two sentences, the first of which presents a negation and optional ellipsis of the negated verb (10).

Thus while, according to Vine, similes introduced by *ná* originate from constructions in which the PAR (verb) preceded the STAND, according to Pinault they stem from constructions with the opposite order of STAND and PAR:

(9) Coordinate negative constructions: PAR - STAND
ná *ta* *indra* *sumatáyo* *ná* *rā́yaḥ*
NEG 3PL.NOM.N Indra.VOC favor(F).NOM.PL NEG rich.NOM.PL
saṁcákṣe ***pū́rvā*** ***uṣáso*** ***ná*** ***nū́tnāḥ***
enumerate.DAT earlier.NOM.PL.F dawn(F).NOM.PL NEG/like recent.NOM.PL.F
'Neither your favors nor your riches, o Indra, can be entirely surveyed, through the previous dawns, nor through the current ones.' > 'Neither your favors nor your riches, O Indra, can be entirely surveyed, just like the previous and the current dawns (cannot be entirely surveyed).' (RV 7.18.20)

(10) Negative parallelism: STAND - PAR
vér ***ná*** *druṣác*
bird.NOM NEG/like wood_sitting.NOM
camúvor *ā́* *asadad* *dhárīḥ*
cup(F).LOC.DU LP seat.AOR.3SG tawny.NOM
'It is not a bird sitting in the wood, the tawny one (Soma) has taken his seat in the two cups.' > 'Like a bird sitting in the wood the tawny one has taken his seat in the two cups.' (RV 9.72.5)

Observing the relative order of standard and parameter separately for *iva* and *ná* similes, we gain some important insights on the origin of these constructions. Table 4 shows that simple similes introduced by *ná* have STAND - PAR order more frequently than those introduced by *iva* (68% vs. 60%). While this difference is statistically only weakly significant (χ^2 test, p-value 0.06), the picture changes if we focus on similes whose standard is composed of one single element, with no modifiers: here, the percentage of STAND - PAR order reaches 78% with standards marked by *ná*, against 63% of standards marked by *iva* (p-value 0.013). On the contrary, no significant difference can be observed in word-order patterns of similes with gapping, since *ná* and *iva* similes of this type show STAND - PAR order in 54% and 52% of cases respectively.

Table 4. N. of STAND - PAR and PAR - STAND orders in simple similes and in simple similes whose standard consists of only one element.

SIMILE TYPE	ALL SIMPLE SIMILES				STANDARD = ONE ELEMENT			
STM / ORDER	*iva* similes		*ná* similes		*iva* similes		*ná* similes	
STAND - PAR	114	**60%**	234	**68%**	65	**63%**	121	**78%**
PAR - STAND	76	40%	108	32%	37	37%	33	22%
TOTAL	190		342		102		154	
p-value (χ^2 test)	**0.06**				**0.013**			

If we assume that, in the absence of other syntactic and pragmatic factors presented under point 1, similes tend to retain the original relative position of standard and parameter, the fact that simple *ná* similes have a more marked preference for the STAND - PAR pattern than *iva* similes may constitute an important clue in favor of their origin from the negative parallelism (Pinault 1985), where the standard always precedes the verb. The fact that the preference for the STAND - PAR order is less marked for *iva* similes, on the other hand, may support the hypothesis of its origin as a marker for syntagmatic comparison, which does not tie the standard to any position with respect to the parameter (see points 2 and 3). Finally, the fact that *ná* and *iva* similes behave in the same way in the presence of gapping would be due to the heaviness of the standard in such constructions.

Turning to semantics, the origin of *ná* equatives from negative parallelism provides some interesting insights on their specialization for figurative comparison: in negative parallelism, the subject of the first clause usually represents a prototype participant of the action or quality expressed by the verb and thus lends itself to figurative readings.[15]

3. Query: equatives whose verb (PAR) is construed with the STAND, and not with CPREE

Query number 2 returns five cases in which the verb is constructed with a standard introduced by *ná* (as in Figure 9) and three cases in which *yáthā* occurs in a hybrid construction, as the one presented in (7). In contrast, the query does not return any case in which a standard marked by *iva* is clearly constructed with the verb. If we interpret such cases as remnants of a stage in which both the comparee and the standard clause could contain a verb, the presence of such evidence in *ná* similes confirms point 2 on the clausal origin of the latter; accordingly, the lack of such evidence in *iva* similes may suggest that *iva* has always introduced syntagmatic comparison.

4. Query: frequency of equatives with gapping structure

If, as suggested by point 3, *iva* similes were always syntagmatic, we can assume that they originally had simpler standards and that only later allowed gapping structure on the model of *ná* similes (which, as suggested by point 2 and 3, originally contained a verb). By dividing the corpus into the ten books that make up the RV, we can check whether similes with gapping became more frequent in younger books (I, VIII-X) than they were in older ones (II-VII). Table 5 reports the frequencies of simple similes and similes with gapping introduced by *iva* and *ná* throughout the ten books; note that, if the whole RV is considered (last raw), the ratio of simple and gapped standards is virtually the same for *iva* and *ná* similes.

Table 5. Percentage of simple similes and of similes with gapping in each book.

	iva similes				*ná* similes			
Book	Simple similes		With gapping		Simple similes		With gapping	
I	22	56%	17	**44%**	63	58%	45	**42%**
II	31	**76%**	10	24%	17	65%	9	**35%**
III	12	**75%**	4	25%	14	67%	7	**33%**
IV	7	**78%**	2	22%	16	**73%**	6	27%
V	19	**90%**	2	10%	13	**72%**	5	28%
VI	10	**67%**	5	33%	32	**70%**	14	30%
VII	10	**67%**	5	33%	27	64%	15	**36%**
VIII	25	62.5%	15	**37.5%**	30	**68%**	14	32%
IX	19	59%	13	**41%**	71	**74%**	25	26%
X	35	55%	29	**45%**	59	**71%**	24	29%
Total	190	**65%**	102	**35%**	342	**67%**	164	**32%**
p-value	**0.01**				**0.024**			

[15] Furthermore, Pinault (1985: 138-143) suggests that the comparative reading of *ná* must have spread thanks to the existence of comparative compounds (e.g. *vā́ta-jūta-* lit. 'wind-swift') and comparisons with an ablative STAND (e.g. *manáso.*ABL *jávīyas* 'swifter that thought'), which shared the STAND - PAR order with the negative parallelism. Comparative compounds are known cross-linguistically for their preference for generic comparisons (Haspelmath and Buchholz, 1998) and, at least within the IE domain, idiomatic ablative comparatives are also often employed in this function (cf. the type Latin *melle dulcior* 'sweeter than honey').

Table 5 suggests that gapping structure did indeed become more common for *iva* similes in younger books: a significant difference can be observed between, e.g., 9% of similes with gapping in book V and 43% in book I, or 45% in book X. Similes introduced by *ná* present a different picture: while book I has indeed the higher percentage of similes with gapping (41%), these were already frequent in old books such as II, III, and VII. In fact, Kruskal-Wallis tests suggest that older and younger books differ from each other in the frequency of *iva* similes with gapping (p-value 0.01) as well as in the frequency of *ná* similes with gapping (p-value 0.02). Due to the low absolute counts reported in Table 5, the tests do not point to clear diachronic differences in the structure of *ná* and *iva* similes and suggest that the issue should be investigated further on a larger data set.

To sum up, with the partial exception of point 4, results obtained from the four queries suggest that equative constructions introduced by *ná* and *iva* probably influenced each other: by systematic ellipsis of the negated verb in the negative parallelism, *ná* similes became syntagmatic and the standard marker *ná* developed a clitic behavior;[16] *iva* similes, on the other hand, specialized for figurative comparison and started to feature gapping structure.

5 Thinking big: cross-linguistic extensions

As anticipated above, the annotation scheme presented in Section 3.2 was developed within a project devoted to the study of Rigvedic similes. As showed in Section 4, the introduction of language-specific extensions made it possible to perform precise, quantitative analyses on the syntax of Rigvedic similes; however, some language-specific extensions would be superfluous if employed in analyses of more general interest or for languages other than Early Vedic.

This suggests that, in view of the next UD release, some extensions might be discarded whereas other might be considered for employment in other treebanks. For instance, the distinction between standard markers of clausal and phrasal equatives, which in the VTB are annotated as `mark` and `mark:sim` respectively, should be discarded as the difference between such constructions results in the presence vs. absence of a verb in the standard. Furthermore, the information stored in the `:manner` and `:grad` extensions should be moved to the MISC field of the CoNLL-U format and assigned on a lexical basis to the parameter, depending on whether it encodes a gradable or non-gradable quality.[17]

More interesting is the possibility of extending the relation subtype `:sim` to standard markers of equative and similative constructions in other languages and construction types. In many languages, standard markers of equative constructions can be identical with conjunctive particles and subordinators (Haspelmath et al., 2017): remaining within the Indo-European domain, cf. Latin *ut* 'as, how', which introduces several other kinds of subordinate clauses. Beside particles and conjunctions, standards of equatives and similatives can be marked by adpositions or by case markers. When the parameter marker is expressed by an adjective or verb, the standard is marked by a case selected by the governing adjective or verb: cf. the Latin adjective *consimilis* in (11) and the Ancient Greek participle *eidómenon* in (12), both governing a dative standard. Figure 10 shows the suggested annotation scheme for example (12).

(11)	*harum*	*est*	***consimilis***	***capris***	*figura*
	this.GEN.PL	be.PRS.3SG	similar.NOM	goat.DAT.PL	shape.NOM
			PM	STAND.STM	CPREE

'their shape (scil. of elks) is similar to [that of] goats' (Caes. *Gall.* 6.27.1; Ittzés 2021: 479)

(12)	*ēlthé*	*moi*	*phásma*	***eidómenon***	***Arístōni***
	come.AOR.3SG	1SG.DAT	phantom.NOM	resemble.PTCP.PRS.NOM	Ariston.DAT
			CPREE	PM	STAND.STM

'A phantom came to me that resembled Ariston.' (Herodotus 6.69.1; de Kreij 2021: 350)

[16] Note that negative *ná*, from which comparative *ná* derives (cf. Pinault, 1985), stands either in clause-initial position or before the predicate.

[17] Note that the CoNLL-U format adopted by the DCS does not include a MISC field. This determined the choice of extending the syntactic relations `obl` and `advcl` of the STAND with semantic information pertaining the whole construction such as `:manner` and `:grad`.

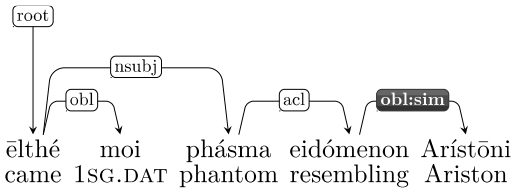
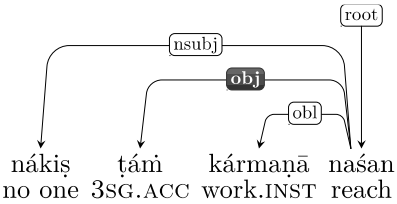


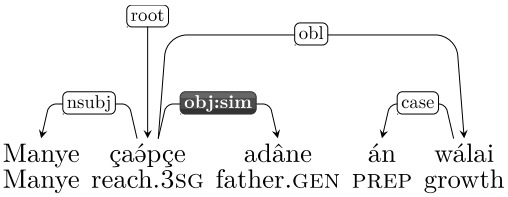
Figure 10. Annotation scheme for example (12).

Extending the relation subtype `:sim` would allow accounting for equative and similative constructions that are otherwise not covered by the UD taxonomy. This is the case, for instance, of reach equatives such as (5), which are tagged like usual transitive clauses in the UD scheme (Figure 11). While in Early Vedic such constructions are sporadic and scarcely grammaticalized (Biagetti 2021),[18] in some languages they constitute a major comparison strategy and extending their annotation scheme would enhance the possibility of studying equative constructions cross-linguistically.[19] The extended annotation for reach equatives is illustrated by Figure 12 from Malgwa (Chadic; Löhr, 2002: 107).



'No one can reach him in his ritual work.'

Figure 11. UD scheme for reach equatives.



'Manye reaches her father in growth.'

Figure 12. Extended scheme for reach equatives.

The reason for adding relation subtypes to standard markers and not to parameter markers of equative constructions is suggested by Haspelmath et al. (2017: 25) Generalization 1, according to which "[n]o language has only a degree-marker, leaving the standard unmarked". In other words, while constructions such as "Kim is **Ø** tall **like** Pat" are cross-linguistically common, constructions such as "Kim is **equally/as** tall **Ø** Pat" are not attested; thus, marking only standard markers with relation subtypes would allow capturing all types of equatives while avoiding redundancy. Finally, assigning the label `:sim` to elements of equative constructions would allow distinguishing them from elements of comparative constructions proper, which encode comparison of inequality (Treis 2017) and are marked by the extension `:cmpr` in some treebanks.[20]

6 Conclusion

By presenting a case study on Rigvedic equative constructions, in this paper I argued that treebanks constitute an important support to research in historical linguistics because they allow to confirm or dismiss previously formulated hypotheses (see especially query 2) and to observe correlations between language phenomena that could hardly be grasped by the naked eye (queries 1, 3, and 4). However, the need to account for formal variations or hybrid constructions that may play a role in language change sometimes calls for more granular and informative annotation schemes. In the case of Rigvedic similes, I suggested implementing the UD scheme for equative and similative constructions with sub-relations; crucially, such extensions are not meant to be language specific and some of them might be adopted by every treebank developer interested in representing equative constructions.

[18] With Dixon (2012), we might say that they constitute comparative strategies rather than constructions proper.

[19] See for instance the examples from Malgwa (Chadic), Malian Tamashek (Berber), or Zay (Semitic) in Haspelmath et al. (2017: 21-22).

[20] Treebanks of Latin, Polish, and Tamil employ `obl:cmpr` for comparative oblique arguments and `advcl:cmpr` for comparative clauses. While the former is limited to comparison of inequality, the latter is instantiated with examples of clausal equatives. In order to increase consistency, I suggest limiting `advcl:cmpr` to proper comparative clauses and adding a new relation subtype (such as `:sim`) for clausal equatives. Note, in passing, that Telugu employs `obl:cmp` and Moksha `obl:comp` with the same purpose of `obl:cmpr`. Finally, Erzya employs `advmod:comp` for adverbs functioning as standard markers in comparatives proper. Cf. https://universaldependencies.org/ext-dep-index.html.

Acknowledgements

I wish to thank the three anonymous reviewers who, with their comments and suggestions, have provided insightful feedback for a substantial improvement of this work. I am grateful to Oliver Hellwig for his constant support, from the first annotation tests to the final stages of this article, and to Salvatore Scarlata for the lengthy discussions on what was the best way to annotate Rigvedic similes. Final responsibility remains my own.

Reference

Andersen, Paul Kent. 1983. *Word Order Typology and Comparative Constructions*. John Benjamins, Amsterdam.

Anthony, Laurence. 2013. A critical look at software tools in corpus linguistics. *Linguistic research* 30(2): 141–161.

Bergaigne, Abel. 1887. La syntaxe des comparaisons védiques. Mélanges Renier, 75-101. Vieweg, Paris.

Biagetti, Erica. 2021. *Ṛgvedic similes: a corpus-based analysis of their forms and functions*. PhD thesis. University of Pavia.

Biagetti, Erica, Oliver Hellwig, Salvatore Scarlata, Elia Ackermann, and Paul Widmer. 2021. Evaluating Syntactic Annotation of Ancient Languages: Lessons from the Vedic Treebank. *Old World: Journal of Ancient Africa and Eurasia* 1, no. 1: 1–32.

Biber, Douglas. 2009. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*, Bernd Heine & Heiko Narrog (eds), 159–191. Oxford University Press, Oxford.

Delbrück, Berthold. 1888. Altindische syntax. Verlag der Buchhandlung des Waisenhauses.

Dixon, Robert M.W. 2012. *Basic Linguistic Theory*. Volume 3. Further Grammatical Topics. Oxford University Press, Oxford.

Eckhoff, Hanne, Bech, Kristin, Bouma, Gerlof, Eide, Kristine, Haug, Dag, Haugen, Odd Einar & Jøhndal, Marius. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1), 29–65.

Gunkel, Dieter, and Kevin Ryan. 2015. Investigating Rigvedic word order in metrically neutral contexts. Handout. Vienna. https://www.academia.edu/40393688/Investigating_Rigvedic_word_order_in_metrically_neutral_contexts

Haspelmath, Martin & Buchholz, Oda. 1998. Equative and similative constructions in the languages of Europe. In *Adverbial Constructions in the Languages of Europe*, Van der Auwera, Johan (ed.), 277–334. Mouton de Gruyter, Berlin.

Haspelmath, Martin & the Leipzig Equative Constructions Team 2017. Equative constructions in world-wide perspective. In *Similative and Equative Constructions: A Cross-linguistic Perspective*, Yvonne Treis & Martine Vanhove (eds.) 9–32. John Benjamins, Amsterdam.

Haug, Dag T. T. 2015. Treebanks in historical linguistics research. In *Perspectives on Historical Syntax*, Carlotta Viti (ed), 187–202. John Benjamins, Amsterdam.

Hellwig, Oliver, Scarlata, Salvatore, Ackermann, Elia & Widmer, Paul. 2020. The Treebank of Vedic Sanskrit. In *Proceedings of The 12th Language Resources and Evaluation Conference* (LREC 2020), Nicoletta Calzolari, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi *et al.* (eds.), 5137–5146.

Hellwig, Oliver and Sven Sellmer. Forthcoming. The Vedic Treebank. In Erica Biagetti, Chiara Zanchi, and Silvia Luraghi, *Building New Resources for Historical Linguistics*. Pavia: Pavia University Press.

Israel, Michael, Jennifer Riddle Harding, and Vera Tobin. 2004. On simile. *Language, culture, and mind* 100.

Jamison, Stephanie W. 1982. Case disharmony in Rigvedic similes. *Indo-Iranian Journal* 24, no. 4: 251–271.

Jamison, Stephanie W. & Brereton, Joel P. 2014. *The Rigveda: the Earliest Religious Poetry of India*. Oxford University Press, New York.

de Kreij, Nina. 2021. 13 Ancient Greek. In Götz Keydana, Wolfgang Hock, and Paul Widmer (eds.), *Comparison and Gradation in Indo-European*. Berlin: De Gruyter Mouton, 349–384.

Kulikov, Leonid. 2021. Gradation in Old Indo-Aryan. In Comparison and Gradation in Indo-European, Keydana, Götz, Wolfgang Hock and Paul Widmer (eds.),385–416. The Mouton Handbooks of Indo-European Typology, 1. De Gruyter Mouton, Berlin / Philadelphia.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Levshina, Natalia, Savithry Namboodiripad, et al. Forthcoming. *Why we need a gradient approach to word order*.

Löhr, Doris. 2002. *Die Sprache der Malgwa (Nárá Málgwa)*. Frankfurt: Peter Lang.

Pinault, Georges. 1885. Négation et comparaison en védique. *Bulletin de la société de linguistique de Paris* 80, no. 1:103–144.

Pinault, Georges. 1997. Distribution des particules comparatives dans la Rik-Samhitâ, *Bulletin d'Études Indiennes* 13-14, 307–367.

Stassen, Leon 1985. *Comparison and Universal Grammar*. Basil Blackwell, Oxford.

Treis, Yvonne. 2017. Comparative Constructions: An Introduction. In Treis, Yvonne & Martine Vanhove (Eds.). 2017. *Similative and equative constructions: A cross-linguistic perspective* (Vol. 117). John Benjamins, Amsterdam.

Vine, Brent. 1978. On the metrics and origin of Rig-Vedic *ná* 'like, as'. *Indo-Iranian Journal* 20, no. 3: 171-193.

Viti, Carlotta. 2008. The verb-initial word order in the early poetry of Vedic and Homeric Greek. In Karlene Jones-Bley, Martin E. Huid, Ângela Della Volpe, and Miriam Robbins Dexter (eds.), *Proceedings of the 19th Annual UCLA Indo-European Conference* (Los Angeles, November 2nd – 3rd 2007), Selected Papers, 89–111.

Witzel, Michael. 1995. Ṛgvedic History: Poets, Chieftains and Polities. *In The Indo-Aryans of Ancient South Asia*, George Erdosy (ed.), 307–352. De Gruyter Mouton, Berlin.

Appendix A: Queries

This Appendix contains all the queries employed for the case study presented Section 4. All queries were written in Udapi query language (https://udapi.github.io).

Query 1:

a. N. of STAND - PAR and PAR - STAND orders in all similes

```
cat RV.conllu | udapy util.See node='node.deprel in ("advcl:manner", "obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā")]) == 1'
```

b. N. of STAND - PAR and PAR - STAND orders in all similes with ellipsis

```
cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case:sim"]) == 1'
```

c. N. of STAND - PAR and PAR - STAND orders in all similes with gapping

```
cat RV.conllu | udapy util.See node='node.deprel in ("advcl:manner") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "mark:sim"]) == 1'
```

d. N. of STAND - PAR and PAR - STAND orders in similes with ellipsis and simple STAND

```
cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case:sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
```

e. N. of STAND - PAR and PAR - STAND orders in similes with ellipsis and complex STAND

```
cat RV.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma in ("na", "iva", "yathā") and x.deprel == "case:sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) >= 1'
```

Query 2:

a. N. of STAND - PAR and PAR - STAND orders in similes introduced by *ná*:

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel in ("case:sim", "mark:sim")]) == 1'
```

b. N. of STAND - PAR and PAR - STAND orders in similes introduced by *iva*:

```
cat RV.conllu | udapy util.See node='node.deprel in (len([x for x in node.children if x.lemma == "iva" and x.deprel in ("case:sim", "mark:sim")]) == 1'
```

c. N. of STAND - PAR and PAR - STAND orders in *ná*-similes with ellipsis

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel == "case:sim"]) == 1'
```

d. N. of STAND - PAR and PAR - STAND orders in *iva*-similes with ellipsis

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "iva" and x.deprel == "case:sim"]) == 1'
```

e. N. of STAND - PAR and PAR - STAND orders in *ná* similes with ellipsis and simple STAND

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma in ("na") and x.deprel == "case:sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
```

f. N. of STAND - PAR and PAR - STAND orders in *iva* similes with ellipsis and simple STAND

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma in ("iva") and x.deprel == "case:sim"]) == 1 and len([x for x in node.children if x.lemma not in ("na", "iva", "yathā")]) == 0'
```

g. N. of STAND - PAR and PAR - STAND orders in *ná*-similes with gapping

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "na" and x.deprel == "mark:sim"]) == 1'
```

h. N. of STAND - PAR and PAR - STAND orders in *iva*-similes with gapping

```
cat RV.conllu | udapy util.See node='len([x for x in node.children if x.lemma == "iva" and x.deprel == "mark:sim"]) == 1'
```

Query 3:

a. STAND constructed with a finite verb

```
cat RV.conllu | udapy -TM util.Mark node='node.lemma in ("na", "iva", "yathā") and node.deprel in ("case:sim", "mark:sim") and node.parent.upos == "VERB" and node.parent.feats["VerbForm"] == ""' | less -R
```

Query 4:

a. N. of *iva* similes with ellipsis in each book

```
cat rv1.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma == "iva" and x.deprel == "case:sim"]) == 1'
```

b. N. of *ná* similes with ellipsis in each book

```
cat rv1.conllu | udapy util.See node='node.deprel in ("obl:manner", "obl:grad") and len([x for x in node.children if x.lemma == "na" and x.deprel == "case:sim"]) == 1'
```

c. N. of *iva* similes with gapping in each book

```
cat rv1.conllu | udapy util.See node='node.deprel == "advcl:manner" and len([x for x in node.children if x.lemma == "iva" and x.deprel == "mark:sim"]) == 1'
```

d. N. of *ná* similes with gapping in each book

```
cat rv10.conllu | udapy util.See node='node.deprel == "advcl:manner" and len([x for x in node.children if x.lemma == "na" and x.deprel == "mark:sim"]) == 1'
```

Is Old French tougher to parse?

Loïc Grobol[1, 2], Sophie Prévost[3], Benoît Crabbé[4]
(1) Modyco, Université Paris Nanterre et CNRS
(2) LIFO, Université d'Orléans and INSA Centre – Val-de-Loire
(3) Lattice, CNRS, ENS, PSL and Université Sorbonne Nouvelle
(4) LLF, CNRS and Université de Paris
sophie.prevost@ens.psl.eu, lgrobol@parisnanterre.fr, benoit.crabbe@linguist.univ-paris-diderot.fr

Abstract

Medieval French is known to be relatively hard to parse, with several possible sources of confusion for automatic parsers, among which its flexible word order and its graphical and syntactic variation, both synchronically and diachronically. In this work, we study in particular the influence of word order, by comparing the performances of two state-of-the-art syntactic parsers trained and evaluated on two treebanks: the Syntactic Reference Corpus of Medieval French (SRCMF), a treebank of Old French (9th—13th century) and the Google Stanford Dependency treebank of contemporary French.

1 Introduction

Parsing Old French is thought to be hard because the language has flexible word order, graphical and syntactic variation. As a result, automatic parsers are underperforming for Old French as compared with most other Romance languages when accounting to the amount of available data (Zeman et al., 2018).

However, while previous studies (Stein, 2014; Stein, 2016; Guibon et al., 2014; Guibon et al., 2015) have investigated the issue with parsing from an intrinsic point of view, to our knowledge, there is no comparative study of the impact of these characteristics on the behaviour of automatic parsers. In particular, there has been no specific study attempting to assess the impact of Old French free word order on parsing.

In this work, we propose a first step in these directions by studying automatic dependency parsing of Old French as compared to Contemporary French. To this end, we train state-of-the art parsers on the closest alter ego in both languages: the Syntactic Reference Corpus of Medieval French (UD-Old-French-SRCMF, henceforth SRCMF), a treebank of Old French (9th—13th century) and the Google Stanford Dependency treebank of contemporary French (UD-French-GSD, henceforth GSD) ; both from the Universal Dependencies (Nivre et al., 2020) projet.

Both corpora have some similarities —comparable sizes and French language— and some dissimilarities as they represent different stages of the French language, with noticeable linguistic differences between them. Our aim is to assess whether those discrepancies have an impact on the scores of the parser and on the types of errors that they make.

We propose a quantitative and qualitative error analysis with a particular focus on word order, with the following intuitive hypothesis: considering the flexibility of word order as well as the morphological variation in Old French, we expect lower scores on the SRCMF treebank than on the GSD treebank and different types of errors.

2 Data

The Universal Dependencies Old French Syntactic Reference Corpus for Medieval French (SRCMF) treebank (Stein and Prévost, 2013) consists of texts spanning from the 9th to the 13th century. In its

most recent version, it consists of 199 700 tokens (including punctuation marks) in 18 030 sentences for an average of 11.1 tokens per sentence.

Most of the development and test data is taken from texts sharing properties with training data, but conversely pre-1100 texts only appear in the training set because they were deemed too small to reserve anything for testing.

The Universal Dependencies French Google Stanford Dependency (GSD) treebank (Guillaume et al., 2019) consists of Contemporary French data, mainly from encyclopedic articles and tourist reviews. It includes 400 399 tokens for 16 341 sentences (averaging 24.5 tokens per sentence). There is no broad chronological span, but the genre disparities may entail significant internal variability. The split sizes for both corpora are reported in table 1.

	Train		**Dev**		**Test**	
Corpus	Tokens	Sentences	Tokens	Sentences	Tokens	Sentences
SRCMF	158 620	14 153	20 554	1888	20 526	1989
GSD	354 662	14 449	35 718	1476	10 019	416

Table 1: Corpus sizes, using their respective standard splits in Universal Dependencies

There are two explanations for the wide gap between the sentence lengths in the two corpora — which might influence the performances of the parser. The first explanation is a linguistic one, as sentences in Contemporary French tend to be more complex (and thus longer) than in Old French, especially because they include more subordinate clauses. The second one is methodological, and lies in a different representation of coordinated clauses: in SRCMF, any finite verb of a main clause gives rise to a sentence and there is no coordination between main clauses. On the contrary, in GSD, main clauses may be coordinated in a single sentence under specific conditions (if the second verb has no overt subject). Hence, the following example is analysed as a single sentence in GSD while it would analysed as two separate sentences in SRCMF: "*Selon Alan «Dave m'a contacté il y a quelques semaines et m'a demandé si je serais prêt à les rejoindre sur scène»*" ("Dave contacted me a few weeks ago and asked if I would accept to join them on stage").

These differences in sizes could have an influence on the global performances of learned parsers, however neither the direction nor the magnitude of the difference is clear from the current state of the art. Grobol and Crabbé (2021) for instance report better performances on the Sequoia treebank (Candito and Seddah, 2012) than on GSD, despite its smaller size, but worse performances on the French Treebank (Candito et al., 2010), which is larger than GSD. However, since our analyses in this work focus on tree-level behaviours rather than word-level performances, our hypothesis will be that since GSD and SRCMF have a similar number of trees, it makes sense to compare parsers trained on these treebanks.

3 Parser

The parser used in this study is HOPS (Grobol and Crabbé, 2021), a neural graph parser/POS tagger with state-of-the-art results on the Universal Dependencies contemporary French corpora. More specifically, HOPS is variant of Dozat and Manning (2018)'s Biaffine graph parser, that takes transformer language model representations as inputs and where POS-tagging is not an explicit step independent of parsing, but it is instead performed jointly with parsing in a hard parameter sharing (Ruder, 2017) multitask formulation as in e.g. Coavoux and Crabbé (2017). Beyond its sheer performances, our choice was also motivated by the versatility of this parser regarding word representations, as it is able to simultaneously use contextual and non-contextual word embeddings along with character-level representations, which — as noted by Smith et al. (2018) — can have a significant impact for parsing languages with rich morphologies and/or flexible graphic systems. In all our experiments, we used the same hyperparameters as Grobol and Crabbé (2021) for their FlauBERT-based models.

In order to parse UD-French-GSD, we retrain a parsing model from scratch, using a French transformer model, FlauBERT-base (Le et al., 2020), for the contextual word embeddings. The results in terms of POS

tagging, unlabelled attachment and labelled attachment F-scores are reported in table 2 and are similar to those reported by Grobol and Crabbé (2021).

Partition	UPOS	UAS	LAS
Dev	98.63	96.71	95.60
Test	98.61	95.90	94.35

Table 2: Performances of the parser (dev-best model out of 5 random seeds) on GSD development and test partitions.

Grobol and Crabbé (2021) show that using Transformer-based contextual word embeddings (Devlin et al. (2019) among others) greatly improves dependency parsing for contemporary French. In order to benefit from comparable advantages when parsing Old French, we derive adapted contextual embeddings in two different ways: by training a small RoBERTa model from scratch (Micheli et al., 2020) and by further training of FlauBERT (Le et al., 2020), in both cases on a corpus of raw Old French and early Middle French of about 10 Mwords[1]. This results in a situation where despite the resources disparities between Old and Contemporary French in general, the parsers have access to comparable resources for both languages.

Development	UPOS	UAS	LAS
HOPS (scratch)	97.14	92.95	89.18
HOPS (FlauBERT)	97.72	93.70	90.93
Test	**UPOS**	**UAS**	**LAS**
Straka et al. (2019)	96.26	91.83	86.75
HOPS (scratch)	96.60	92.20	87.95
HOPS (FlauBERT)	97.59	93.73	90.98

Table 3: Performances of the parser (dev-best model out of 5 random seeds) on SRCMF development and test partitions.

Table 3 reports the results obtained using these two strategies to obtain contextual word embedding. We note that our parser obtains rather good scores and improves on the state-of-the-art with a considerable margin (which is not very surprising, since unlike Straka et al. (2019) we could rely on specific monolingual contextual word embeddings). Considering these results, the rest of our analyses focus on the better-performing FlauBERT-based model. To preserve the opacity of the test partition and avoid design overfitting (van der Goot, 2021), we will only consider the development set of both corpora in the rest of this work.

4 Comparative analysis

As interesting as UPOS, UAS, LAS may be from a computational point of view, when using automatic parsing as a preprocessing step for a large-scale linguistic analysis, the proportion of trees that are fully correctly parsed is also relevant.

For SRCMF, a total of 1100 sentences (58.26 %) of the sentences are parsed completely correctly and 788 sentences (41.74 %) have at least one parsing error (either a wrong attachment or a wrong dependency label. For GSD, we find 660 sentences (44.72 %) of completely correct parses.

Therefore, somewhat unexpectedly, the parser obtains better results on SRCMF than on GSD for these metrics. However, the picture is different with a more refined analysis. If we focus on the major constituents, that is Subject, Object, Root and Copula, the picture is quite different. There are 1693 sentences

[1]This corpus consists of texts from the BFM (Guillot et al., 2018), AND (Trotter, 2012), NCA (Kunstmann and Stein, 2008), Chartes Douai (Glessgen et al., 2010), OpenMedFr (Wrisley, 2018), Geste (Camps et al., 2019), MCVF (Martineau, 2008) and Chartes Aube (Van Reenen et al., 2006) corpora.

(88.67 %) in SRCMF where there are no errors on these syntactic functions, whereas in GSD this amounts to 1423 sentences (96.21 %).

These results show that the parser is better for non major constituents (such as adverbial phrases or clauses, or internal structure of NPs) on Old French, and for major constituents in contemporary French.

Two complementary observations could explain this difference. First, Old French is characterised by high variations. A word can be spelled in many different ways, null subjects are very frequent, and word order has a considerable flexibility, allowing for preverbal objects, postverbal subjects, and all six permutations of S, V and O (SVO, SOV, OVS, OSV, VSO and VOS), even though SVO is the prevalent form very early on. This multi-dimensional variation probably makes it more complex for the parser to correctly identify subjects, objects (or even verbs) in SRCMF than in GSD. Secondly, as mentioned above, sentences are longer and more complex in GSD than in SRCMF, with either more peripheral elements and/or more complex NPs, which may be difficult to correctly analyse for the parser.

From now on, we will focus on two major constituents, subjects and objects, —both nominal or pronominal— in main declaratives. We thus examine cases of parsing errors of the nsubj and obj relations[2], while taking into account the respective order of the main constituents in the sentence.

We focus first on the orders which are common to both treebanks, in order to determine whether there are, for instance, more errors on the attachment and/or the label of the object in OVS than in SVO in either treebank. Then we will turn more specifically to SRCMF, taking also into account unattested combinations in, trying to highlight some correlations between wrong parsing and the different combinations. One important concern, as historical linguists, is not to miss alternative orders to SVO: even though they are much rarer, they constitute a major feature, and their decrease represents a very important evolution in the history of French. Table 4 reports the error rates for subject and object relations according to the different word orders in both corpora

	SRCMF			**GSD**		
Order	sentences	nsubj errs (%)	obj errs (%)	sentences	nsubj errs (%)	obj errs (%)
V	472	2.12	5.72	105	1.90	0.00
SV	424	4.25	1.65	895	0.78	1.67
VS	173	6.94	2.89	26	3.85	0.00
SVO	119	4.20	16.81	384	0.52	2.34
SOV	109	5.50	5.50	52	0.00	11.54
OVS	71	14.08	15.49	1	0.00	0.00
VO	250	4.80	8.00	13	0.00	0.00
Total	1618	4.80	5.93	1476	0.81	2.24

Table 4: Comparison between the error rates for the nsubj and obj relations in both corpora and in the common word orders.

Table 4 shows that i) in both corpora the parser tends to be more high-performing for nsubj than for obj, with exceptions for the SV, VS and V orders; ii) the parser always performs better on Modern French than on Old French (except for obj in SV, but the difference is insignificant : 1,67 vs 1,65). We now turn to a qualitative analysis of the errors that both corpora have in common: wrong nsubj in V, SV, VS and SVO and wrong obj in SV, SVO and SOV.

4.1 Errors on subjects

V order in GSD expl:subj (PRO *ce*) are parsed as nsubj twice. In SRCMF, we find three main types of errors: either a preverbal oblique is parsed as a nsubj, nsubj is wrongly attached to the root, or nsubj is correctly attached to a wrong root.

[2]We restrict our study to nouns and pronouns, since clausal constituents obey different mechanics. Therefore, we leave csubj and ccomp aside in this work.

SV order in GSD there are two main types of errors: in a complex NP a dependent element is labelled as the head, i.e. as nsubj; nsubj is attached to a wrong root. In SRCMF, there are three types of errors: most often, nsubj is wrongly labelled (as xcomp, root, obj, vocative, csubj, nmod) which results in the absence of any nsubj; in a few cases, an element is wrongly labelled as a nsubj, which results in a double nsubj; in another few cases, nsubj is attached to a wrong root.

VS order in GSD nsubj is wrongly labelled only once, as an xcomp while an apposition of an obl is labelled as nsubj. In SRCMF, nsubj is wrongly labelled as an obj or an obl in half cases, but also as a root or a case, or attached to a wrong root. In most cases this entails the absence of nsubj; an amod is once wrongly labelled as an nsubj, which results in a double nsubj.

SVO order in GSD there are two errors: nsubj is labelled as flat (and flat as nsubj) or nsubj is attached to a wrong root. In SRCMF, either nsubj is labelled as an obl or an obj or a root, or it is attached to a wrong root.

SOV, OVS and VO orders nsubj are all correctly parsed in GSD. In SRCMF, in SOV and OVS, most often nsubj is wrongly labelled as obj (which sometimes entails a double obj), obl, xcomp, disloc, apposition; sometimes it is attached to a wrong root; exceptionally an obj is labelled as a nsubj (leading to a double nsubj). In VO, on the contrary, the most frequent error is the wrong labelling of a category (mainly obj) as a nsubj (with cases of double nsubj).

4.2 Errors on objects

SV order in GSD, most errors consist in the wrong analysis as an obj of *se* (PRO), which is expected to be an expl:pass. In other cases, an obl or xcomp is wrongly labelled as an obj. In SRCMF, a nsubj or xcomp is wrongly labelled as an obj, or obj is attached to a wrong root.

SVO order in GSD most errors result from the analysis of obj as xcomp in existential contructions such as "*cette disparition reste une énigme"* ("this disappearance remains a puzzle" (let it be noticed that the analysis as obj is not uncontroversial, as one could have expected to find an obl instead). In a few cases obj is wrongly parsed as obl. In SRCMF errors are far more diversified. Most often, obj is wrongly analysed (obl, advmod, amod, advcl, root, nsubj, hence a double nsubj in an unexpected order SVS) ; in a few cases, obj is wrongly attached.

SOV order in GSD, obj can only be a pronoun. Most errors result from the analysis of reflexive *se* as an expl: pass instead of an obj, both analyses being actually acceptable. In SRCMF, we find nominal objects (albeit rarely: "*Li rois Tristran menace"*). In some cases, obj (nominal or pronominal) is wrongly parsed (obl or flat), or a category is wrongly parsed as an obj, in addition to the correct obj (hence a double obj).

V, VS, OVS and VO orders obj are all correctly parsed in GSD. In SRCMF errors in V and VS orders necessarily involve a category being wrongly parsed as an obj. In V order, in most cases, an obl is wrongly analysed as an obj in existential constructions (where the SRCMF scheme expects obl). In VS, most often nsubj is wrongly parsed as an obj. In VO, most often obj is wrongly analysed (nsubj, obl, nmod, flat, ccomp). It is rarely wrongly attached and the root is correct in most cases. The same holds true for OVS (obj wrongly parsed as obl, advmod, iobj, root, nsubj, hence a double nsubj), though we also find one nsubj parsed as obj, hence a double obj.

Finally, to summarise these analyses, we can note a few main trends:

- both treebanks display both types of errors for nsubj and obj: the absence of a correct label (recall) and/or the presence of a wrong label (precision);
- in GSD a wrong label is never correlated to a wrong part-of-speech, whereas this is the case in 10 % of cases (16/169) in SRCMF;
- not only are the scores better in GSD than in SRCMF, but the errors are usually of a different nature, and less damaging: wrong parsing of obj is always at the benefit of a close category (obl, xcomp,

expl: pass), and this also holds true for nsubj (flat, appos, amod, expl:subj). Wrong attachments or wrong roots are exceptional[3]. On the contrary, in SRCMF, wrong roots and wrong attachments are not an exception[4], and wrong parsing of nsubj and/or obj often results in distant categories, with even possible inversions between nsubj and obj.

5 Influence of word order frequencies for parsing SRCMF

We now turn more specifically to SRCMF, in order to highlight a few correlations between frequencies of word orders and performances of the parsing. Table 5 reports the error rates for the nsubj, obj and root relations in all eleven attested combinations.

	Prevalence		**Error rates** (%)				
Order	sentences	%	root	nsubj	obj	core	any
V	472	25.02	2.75	2.12	5.72	9.32	40.04
SV	424	22.48	2.36	4.25	1.65	6.60	36.79
VS	173	9.17	2.89	6.94	2.89	8.67	36.42
SVO	119	6.30	4.20	4.20	16.81	16.81	51.26
SOV	109	5.77	0.92	5.50	5.50	7.34	42.20
OVS	71	3.76	1.41	14.08	15.59	19.72	53.52
OSV	17	0.90	0.00	5.88	0.00	5.88	47.06
VSO	23	1.21	0.00	0.00	30.43	30.43	69.57
VOS	7	0.37	14.29	14.29	0.00	28.57	71.43
VO	250	13.25	1.20	4.80	8.00	10.40	42.00
OV	221	11.71	0.00	2.71	6.79	6.79	44.80
Total	1888		2.12	4.34	6.30	9.54	41.74

Table 5: Comparison between the error rates for the root, nsubj and obj relations in SRCMF in all the occurring word orders. The "core" column is the ratio of trees where are at least one relation among root, nsubj and obj has an error.

From table 5, it appears that there is no significant correlation between word order frequency and parser performances: the five most frequent orders (V > SV > VO > OV > VS) respectively rank as 6, 3, 7, 2 and 5 in terms of total error rate. On the contrary SOV and OSV, ranked respectively 7 and 10 in terms of frequency, are respectively in position 4 and 1 position in terms correctness.

From a qualitative point of view, we may account for the high performance for SOV by the fact that obj is most often a pronoun (*le*, *la*, *les*, …), with an unambiguous function (versus an NP, which can be obj or nsubj), which probably reduces options and hence errors. For OSV, in most cases (14/17), either nsubj or obj (or both) is a unambiguous pronoun. It should also be noticed that the verb is always a complex one, hence a structure such as: obj aux nsubj verb.

Furthermore, orders with both nsubj and obj are the least frequent ones, and, globally, those show the worst parsing performances, which can be accounted for by the higher complexity of the tree. On the contrary, the high score of wrong parsing in VO is unexpected as i) VO is not infrequent (13.25 %, 3rd position) and ii) obj is often wrongly parsed as an nsubj, with results in a VS order, less frequent (9.17 %). Globally speaking, the highest rate of wrong parsing concerns obj (6.3 %), followed by nsubj (4.34 %) then by root (2.12 %). However this hierarchy varies according to word orders, at least as concerns nsubj and obj, since root always displays the lowest rate of errors (except in V order). Getting back to linguistic concerns, we observe that 3 out of the 4 very rare (less than 5 %) word orders are very badly parsed, with error rates over 20 %: OVS, VSO and VOS, which is of course more damaging for subsequent linguistic studies than when it happens with more frequent orders.

[3]Wrong parsing of roots in general in the seven orders amounts to 0.54 %.

[4]Wrong parsing of roots in general in the seven orders amounts to 2.35 %.

6 Conclusion

Our results suggest that parsing Old French is indeed harder than contemporary French, at least in the current state of existing treebanks (in terms of amount of text and heterogeneity). This is manifested both in word-level metrics and in terms of exact match for major constituents. In term of exact match for all words, our results favour Old French, but this metric is very likely to be influenced by the significantly smaller sentence lengths. Qualitative analyses concur with this trend: parsing errors seem less severe in general in GSD. Finally, rather surprisingly, while errors are not homogeneous among word orders, the most common word orders are not necessarily those that are best dealt with, although the least common word orders are those where there are the most errors.

These are provisional conclusions which deserve further investigations, especially in order to refine the correlations between word orders and wrong parsing, be it as regards the types of errors or the factors likely to be of influence. Orthogonally to these considerations, a broad study of the impact of treebank sizes and sentence lengths on parsers' behaviours could also be a useful complement of this work.

We reserve for future work transverse analyses with other facets such as time period and genre that we have abstracted over in this work. Going forward, being able to narrow down the sources of errors could help design parsers with better handling of rare phenomena, which would be crucial to support fine-grained quantitative linguistic analyses.

References

Jean-Baptiste Camps, Elena Albarran, and Alice Cochet. 2019. Geste: un corpus de chansons de geste, 2016-…, April.

Marie Candito and Djamé Seddah. 2012. Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2, Grenoble, France, June. Association pour le Traitement Automatique des Langues.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical French dependency parsing: treebank conversion and first results. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 1840–1847, Valletta, Malta, May. European Language Resources Association.

Maximin Coavoux and Benoît Crabbé. 2017. Multilingual Lexicalized Constituency Parsing with Word-Level Auxiliary Tasks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 331–336, València, España, April. 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, June. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia, July. Association for Computational Linguistics.

Martin-Dietrich Glessgen, Dumitru Kihaï, and Paul Videsott. 2010. L'élaboration philologique et linguistique des Plus anciens documents linguistiques de la France. *Bibliothèque de l'École des chartes*, 168(1):5–5.

Loïc Grobol and Benoît Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés. In *28e Conférence sur le Traitement Automatique des Langues Naturelles*, Lille, France, June. Association pour le Traitement Automatique des Langues.

Gaël Guibon, Isabelle Tellier, Mathieu Constant, Sophie Prévost, and Kim Gerdes. 2014. Parsing Poorly Standardized Language Dependency on Old French. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories*, pages 51–61, Tübingen, Deutschland, December.

Gaël Guibon, Isabelle Tellier, Sophie Prévost, Mathieu Constant, and Kim Gerdes. 2015. Searching for Discriminative Metadata of Heterogenous Corpora. In Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories*, pages 72–82, Warszawa, Polska, December.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Traitement Automatique des Langues*, 60(2):71.

Céline Guillot, Serge Heiden, and Alexei Lavrentiev. 2018. Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques. Revue de Linguistique française diachronique*, (7):168.

Pierre Kunstmann and Achim Stein. 2008. Le Nouveau Corpus d'Amsterdam. *Corpus*, (7), November.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

France Martineau. 2008. Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, (7), November.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online, November. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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, May. European Language Resources Association.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv:1706.05098 [cs, stat]*, June.

Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2018. An Investigation of the Interactions Between Pre-Trained Word Embeddings, Character Models and POS Tags in Dependency Parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2711–2720, Bruxelles, Belgique, October. Association for Computational Linguistics.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts: the Syntactic Reference Corpus of Medieval French (SRCMF). In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpora*, Corpus Linguistics and International Perspectives on Language, pages 275–282, Manchester, UK, September. Gunter Narr Verlag.

Achim Stein. 2014. Parsing Heterogeneous Corpora with a Rich Dependency Grammar. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2879–2886, Reykjavík, Ísland, May. European Language Resources Association.

Achim Stein. 2016. Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 707–713, Portorož, Slovenija, May. European Language Resources Association.

Milan Straka, Jana Straková, and Jan Hajič. 2019. Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. *arXiv:1908.07448 [cs]*, August.

David Trotter. 2012. Bytes, Words, Texts: The Anglo-Norman Dictionary and its Text-Base. *Digital Medievalist*, 7(0), February.

Rob van der Goot. 2021. We Need to Talk About train-dev-test Splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Pieter Van Reenen, Evert Wattel, and Margôt van Mulken. 2006. *Chartes de Champagne en français conservées aux Archives de l'Aube, 1270-1300*. Éditions Paradigme.

David Wrisley. 2018. The Open Medieval French Initiative (OpenMedFr).

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, October. Association for Computational Linguistics.

Annotation guidelines of UD and SUD treebanks for spoken corpora: a proposal

Sylvain Kahane*, Bernard Caron, Emmett Strickland*, Kim Gerdes*****
*Modyco, Université Paris Nanterre & CNRS
**Llacan, CNRS & INALCO
***Lisn, Université Paris Saclay & CNRS

Abstract

This paper presents practical and theoretical guidelines for the development of treebanks for spoken languages in the UD and SUD annotation schemes. We discuss text-sound alignment, segmentation into "sentences", use of "punctuation", paradigmatic lists, disfluencies, and paratactic constructions. This proposal is based on the development of (Surface-Syntactic) Universal Dependencies treebanks for spoken French, Naija, and Beja.

1. Introduction

This paper presents our recommendations for the development of treebanks for spoken languages, based on our experience in the development of several treebanks for spoken French (Gerdes & Kahane 2009, Lacheret et al. 2014, 2019), Cantonese (Wong et al. 2017), and more recently a large treebank for spoken Naija, an English-lexifier pidgincreole from Nigeria which is spoken by more than 100 million people (Caron et al. 2019), as well as a small treebank for Beja, an Afro-Asiatic language spoken in Sudan (Kahane et al. 2021). All of these treebanks are dependency-based, and our more recent treebanks use the SUD (Surface-Syntactic Universal Dependencies) framework (Gerdes et al. 2018, 2019, 2021), which can be automatically converted into UD (Universal Dependencies) (de Marneffe et al. 2021). Previous attempts to provide guidelines for UD-based spoken treebanks include Dobrovoljc & Nivre (2016) and Øvrelid et al. (2018).

The need for special guidelines for spoken language treebanks arises from several particularities of spoken language corpora and spoken language grammar.

Spoken corpora are transcriptions of audio and video recordings. We will not discuss the issue of these transcriptions, as conventions can vary from language to language.[1] However, the relationship between the written text and the sound file which characterizes oral corpora introduces some specificities in the treebank itself, which are discussed in Section 2. The identification of speakers and the marking of overlaps between speakers are also considered.

The next challenge in the transcription of spoken languages is the fact that recorded speech does not contain any consistent sentence boundaries. A segmentation based on prosody is sometimes adopted, especially for interlinear glossed texts (IGTs), but prosodic units can be strongly divergent from syntactic units (Beliao et al. 2015, Kahane & Lacheret 2019). Section 3 is devoted to segmenting

[1] Each language we considered is bound by different constraints. French has a strong tradition of orthographic normalization (even if there remains some room for discussion, see Dister et al. 2019). Naija orthography has yet to be codified, even if the language now enjoys some official written media presence through outlets like BBC News Pidgin (https://www.bbc.com/pidgin). We therefore gave our transcriptors the freedom to choose how best to represent the language in writing. Beja has no written tradition whatsoever, which means that our Beja corpus is an interlinear glossed text with a phonetic transcription and a word segmentation done by a linguist (Vanhove 2014).

spoken texts into maximal syntactic units. Following the French tradition of studies of the syntax of spoken French (Blanche-Benveniste 1990, Berrendoner 1990), we consider two levels of analysis, referred to as *microsyntax* and *macrosyntax* in the French scholarly tradition: microsyntax, which is the syntax of government (Fr. *rection*), describes marked relations where one word imposes its form, category, or position on another; macrosyntax concerns looser relations, such as those involving detached/dislocated units, parentheses, inserts, discourse markers, etc. Macrosyntactic units generally correspond to units that are delimited by punctuation in written texts. The choice of the delimiters also requires some discussion.

One particularity of spoken productions is the great number of paradigmatic lists or "piles", which are successions of units piled up in the same syntactic position. This is the case with coordination, but also reformulation, or apposition. Section 4 proposes a homogeneous treatment of these constructions.

Section 5 is devoted to the analysis of disfluencies due to incomplete units.

The last characteristic of spoken languages we examine is the variety of paratactic constructions used in spoken production. In Section 6 we propose splitting the *parataxis* relation of UD into seven different subtypes.

2. Text-sound alignment and speech turns

UD-based treebanks are encoded using the CoNLL-U format (https://universaldependencies.org/format.html). In this tabular format, each sentence is encoded separately. Each token occupies a row and information associated with that token is divided into 10 columns (see Fig. 1).[2] Metadata associated with a given sentence precede the table describing the dependency tree. These metadata must contain the text of the sentence (# text) and a sentence id (# sent_id). For spoken corpora, two main pieces of information should be added: a (permanent) link to the sound file (# sound_url) and, if the corpus is not a monologue, an id for each speaker (# speaker_id).

```
#sound_url = http://www.tal.univ-paris3.fr/trameur/iTrameur-naija/mp3/KAD_22_Chatting-At-The-Restaurant_DG.mp3
#speaker_id = Sp205
#text = na wa for dat woman o !//
#text_en = It's a shame for that woman!
#text_ortho = Na wa for dat woman o!
1   na  na  PART    _   PartType=Cop   0   root    _   AlignBegin=15240|AlignEnd=15382|ExtPos=INTJ|Gloss=be1|Idiom=Yes
2   wa  wa  INTJ    _   PartType=Cop   1   comp:pred  _   AlignBegin=15382|AlignEnd=15523|Gloss=wow|InIdiom=Yes
3   for for ADP _   _   1   comp:obl    _   AlignBegin=15523|AlignEnd=15665|Gloss=for
4   dat dat DET _   Number=Sing|PronType=Dem  5   det _   AlignBegin=15665|AlignEnd=15807|Gloss=SG.DEM
5   woman   woman   NOUN    _   _   3   comp:obj    _   AlignBegin=15807|AlignEnd=15948|Gloss=woman
6   o   o   PART    _   PartType=Disc  1   mod:emph    _   AlignBegin=15948|AlignEnd=16090|Gloss=EMPH
7   !// !// PUNCT  _   _   1   punct   _   AlignBegin=16090|AlignEnd=16090|Gloss=PUNCT
```

Figure 1. Encoding of text-sound alignment (from SUD_Naija-NSC)

With the sound url and an additional temporal alignment, it is possible to listen to the sound associated with each word. In (S)UD_Naija-NSC and in (S)UD_Beja-NSC, each word is time-aligned using the features AlignBegin and AlignEnd, with a value in milliseconds from the start of the sound file.[3] If the corpus features punctuation marks, the two features should typically be equal in value unless the symbol corresponds to a prosodic break. These word-level features also allow one to determine the timespan of a given utterance: the AlignBegin value of the first token correspond to the beginning of

[2] SUD_Naija-NSC uses a macrosyntactic markup (see Section 3) with special punctuation signs, such as < or //. This markup is part of the annotated text, given in #text. An orthographic variant is proposed in #text_ortho, with traditional punctuation signs and an uppercase at the beginning of the sentence. Moreover an English translation is given in # text_en, as well as a gloss for each token.

[3] In the case of SUD_Naija-NSC, only some time boundaries were stored during the transcription process; others have been roughly assessed by dividing each time segment by the number of words. It would have been also possible to keep the features on some words only.

the utterance, while the AlignEnd value of the final token corresponds to the end. For corpora with only a sentence-level alignment, we recommend adding a single AlignBegin feature to the first token, and an AlignEnd feature to the last.

For the time being, (S)UD_Naija-NSC, (S)UD_Beja-NSC, and (S)UD_French-ParisStories are the only (S)UD treebanks that provide a link towards a sound file. UD_English-GUM (Zeldes 2017), a third of which represents spoken data, contains # speaker and # addressee features, which are, to our knowledge, the only way to distinguish spoken from written utterances. UD_Polish-LFG (Patejuk & Przepiórkowski 2018) contains a feature # genre = spoken (prepared) to distinguish the spoken data. UD_Frisian_Dutch-Fame (Braggaar & Vander Goot 2021) and UD_Scottish_Gaelic-ARCOSG (Batchelor 2019) have a # speaker feature, while UD_Norwegian-NynorskLIA (Øvrelid et al. 2018) has metadata including compound features like # dialect: eidsberg speakerid: eidsberg_uio_03, which include both a location and a speaker id. UD_Latvian-LVTB (Pretkalniņa et al. 2018) has a feature # newpar id, which probably contains the speaker id of new participant. Other treebanks of spoken languages (UD_Turkish_German-SAGT, UD_Slovenian-SST, UD_Cantonese-HK, UD_Komi_Zyrian-IKPD, UD_Swedish_Sign_Language-SSLC, UD_Chukchi-HSE)[4] do not contain any specific metadata, while some treebanks (UD_English-Lines, UD_Greek-GDT, UD_Persian-Seraji) that are partially composed of spoken data do not contain any metadata allowing one to distinguish between speech and writing. For mixed corpora, we think that a feature # genre = spoken should be used to facilitate the identification of speech. Outside of (S)UD, the CHAT transcription system used in the CHILDES database of childhood speech marks each utterance with a three-letter speaker identifier, each of which is associated with a name and role (i.e., father, mother, child) in the file header. Additional information may be provided about a speaker's background (age, socioeconomic status) and the context of the recording, while special punctuation can be used to designate overlapping speech (MacWhinney 2000).

Note that in (S)UD corpora only one speaker and one # speaker_id is allowed for each utterance. In instances of co-constructions between two speakers, we use the special features AttachTo and Rel. By co-constructions, we mean two trees T1 and T2 from two different speakers that form a cohesive syntactic construction, as in (1) from SUD_French-Rhapsodie.[5]

(1) *$L3 ^et c'est récupéré (bien sûr) par l'équipe $- argentine //+*
$L2 qui -$ sont de nouveau en possession (les Argentins) du "euh" ballon //
'$L3 ^and it is recovered (of course) by the $- Argentinian team //+.
$L2 who -$ are again in possession (the Argentines) of the "uh" ball //'

One token from T1 is the governor of the root of T2 in this new construction. In (1), L2's utterance is a relative clause that modifies the noun *équipe* 'team' from L3's utterance. This is indicated on the root of the second tree by the feature AttachTo=11@Rhap_D2003-92bis indicating that this utterance could be attached to the 11th token of tree Rhap_D2003-92bis, and by the feature Rel=mod@relcl indicating the relation between the two tokens (see Fig. 2 and 3).[6]

[4] UD-Chukchi-HSE (Tyers & Mishchenkova 2020) has a feature # text[phon], with a phonetic transcription, but it is not a feature specific to spoken corpora.

[5] SUD_French-Rhapsodie is the former SUD_French-Spoken, which was initially the Rhapsodie treebank (Lacheret et al. 2019). The treebank has been renamed due to the introduction of a new treebank of spoken French, SUD_French-ParisStories. Rhapsodie uses a markup, where sentence boundaries are indicated by // and co-constructions by //+. Overlaps are indicated by $- …-$. But contrary to SUD_Naija-NSC, # text is based on standard punctuations and the macrosyntacic markup is stored in # macrosyntax.

[6] UD requires the root of a tree to have the relation root.

```
# macrosyntax = ^et c'est récupéré ( bien sûr ) par l'équipe $- argentine //+
# text_ortho = et c'est récupéré, bien sûr, par l'équipe, euh, argentine.
# speaker = L3
# sent_id = Rhap_D2003-92bis
1  et  et  CCONJ  _  _  3  cc  _  _
2  c'  ce  PRON  _  Gender=Masc|Number=Sing|Person=3|PronType=Dem  3  subj@pass  _  _
3  est être  AUX _  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin  0  root  _  _
4  récupéré  récupérer  VERB  _  Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part  3  comp:aux@pass  _  _
5  ,  ,  PUNCT  _  _  6  punct  _  _
6  bien  bien  ADV _  _  7  mod _  ExtPos=ADV|InIdiom=Yes
7  sûr sûr ADJ _  Gender=Masc|Number=Sing  3  mod _  PhraseType=Idiom
8  ,  ,  PUNCT  _  _  9  punct  _  _
9  par par ADP _  _  4  comp:obl  _  _
10 l'  le  DET _  Definite=Def|Number=Sing|PronType=Art  11  det _  _
11 équipe  équipe  NOUN  _  Gender=Fem|Number=Sing  9  comp:obj  _  _
12 ,  ,  PUNCT  _  _  13  punct  _  _
13 euh euh INTJ  _  _  11  discourse  _  _
14 ,  ,  PUNCT  _  _  15  punct  _  _
15 argentine  argentin  ADJ _  Gender=Fem|Number=Sing  11  mod _  Overlap=Rhap_D2003-92ter
16 .  .  PUNCT  _  _  3  punct  _  _

# macrosyntax = qui -$ sont de nouveau en possession ( les Argentins ) du "euh" ballon //
# text_ortho = qui sont de nouveau en possession du, euh, ballon.
# speaker = L2
# sent_id = Rhap_D2003-92ter
1  qui qui PRON  _  _  2  subj  _  Overlap=Rhap_D2003-92bis
2  sont  être  VERB  _  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  0  root  _
AttachTo=11@Rhap_D2003-92bis|Rel=mod@relcl
3  de  de  ADP _  _  5  mod _  ExtPos=ADV|PhraseType=Idiom
4  nouveau nouveau ADJ _  Gender=Masc|Number=Sing  3  unk _  InIdiom=Yes
5  en  en  ADP _  _  2  comp:obl  _  _
6  possession possession NOUN  _  Gender=Fem|Number=Sing  5  comp:obj  _  _
7  les le  DET _  Definite=Def|Number=Plur|PronType=Art  8  det _  _
8  Argentins  Argentin  PROPN  _  _  2  dislocated _  _
9-10  du  _  _  _  _  _  _  _  _
11 de  de  ADP _  _  6  udep  _  _
12 le  le  DET _  Definite=Def|Gender=Masc|Number=Sing|PronType=Art  16  det _  _
13 ,  ,  PUNCT  _  _  12  punct  _  _
14 euh euh INTJ  _  _  16  discourse  _  _
15 ,  ,  PUNCT  _  _  14  punct  _  _
16 ballon  ballon  NOUN  _  Gender=Masc|Number=Sing  11  comp:obj  _  _
17 .  .  PUNCT  _  _  2  punct  _  _
```

Figure 2. Encoding of a co-construction (from SUD_French-Rhapsodie)

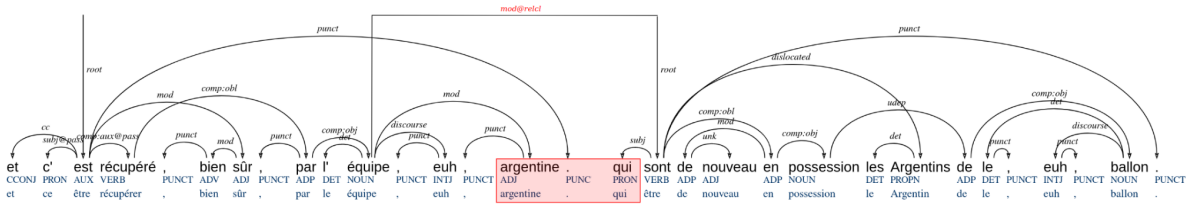


Figure 3. Visualization of the encoded information. In red: the AttachTo link and the Overlap.

In cases of overlap between two speech turns, words that overlap have a feature Overlap whose value is the id of the overlapping sentence. In our previous example, *argentine*, of sentence Rhap_D2003-92bis, has a feature Overlap=Rhap_D2003-92ter, indicating that it overlaps with sentence Rhap_D2003-92ter, more specifically with the word *qui* 'who', which itself carries the feature Overlap=Rhap_D2003-92bis. Overlaps can also be deduced from time alignment, but some corpora indicate overlaps without also containing temporal information, and for most query engines it is much simpler to request overlaps if a feature is present.

3. Sentence segmentation and punctuation

With the exception of rehearsed speeches and other kinds of highly prepared oratory, oral language cannot typically be segmented into units equivalent to the written sentence. Instead, we segment our corpora into illocutionary units (IUs) (Cresti 1995, Pietrandrea et al. 2014). An IU is defined as a speech

segment that corresponds to a single speech act, i.e., a single question, declaration, or command. Consider the following examples from French and Naija.

(2) *ceux qui sont en location < la moyenne < c'est environ trois ans //*
those who are on lease < the average < it is about three years //
'For those that rent, the average lease is about three years.'

(3) *e get one lady < hm she just enter inside shop o //*
there was one lady < hm she just came into the shop //

These utterances represent cohesive linguistic units serving to express a single primary idea. One can split these into internal units sharing no marked relations of dependency with one another, separated in these examples by the character <. The cohesiveness of these units can nevertheless be demonstrated by their varying degrees of autonomy. For example, the utterances *c'est environ trois ans* 'it is about three years' and *hm she just enter inside shop o* 'hm she just came into the shop' form perfectly acceptable utterances in isolation that carry the same basic message as the original utterance. However, *la moyenne* 'the average' or *e get one lady* 'there was one lady' would either be unacceptable or serve a very different illocutionary purpose in isolation. In oral corpora, we recommend preserving this cohesiveness by using the IU as the primary unit of segmentation. Within each IU, internal components can be delimited through special symbols, either placed directly in the list of tokens, or in a dedicated sentence feature containing the macrosyntactic annotation (and replaced in the text by standard punctuation signs; see Note 5).

In our corpora, the end of each IU is marked by a symbol indicating an IU boundary. In the corresponding CoNLL-U file, these may be integrated directly into the list of tokens and annotated as punctuation marks. In order to better identify the modality of each IU, we recommend introducing a separate boundary marker for each type of utterance. In our corpus of spoken Naija, the symbol // is used to mark the end of an assertion, ?// the end of a question, and !// the end of an order or an exclamation. The symbol &// is used to mark the end of an interrupted or incomplete utterance. However, more traditional symbols such as periods and question marks may also be used so long as they are used consistently as IU boundary markers.

(4) *I no pass all di subjects //*
'I didn't pass all of the subjects.'

(5) *wetin be Ponzi Scheme ?//*
'what is a Ponzi Scheme?'

(6) listen attentively !//
'listen attentively!'

(7) *e get things wey you &//*
'there are things that you…'

As demonstrated previously in examples (2) and (3), IUs can contain internal units with varying degrees of autonomy. All completed utterances must contain at least one (and typically only one) nucleus, a fully autonomous macrosyntactic unit capable of forming an acceptable utterance when spoken alone in the same discursive position (Pietrandrea & Kahane 2019). Examples (4-6) are each composed of one nucleus, but IUs can also contain multiple adnuclei located to the left or right peripheries of the nucleus. These are illocutionarily dependent on the nucleus and typically lack any marked microsyntactic relationship with the elements of the nucleus. These can be divided into prenuclei, located to the left of the nucleus, and postnuclei, located to the right. Like the IU boundaries, we recommend explicitly delimiting these units using a dedicated set of symbols. In our corpora, we favor using < to mark the ends of prenuclei (see ex. 8-9) and > to mark the beginning of postnuclei (ex. 10-11). However, more traditional symbols such as commas may also be used.

(8) *les chaises < il faut me les donner //*
'the chairs, you need to give them to me.'

(9) *en ce qui me concerne < j'aimerais enseigner dans un établissement public //*
'as for me < I would like to teach in a public school'

(10) *dat's what de use to do > some of dem o //*
'that's what they used to do, some of them.'

(11) *good evening > my daughter //*
'good evening, my daughter.'

Note that it is possible for an IU to contain strings of adnuclei

(12) *donc < alors < ça date de quand à peu près > ce fauteuil-là ?//*
'so, then, it dates from when approximately, this armchair?'

(13) *Osas < dis your wedding wey you dey prepare < me < I dey look forward to di ting o !//*
'Osas, that wedding of yours that you are preparing, me, I'm looking forward to it!'

These adnuclei are typically connected to the root of the utterance using the UD relations *dislocated, vocative, discourse,* and in some cases *parataxis*. These are discussed in Section 6.[7]

It is important to note that our segmentation generally avoids coordination between main verbs in a single IU, since each part can form an autonomous utterance. (13) shows a string of consecutive IUs each containing a single main verb.

(14) *vous oubliez vos produits habituels.*
et vous mettez à l'intérieur la boule magique, la boule de lavage avec vos vêtements.
et vous allez voir le résultat.
'you forget your usual products.
and you put inside the magic ball, the washing ball with your clothes.
and you will see the result.'

4. Paradigmatic lists

The annotation of paradigmatic structures is challenging in any analysis that marks heads, such as dependency- and X-bar-based phrase structure, because the idea of paradigm itself supposes that two or more elements jointly qualify as the head of a phrase. SUD and UD use the same set of basic relations to encode paradigmatic structures: The *conj* and *cc* relations for standard coordination, the *list* relation that is absent from the analysis of spoken data that we propose,[8] and finally the *orphan* and *reparandum* relations used respectively for ellipsis and disfluency.

As shown in Blanche-Benveniste et al. (1990) and Gerdes & Kahane (2009), reformulation is hard to delimit on a spectrum from coordination to actual repairs, where a second phrase replaces the first. Are (15a) and (15b) coordinations or reformulations?

(15) a. *et puis après, ben, j'ai travaillé sur* ***les micro-processeurs***, ***l'informatique***. (Rhap_D0005-9)
'and then, well, I worked on **microprocessors, computers**.'
b. *mais comme toujours, l'acte d'écrire peut prendre* ***différents masques, différentes valeurs***. (Rhap_2009-2)
'but as always, the act of writing can take on **different masks, different values**.'

[7] Fillers uttered by the listener, such as Fr. *mh* 'hum', are considered discourse markers attaching to the speaker's utterance. They have their own tree but carry the features AttachTo and Rel=discourse.

[8] *list* encodes a list containing only fragments such as address information or phone numbers.

Reformulations are not necessarily repairs but rather elaborations that are also frequently used by rhetorically skilled speakers.

(16) a. *from dat day < as I go meet am for shop < im con { **dey observe me** || **dey observe my attitude** } //* (ABJ_GWA_12_Accident_MG__123)
'From that day, as I went to meet him in the shop, he started **observing me, observing my attitude**.'
b. *mais le, le dis~, **le commissaire du district, le préfet du lieu** ne voulait pas de, de femme non mariée.* (Rhap_D2004-11)
'but the, the dis~, **the district commissioner, the prefect of the place** did not want any, unmarried woman.'
c. *euh, **deux petites phrases, deux vraies options** qui dessinent votre route, une route qui témoigne **d'une certaine, d'une bonne, d'une très bonne** conduite*. (Rhap-D2001-3)
'uh, **two small sentences, two real options** that draw your road, a road that testifies to **a certain, a good, a very good** conduct.'

We excluded the *reparandum* relation from the analysis of spoken data in our SUD annotation scheme, and instead use the concept of paradigmatic lists developed by Blanche-Benveniste et al. (1990) and produce a simple typology of paradigmatic lists (Kahane & Pietrandrea 2012) by annotating three sub-relations of the *conj* dependency:

- *conj:coord*: the standard coordination, where each conjunct denotes a different referent;
- *conj:dicto*: used for several denotations of the same referent, which are repetitions or reformulations of the same denotation, as in (16), including disfluencies as in (16b);
- *conj:appos:* used for double denotations of the same referent, as in (17).[9]

We decided not to distinguish reformulation and disfluencies because the line between voluntary and unvoluntary reformulation or correction is very difficult to draw.[10] In the transfer to UD annotation, *conj:dicto* is renamed *reparandum*, even if it is not exactly a standard *reparandum* annotation, because the first conjunct remains the head.[11]

(17) a. ***match nul, zéro partout** pour Lille au Mans.* (Rhap_M2006-72)
'**draw, zero all** for Lille at Le Mans.'
b. *et c'est dans cet esprit qu'est proposée par Szymanoski **l'ouverture du concert, son opus douze**.* (Rhap_D2012-47)
'and it is in this spirit that Szymanoski proposes the opening of the concert, his opus twelve.'
c. *eh bien, le secret, il est **là, dans ce petit bruit que l'on entend***. (Rhap_D2011-57)
'well, the secret is **there, in this small noise that we hear**.'

In the case of a partial answer to a *wh*-question, we also consider that the answer forms a *conj:appos* pile with the wh-word, encoded using the AttachTo and Rel features.

[9] UD uses the same *appos* relation for appositions that form paradigmatic lists and appositions where one unit modifies the other. We propose encoding the latter as a standard modification (*mod* in SUD or *nmod* in UD) or as subtype of modification (*mod:appos* in SUD or *nmod:appos* in UD) since the two constructions have different semantic and prosodic properties: in *conj:appos*, the two conjuncts are separate prosodic units, the apposed unit being in a different register, while in *mod:appos*, the two elements form one cohesive prosodic unit (*my brother John, le journal Le Monde*).

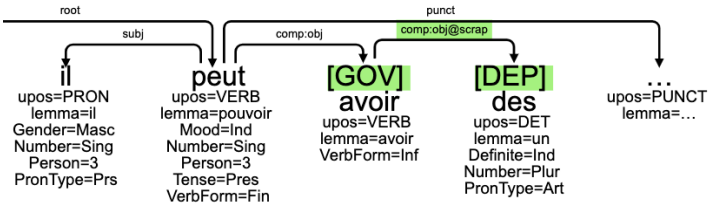
[10] The initial Rhapsodie annotation considered seven different cases of paradigmatic lists (Kahane & Pietrandrea 2020). The distinction between disfluencies and reformulation was based on purely formal criteria (repetition of a unit) to avoid poor inter-annotator agreement (Bawden et al. 2014).

[11] The notion of *reparandum* presupposes that the second conjunct repairs the first (Shriberg 1994). Following Blanche-Benveniste (1990), we instead consider that, in reformulations, even involuntary ones, information is cumulative, and nothing that has been said can be completely forgotten.

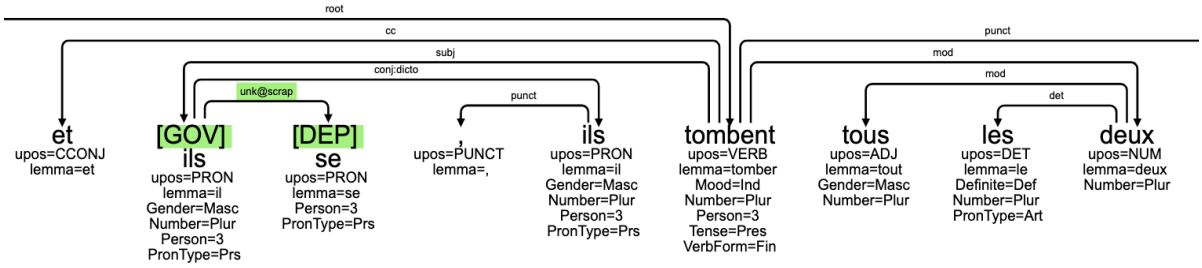
(18) $*L1 Magalie utilise* ***depuis combien de temps*** *notre boule magique?*
$L5 ***depuis deux mois***.
'$L1 **how long** has Magalie been using our magic eight ball?
$L5 **for two months**.'

5. Disfluencies

One kind of disfluency, often called repair, is observed when a unit, then called the reparandum, gets overridden by a new unit, the repair. As shown in section 4, this is a subtype of the common mechanism of listing, which is not necessarily disfluent and which we formalize with the relation *conj:dicto*. Another case of disfluency corresponds to incomplete units. In this case, UD chooses to promote one element as the head of the unit and to attach the other elements to it (Dobrovoljc & Nivre 2016). When this produces nonstandard relations, we propose adding a feature. In SUD, we add the feature *scrap* on these relations using @ as a separator (*@scrap*) This is converted into UD by inserting the Scrap=Yes feature on the dependent. Fig. 4 and 5 show some uses of *@scrap*. This feature is particularly useful for error mining: for instance, a relation between a verb and a determiner, as in Fig. 4, should not be allowed without a *@scrap*.



'it can have some …'
Figure 4. Incomplete object (comp:obj@scrap).

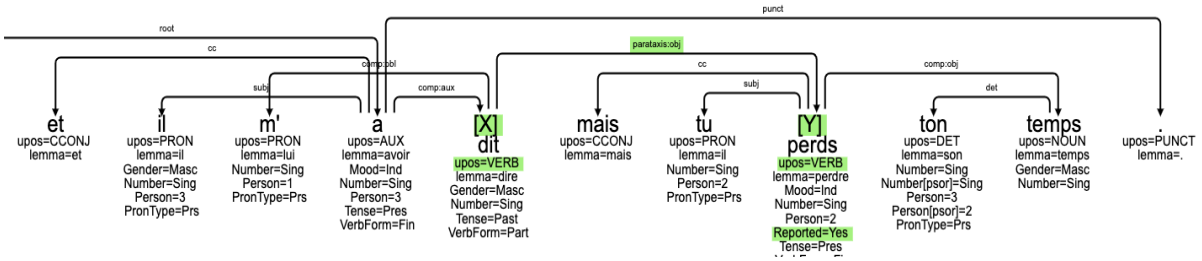


'and they REFL, they both fall'
Figure 5. Unknown relation inside a reparandum (unk@scrap).

6. Paratactic constructions

Paratactic constructions are particularly frequent in spoken corpora. The UD *parataxis* relation covers all cases where a finite verbal construction is neither the root of an utterance, nor a governed clause. In our corpora, we have identified seven different situations that we propose to distinguish: *parataxis:obj, parataxis:discourse, parataxis:parenth, parataxis:insert, parataxis:dislocated, parataxis:mod, parataxis:conj*.

The first occurs when an illocutionary unit occupies a governed position. The most common situation is reported speech in the object position of a verb of saying. In such cases, UD uses the relation *parataxis*, which we propose replacing with *parataxis:obj*.



'and he told me but you are wasting your time.'

Figure 7. Reported speech.

We add the feature Reported=Yes on the head of the reported IU, because some of them are not directly governed, such as the second reported IU in (19), *ça pourrait être grave*, encoded as a separate sentence.[12]

(19) *après, elle a dit, on est aux affaires. ça pourrait être grave.*
'then she said, we're in business. it could be serious.'

Thanks to the feature Reported=Yes, a standard *comp:obj* relation can be used in SUD and should be preferred because reported speech can commute with a complementizer phrase and, in accordance with SUD principles, two units that are in the same commutation paradigm must have the same syntactic function.

Another case of an IU in a governed position is called a graft by Deulofeu (1999). A graft is an IU that is produced in a position where a noun phrase is expected, such as the IU *disons ma carrière pour simplifier* 'let's say my career to simplify', which occupies a subject position.

(20) *vous avez dit que, euh,* ***disons ma carrière pour simplifier,*** *témoigne de ma bonne conduite.*
'you said that, uh, **let's say my career to simplify**, shows my good behavior.'

In such a case we use a plain *subj* relation rather than a *parataxis* relation, but we add a feature Graft=Yes. Some graft constructions are lexicalized such as the construction (21) from SUD_French-ParisStories, where an IU is introduced by the idiom *en mode* 'in the mode', or the construction (22), where a question is apposed to the word *question* 'question'.

(21) *et là lui il l'a regardée en mode* ***euh mais ça va madame ?***
'and there he looked at her in the mode **uh but is it ok miss ?**'

(22) *alors on pourrait poser la question* ***les écrivains ont-ils existé ?***
'then one could ask the question **did the writers exist?**'

UD uses the relation *discourse* for discourse markers, except for verbal expressions, such as *I mean, I guess, you know,* etc., as in (23), as well as tag questions, where *parataxis* is used.[13] This is a category-based distinction that we believe is unwarranted.[14] In SUD, we decided to extend *discourse* for these cases, which is automatically converted into *parataxis:discourse* in UD. The relation has already been introduced in the UD spoken Slovenian treebank (Dobrovoljc & Nivre 2016).

[12] Another solution could be to add quotation marks during the transcription, but this might become unreadable as the reported speech can span over a whole set of illocutionary units.

[13] As shown by Kahane & Pietrandrea (2009), verbal discourse markers have properties that distinguish them from parentheses. They do not accept tense modifications and they cannot be modified. And they have a transitive verb without an overt object, but which takes its host as its object argument.

[14] Many distinctions for relations are category-based in UD, such as *nsubj* vs *csubj*, for nominal vs clausal subjects, *obj* vs *ccomp* for nominal vs clausal objects, or *amod* vs *advmod* vs *nmod*, for adjectival vs adverbial vs nominal or adpositional modifiers.

(23) ***you know*** *gote na kind of delicacy weh dem dey prepare //*
you know gote is a kind of delicacy which they prepare.'

The third case is parenthetical clauses or parentheses, which are real illocutionary units that could be autonomous, but which are inserted in the middle of another illocutionary unit, as *my mama dey dat time* 'my mum was alive at the time' in the Naija sentence (24) or *je suis désolée* 'I'm sorry' in the French sentence (25). We note them as *parataxis:parenth*.

(24) *sometimes (**my mama dey dat time** //) she go help me carry small //*
'Sometimes, **my mum was alive at the time**, she would help me to carry a little.'

(25) *sauf que le clip, enfin, **je suis désolée**, mais il faisait pas très professionnel, hein*.
(ParisStories_2020_concoursInstagram_49)
'except that the clip, well, **I'm sorry**, but it was not very professional, eh'

The fourth situation is that of incises or inserts, which are distinguished from parenthetical ones by the fact that they are not saturated and could not form an independent statement (Bonami & Godard 2008). They belong to a more formal register (news, sermons, stories, etc.). In French, they are also characterized by the inversion of the subject. We annotate them as *parataxis:insert*. An insert can be found at the end of an illocutionary unit.

(26) *vraiment, **dit Job**, la vie de l'homme sur terre est une corvée*. (Rhap_M2003-33)
'truly, **says Job**, man's life on earth is a chore.'

(27) *alors, que va faire maintenant le président ? **se demande Le Progrès à Lyon***.
(Rhap_D2013-45)
'So what will the president do now? **wonders Le Progrès in Lyon**.'

The fifth case is that of verbal constructions in the prenucleus position where they function as dislocations. In example (28), *j'ai des copines actuellement* 'I have friends at present' is clearly not an illocutionary unit (it is not asserted by the speaker that she has friends). It is a prenucleus, which can switch with a nominal phrase like *mes copines* 'my friends'. For the time being, we use *dislocated* for these cases in SUD and UD, but we consider revisiting it as *parataxis:dislocated* in UD.

(28) ***j'ai des copines euh actuellement*** *euh, je m'entends super bien avec*.
(ParisStories_2019_experienceFac_37)
'**I have friends uh at present** uh, I get along great with.'

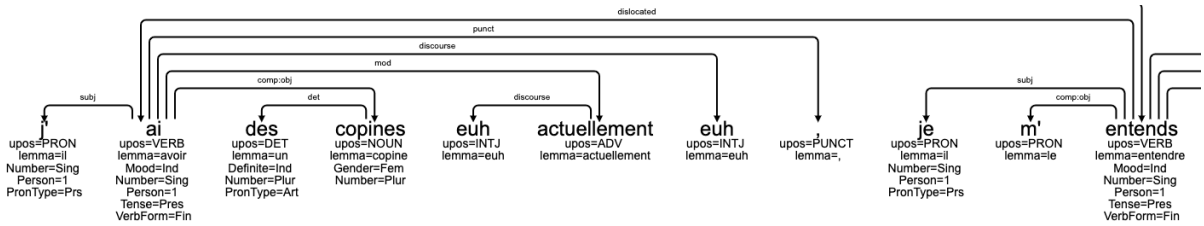


Figure 8. Parataxis:dislocated.

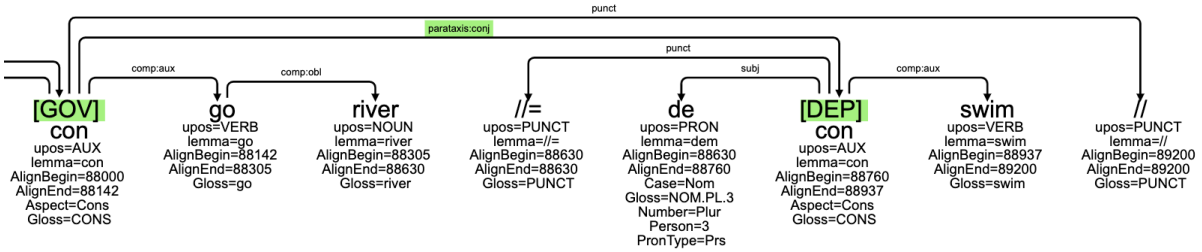
The sixth case of paratactic construction is the case of clausal modifiers that are not introduced by a subordinating conjunction. As in the previous case, the first clause *creche no dey my office*, lit. 'daycare is not (in) my office', is not asserted by the speaker and is a clear prenucleus. For such cases, we use *parataxis:mod* or simply *mod* in SUD (equivalent to UD *advcl*).

(29) *creche no dey my office now < e for no dey easy //*
'If I hadn't had a daycare in my office, it wouldn't have been easy.'

UD makes a seventh use of the *parataxis* relation, to link juxtaposed propositions, as in (19), taken from the spoken part of UD_English-GUM.

(30) *God, I didn't spend the night, that's what makes me so mad, I'm grounded for nothing.*

Our IU segmentation avoids this case most of the time; thus (30) can be perfectly split into three autonomous assertions. Nevertheless, in cases where we do not separate two juxtaposed or coordinated statements, we propose using *parataxis:conj*. In SUD_Naija-NSC, this relation has been used for sequences of parallel IUs.



'then they went to the river, and they swam.'[15]

Figure 9. Parataxis:conj.

So far, we have not found any examples of parataxis which do not fit into one of the seven cases considered, and as we have shown, quite a few can be done away with in SUD (e.g. *parataxis:obj, parataxis:discourse* and *parataxis:mod*).

7. Conclusion

The goal of this paper was to present the additional features we have introduced in our spoken treebanks and to give some proposals for future developments of spoken treebanks, in order to converge on a common set of features. This concerns both practical recommendations for text-sound alignment (# sound_url, BeginAlign, EndAlign), and theoretical propositions concerning text segmentation into sentences, paradigmatic lists, and paratactic constructions.

Our set of relations and tags are meant to be extended to the annotation of constructions typical of spoken texts in other languages as well. This can only be verified by the elaboration of more spoken treebanks for typologically different languages. We hope that the phenomena presented in this paper will motivate other linguists to work on spoken language treebanks and that this paper can serve as a first guide in this endeavour.

Acknowledgements

We would like to thank our reviewers for their valuable remarks. Several people have participated in the development of our annotation schemes for spoken language. We are particularly grateful to Marine Courtin, Vanessa Gaudray-Bouju, Menel Mahamdi, and Mariam Nakhlé.

References

Batchelor C. (2019). Universal dependencies for Scottish Gaelic: syntax. In *Proceedings of CLTW2019 at Machine Translation Summit XVII*.

Bawden R., Botalla M.-A., Gerdes K., Kahane S. (2014) Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie. In *Proceedings of the 9th Language Resources and Evaluation Conference* (*LREC*).

[15] The auxiliary *con* marks the consecutivity of two events.

Beliao J., Lacheret A., & Kahane S. (2015) Interface intono-syntaxique en français parlé : compter quoi, compter comment, compter pourquoi ?. In S. Loiseau (ed), *La fréquence textuelle* [*Langage,* 197], 129-153, Larousse.

Berrendonner A. (1990). Pour une macro-syntaxe. In *Données orales et théorie linguistique* [*Travaux de linguistique (Gent)*, 21], 25-36.

Blanche-Benveniste C. (1990). *Le français parlé (études grammaticales)*. Editions du CNRS.

Bonami O., & Godard D. (2008). On the syntax of direct quotation in French. In *Proceedings of the 15th International Conference on HPSG*, CSLI Publications, 358-377.

Braggaar A. & van der Goot R. (2021). Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*

Caron, B, Courtin, M., Gerdes, K., & Kahane, S. (2019) A Surface-Syntactic UD Treebank for Naija. In *Proceedings of the 17th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

Cresti E. (1995). Speech act units and informational units. In E. Fava (ed), *Speech Acts and Linguistic Research,* Proceedings of the Workshop. Center for Cognitive Science, SUNY at Buffalo, 89-107.

Deulofeu, J. (1999). *Recherches sur les formes de la prédication dans les énoncés assertifs en français contemporain (le cas des énoncés introduits par le morphème que)*. Thèse d'état, Université Paris 3.

de Marneffe M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308.

Dister A., Goldman J.-P., & Marlet R. (2019) Orthographic and phonetic transcriptions of Rhapsodie recording. In Lacheret-Dujour et al. (2019), 21-34.

Dobrovoljc, K. & Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC),* 1566-1573.

Gerdes K., Guillaume B., Kahane S., & Perrier G. (2018) SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Universal Dependencies Workshop (UDW)*, EMNLP.

Gerdes K., Guillaume B., Kahane S., & Perrier G. (2019) Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic functions and deep-syntactic features. In *Proceedings of the 18th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

Gerdes K. & Kahane S. (2009) Speaking in piles: Paradigmatic annotation of French spoken corpus. In *Proceedings of the 5th Corpus Linguistics Conference*, Liverpool, http://ucrel.lancs.ac.uk/publications/cl2009.

Kahane S., & Lacheret A. (2019) Syntax and prosody mapping: What and how? The case of intonational periods and illocutionary units. In Lacheret-Dujour et al. (2019), 339-363.

Kahane S., Pietrandrea P. (2009) Les parenthétiques comme « Unités Illocutoires Associées » : une perspective macrosyntaxique, in M. Avanzi & J. Glikman (eds), *Les Verbes Parenthétiques : Hypotaxe, Parataxe ou Parenthèse ?* [*Linx*, 61], 49-70.

Kahane S., Vanhove M., & Ziane R. (2021) A morph-based and a word-based treebank for Beja. In *Proceeding of the 20th Proceedings of the 18th international conference on Treebanks and Linguistic Theories (TLT)*, SyntaxFest.

Lacheret A., Kahane S., Beliao J., Dister A., Gerdes K., Goldman J.-P., Obin N., Pietrandrea P., & Tchobanov A. (2014) Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. In *Actes du 4ème congrès mondial de linguistique française* (*CMLF*), *SHS Web of Conferences*, vol. 8, EDP Sciences, 2675-2689.

Lacheret-Dujour A., Kahane S., & Pietrandrea P. (eds) (2019) *Rhapsodie – A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Øvrelid L., Kåsen A., Hagen K., Nøklestad A., Solberg P. E., & Johannessen J. B. (2018). The LIA treebank of spoken Norwegian dialects. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC),* 4482-4488.

Patejuk A. & Przepiórkowski A. (2018). *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw. Downloadable from http://nlp.ipipan.waw.pl/Bib/pat:prz:18:book.pdf.

Pietrandrea P. & Kahane S. (2019) Macrosyntactic annotation. In Lacheret-Dujour et al. (2019), 97-125.

Pietrandrea P., Kahane S., Lacheret A., & Sabio F. (2014) The notion of sentence and other discourse units in corpus annotation. In T. Raso, H. Mello, M. Pettorino, *Spoken Corpora and Linguistic Studies*, Benjamins.

Pretkalniņa L., Rituma L., Saulīte B. (2018) Deriving enhanced Universal Dependencies from a hybrid dependency-constituency treebank. In *Proceedings of the 21sh International Conference Text, Speech, and Dialogue*, LNCS, Vol. 11107, Springer Link, 95-105.

Shriberg E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California, Berkeley.

Tyers, F. M. and Mishchenkova, K. (2020) Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW)*, 195-204.

Vanhove M. 2014. The Beja Corpus. In Mettouchi, A. and C. Chanard (eds.). *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. http://dx.doi.org/ 10.1075/scl.68.website.

Wong T., Gerdes K., Leung H. & Lee J. S. (2017) Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 266-275.

Zeldes, A. (2017) The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3), 581–612.

A morph-based and a word-based treebank for Beja

Sylvain Kahane*, Martine Vanhove, Rayan Ziane***, Bruno Guillaume******
*Modyco, Université Paris Nanterre & CNRS
**Llacan, CNRS, INALCO, & EPHE
***Llacan, CNRS, INALCO, & Université d'Orléans
****Sémagramme, INRIA Nancy Grand Est

Abstract

The paper presents the first UD treebank for Beja, a Cushitic language spoken in Sudan. It has been built from the conversion and enhancement of an Interlinear Glossed Text (IGT). The paper's objectives are three-fold: we explain our choice to use a morph-based annotation and its consequences, we describe the processing chain from an IGT to a morph-based dependency treebank and a word-based treebank, and we present several interesting constructions in Beja.

1 Introduction

This paper presents a small treebank for Beja, a Cushitic language spoken in Sudan. Initially developed in SUD (Surface-Syntactic Universal Dependencies) (Gerdes et al. 2018, 2019, 2021), the treebank is also available in UD (de Marneffe et al. 2021). It has been built from an Interlinear Glossed Text (IGT) (Comrie et al. 2008) developed by Martine Vanhove (2014). The original corpus contains 5899 words and 12507 morphs representing slightly less than one hour of recordings divided into 18 files. Two files, containing 1101 morphs, 418 stems, and 56 sentences, have been completely annotated and constitute the UD2.8 Beja-NSC treebank.[1]

One of our goals was to avoid losing information contained in the original resource, which led us to adopt a morph-based tokenization rather than a word-based annotation.

The paper focuses on three aspects of developing this treebank. Section 2 presents the Beja corpus and the IGT annotation we started with, the UD annotation scheme, and the adjustments to it which were necessary for our morph-based annotation. An overview of the conversion to a word-based treebank is also provided. Section 3 explains the processing chain from an IGT to a UD treebank and the optimization of this chain. Section 4 introduces some challenges faced during the syntactic annotation of Beja.

2 A morph-based annotation for Beja

2.1 Beja and CorpAfroAs

Beja is the sole member of the North Cushitic branch of the Afroasiatic phylum. It is mostly spoken in eastern Sudan, as well as in southern Egypt and northern Eritrea. In Sudan, the country where the data were collected, the number of speakers is about 2,000,000, but the language has no official recognition and exists purely as an oral language. As explained by Martine Vanhove (2006), Beja is not poorly described compared to other Sudanese languages, and the most recent grammar, published in French, goes back to 2017 (Vanhove 2017). However, some elements required for a complete description of the language remain unavailable.

The data used for the development of this treebank comes from the CorpAfroAs project (Mettouchi & Chanard 2010). CorpAfroAs is a multilingual corpus which aimed at providing a structured database of natural records of Afroasiatic languages, transcribed, translated and annotated to allow for complex

[1] The first version of the treebank published on May 1st, 2021, for UD2.8, and released on November 1st, 2021, for UD2.9, is a morph-based treebank. New modifications have been done for the publication of this article that will be incorporated for UD2.10 on May 1st, 2022. We also plan to publish the word-based version of the treebank on the same occasion.

requests. CorpAfroAs is organized around two axes: prosodic analysis and morphosyntactic glossing. It is this morphosyntactic glossing that served as the raw material for our work.

2.2 Beja's Interlinear Glossed Text

All CorpAfroAs corpora use a common format for IGTs presented in Comrie (2015).

(1) *w=ʔi:d arraf-i / a-di=t a- ba i-ni //*
DEF=Aid congratulate-AOR.1SG 1SG-say\PFV=COORD 1SG-go 3SG.M-say\PFV
'I went to wish him a blessed Aid, he said.'

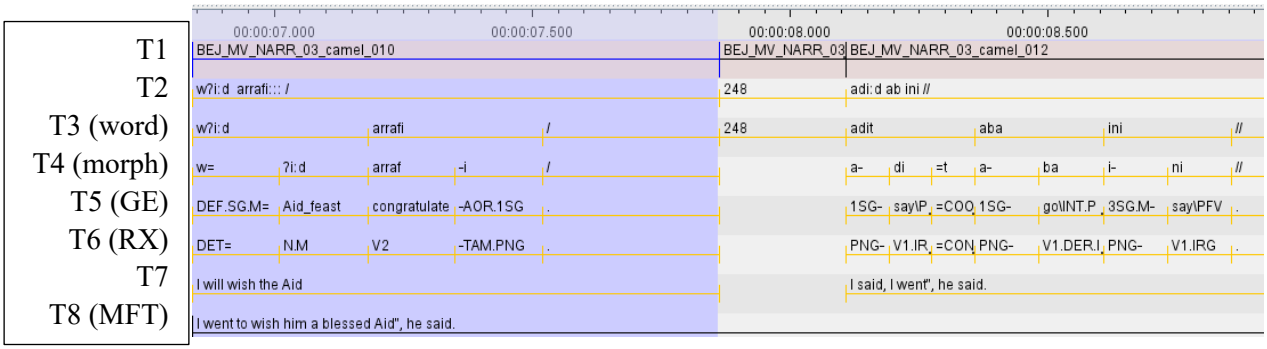


Figure 1. ELAN screenshot of the CorpAfroAs IGT for (1)

The top of Figure 1 contains the timeline. Tier 1 is the segmentation into prosodic units. Tier 2 is a broad phonetic transcription; / marks a minor prosodic break (rising final continuative intonation) and // a major prosodic break (falling final intonation). Tier 3 contains the prosodic words and Tier 4 a tokenization into the smaller units with a non-segmentable signifier (i.e. the morphs, as defined by Haspelmath 2020). The tokens considered are lexical stems, inflected forms of stems (when the inflectional "morphemes" are not affixes), inflectional affixes, and derivational affixes. For the sake of simplicity, we will call these tokens *morphs*, even if some of them are a combination of morphemes, such as *ni*, which is the perfective form of the stem of the verb *di* 'to say'. This inflected stem also combines with a prefix *i-*, which is a subject agreement morph, giving the verbal form *ini* at the end of example (1).

Tier 5, labelled GE, is a gloss. Tier 6, labelled RX, contains morphosyntactic features, including POS (DET, N, V1…) and inflectional categories (TAM, PNG for Person-Number-Gender, …).

Tier 7 is a translation of each prosodic unit and Tier 8, called the "major free translation" (MFT), is a translation based on larger units, allowing for better translations. It is this last tier which was used as the basis for the sentence segmentation. Since the end of each MFT unit does not necessarily corresponds to the end of a sentence, understood as a coherent syntactico-semantic unit, we copied the MFT tier (Mft-cp) on which we signaled the end of each sentence by a # sign (Figure 2).[2]

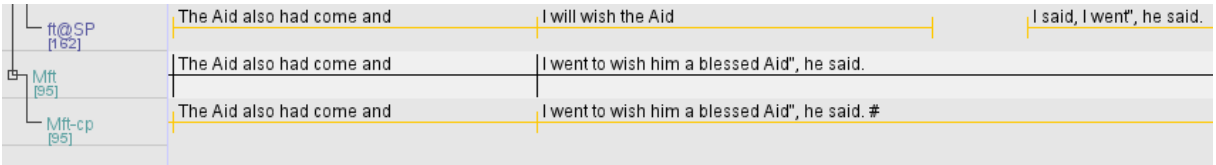


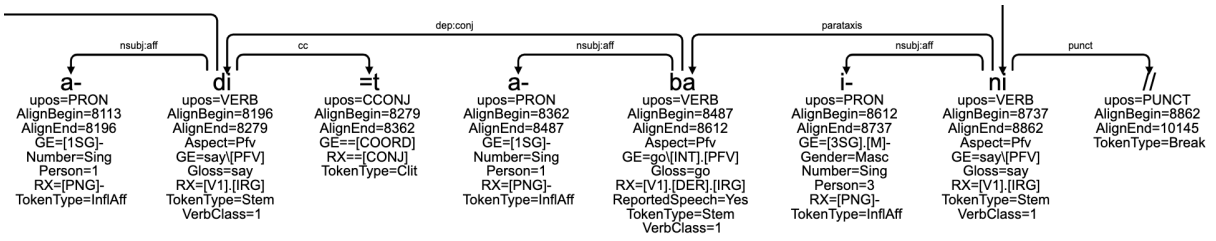
Figure 2. ELAN screenshot of the sentence segmentation

The IGT does not contain lemmas (which supposes a more advanced description and a lexicon), but contains glosses and translations. The corpus contains a time alignment and each IGT sample is coupled with a sound file accessible on the CorpAfroAs website.

[2] The segmentation of spoken corpora into major syntactic units (often called *sentences*, even if the notion can be problematic for spoken production) is a complex question that will not be addressed here. See Kahane et al. (2021) for some guidelines and Pietrandrea et al. (2014) for a more comprehensive study.

2.3 A morph-based annotation scheme for SUD and UD

We use the "morph" segmentation for tokenization. The content of the tiers GE and RX is kept in features GE and RX. The time alignment gives us the features AlignBegin and AlignEnd of each token, including the prosodic breaks (Figure 3) (see Kahane et al. 2021 for the conventions we use for spoken corpora).



'I said and I went, he said'

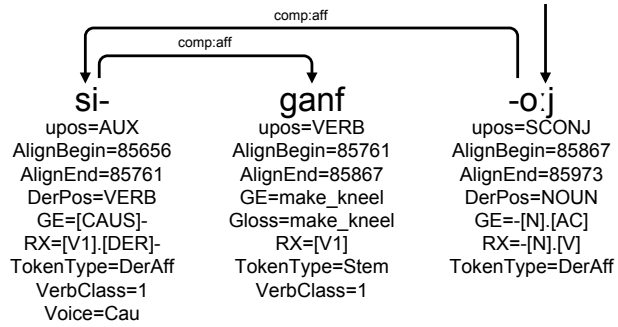
Figure 3. The UD annotation for the end of (1)

Some pieces of information of the GE tier are used to instantiate morphosyntactic features, such as Number, Person, Gender, Aspect, Definiteness… (see Section 3.2). A TokenType is added on each token, with five values: Stem for stems (including some inflected stems), Clit for clitics, InflAff for inflectional affixes, DerAff for derivational affixes, and Break for prosodic breaks.

In addition to the usual #text and #sent_id features (see online UD guidelines or de Marneffe et al. 2021), the metadata contain a #sound_url feature for the URL of the sound file corresponding to the transcription and #phonetic_text for the phonetic transcription from Tier 2 (Kahane et al. 2021). The #text value, which is the concatenation of the tokens, including simple and double hyphens, is distinct from the #phonetic_text value.

We decided to give each token a POS as if it was a word. In consequence, pronominal affixes are PRON, verb nominalizers are SCONJ, TAM or causatives are AUX, case markers are ADP, plurals are DET, and purely phonetic signs are PART.

In our SUD annotation, SCONJ, ADP, and AUX affixes are treated as governors of their base, as if they were words (Gerdes et al. 2018). We use the corresponding SUD syntactic relation with the *aff* extension: In other words, subject pronominal affixes are *subj:aff* of their base, plurals are *det:aff*, while for SCONJ, ADP, or AUX, the base is *comp:aff* of the affix. Figure 4 gives the example of the deverbal noun *siganfo:j* 'settling' where the verbal stem *ganf* 'make kneel' combines with two derivational affixes, the causative prefix *si-* and the nominalizer *-o:j*,



'settling'

Figure 4. Derivational affixes (SUD-style)

Note that the fact that some affixes are treated as heads allows us to indicate in which order they combine. In the case of *siganfo:j*, *ganf* combine first with the causative *si-* and then with the nominalizer *-o:j*. Derivational affixes receive an additional feature *DerPos* indicating the POS of the derived form: AUX affixes have a DerPos=VERB and SCONJ affixes have a DerPos=NOUN. For inflected forms, the POS of the inflected form remains the POS of the stem. Note also that in the case of an inflected

form, the base can have its own dependents, while in the case of a derived form all dependents are on the derivational affix.[3]

In the UD version of the morph-based, all affixes are dependent of the stem. We use the UD syntactic relation corresponding to their functional role, with an additional *aff* extension: subject pronominal affixes are *nsubj:aff*, case markers are *case:aff*, nominalizers are *mark:aff*, AUX affixes are *aux:aff*, plurals are *det:aff*. Figure 5 gives the UD version of the two words of Figure 4. (See Figure 3 of the example of pronominal affixes in UD-style.) Note that the order in which the affixes combine with the stem is lost in the UD version. This is a problem for the conversion to the word-based version, because we cannot easily determine whether *siganfo:j* is a noun or a verb (see Gerdes et al. 2021 for a similar discussion about the fact that UD underspecifies the internal structures of nuclei). For this reason, the word-based UD version of the treebank is derived from the morph-based and word-based SUD versions and not from the morph-based UD version.

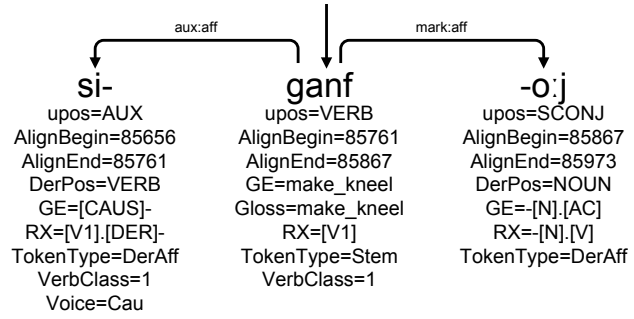


Figure 5. Derivational affixes (UD-style)

2.4 From a morph-based treebank to a word-based treebank and back

UD theoretically requires treebanks to be word-based. However, we think that the fact that our treebank is morph-based does not pose a problem because the morph-based annotation is explicit, due to the different features we introduced (*TokenType* on tokens and *aff* on relations) and because it is not difficult to merge every stem with its affixes to obtain a word-based treebank. Note that the question to have morph-based (generally called morpheme-based) treebank rather than word-based treebank has been discussed several times for different languages: see Tsarfaty & Goldberg (2008) for Modern Hebrew, Vincze et al. (2010) for Hungarian, Zhan et al. (2014) for Chinese, or Park (2017) for Korean.[4]

For the conversion into a word-based treebank, the lists of morphosyntactic features attached to the different parts of a word must be merged in different ways. Some features had to be concatenated, such as the feature *form*, containing the form, and the features *GE* and *RX* containing the morphosyntactic glosses. See Figure 6 for the word-based version of Figures 4 and 5.

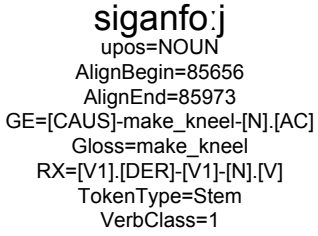


Figure 6. The word-based annotation of the derived word *siganfo:j*

[3] Compare in English:

(i) *He cleaning the table was impressive.*
(ii) *His cleaning of the table was impressive.*

In (i), *cleaning* is an inflected form and *clean* can have a subject and a direct object, while, in (ii), *cleaning* is a derived noun with a determiner and a noun complement.

[4] We would like to thank one of our reviewers, who pointed out that our segmentation was morph-based rather than morpheme-based.

The most challenging feature is *upos* (the UD feature for the "universal" POS) for derived forms. Thanks to the SUD annotation where the derivational affixes are head and to the *DerPos* feature, it becomes trivial to compute the *upos* of a derived form: it is the *DerPos* of the topmost derivational affix. Except for the features *form, GE,* and *RX*, which are concatenated, and *upos*, which is replaced by *DerPos*, the features of the derived form are the features of the topmost derivational affix.

For inflected forms, the *upos* of the word is the *upos* of the stem. The features *form, GE,* and *RX* are concatenated as for derived forms, but contrary to derived forms, other features are unified for inflected forms. Figure 7 shows the word-based version of the morph-based analysis of Figure 3, where we can see that the *Person* and *Number* features coming from the pronominal affixes are reported on the verb forms.

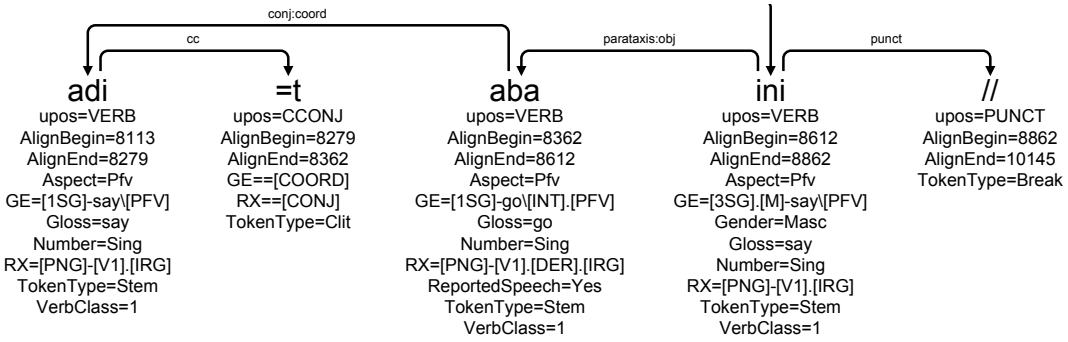
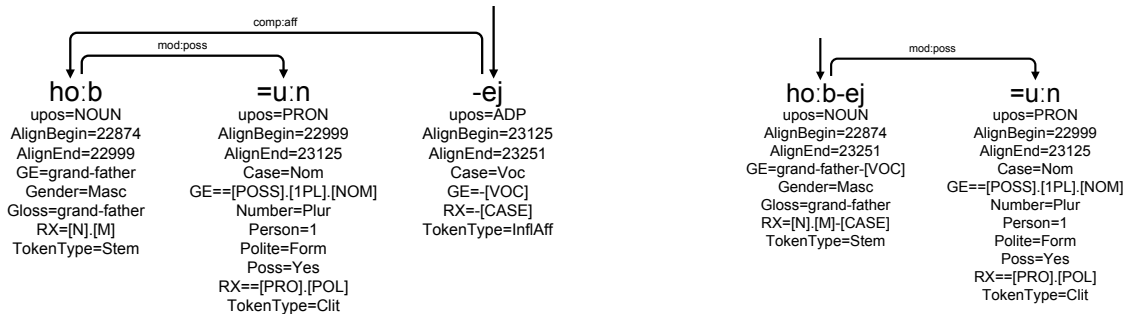


Figure 7. The word-based UD annotation for the end of (1)

We faced one unexpected problem with a clitic placed between a stem and an inflectional affix. We analyzed this case as an amalgam with only one word corresponding to two lexemes (Figure 8).



'our grand-father' 'our grand-father'

Figure 8. The morph-based and word-based SUD annotation of an incorporated clitic

We must also compute the *AlignBegin* and *AlignEnd* features of words, which are easily deductible from the corresponding features of the morphs. Note that in the word-based version of the treebank, some information is lost and it will not be possible to recover the segmentation into morphs, as well as the features associated to this morphs. This is why we decided to distribute the Beja UD@2.8 treebank in a morph-based version. On our side, we will maintain the morph-based SUD version, which is the most informative one (see in particular the discussion of Section 3 when the word contains two derivational affixes).

3 From IGT to UD

The construction of the UD treebank was carried out in three steps: the conversion of the IGT into a CoNLL-U (Section 3.1); the automatic pre-annotation by enrichment of the CoNNL-U (Section 3.2); the manual SUD annotation, the conversion to UD and the validation of the treebank (Section 3.3).

3.1 From IGT to CoNLL-U

The first obstacle to the conversion to CoNLL-U is the fact that the corpus is segmented into prosodic units that do not necessarily correspond to syntactic units. We decided to base our major segmentation on the "major free translation" segmentation, which corresponds more or less to illocutionary units as illustrated in Figure 2 in section 2.2. The tokenisation is based on the segmentation into "morphs".

Once the choice of the tiers for the tokenization and the segmentation into sentences is made, the conversion of the IGT to a CoNLL-U is straightforward and loses no information from the IGT format. For each token, the time alignment is stored in the features *AlignBegin* and *AlignEnd*, the content of the GE and RX tiers is stored in the features *GE* and *RX*.

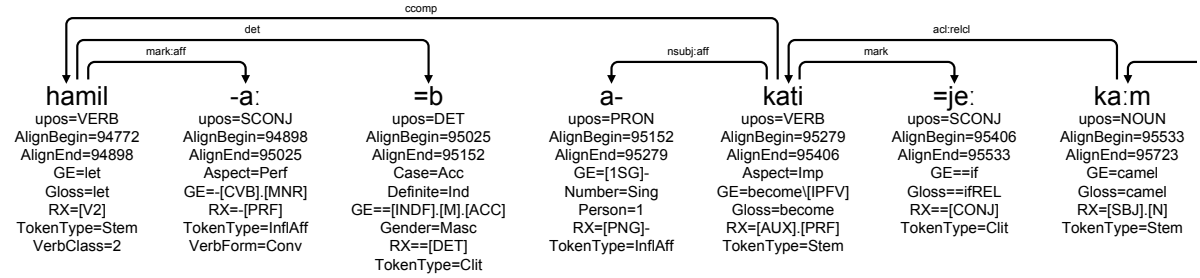
3.2 Automatic pre-annotation

The first CoNLL-U we obtain is almost similar to the IGT. The second step consists in enriching this CoNLL-U by transferring the content of the GE and RX tiers into UD features in order to fit the UD annotation scheme. The annotation specific to the morph-based level was introduced entirely automatically.

As the GE and RX formats of CorpAfroAs IGTs are enriched versions of the Leipzig Glossing Rules (Comrie 2015), they allow us to infer the UD POS and all the UD morphosyntactic features that must be associated with the tokens. We built a lexicon that proposes a translation into a UD feature for each label used in the GE and RX tiers. It was also possible to infer the syntactic relation for many tokens. The Grew tool (Guillaume, 2021), through its graph rewriting function, makes it possible to write a grammar of rules matching elements within dependency trees.

The feature *TokenType*, which distinguishes stems, affixes, clitics, and prosodic breaks, is based on the form of the token: As usual in IGTs, affixes have a hyphen (*a-* or *-a*) and clitics a double hyphen (*ba=* or *=i*), while prosodic breaks are assigned to special symbols (/ and //). For affixes and clitics, the governor was the closest stem and the positions of the hyphens indicate if the stem occurs after or before them. The distinction between inflectional and derivational affixes can be computed from the syntactic *RX* feature (most derivational affixes have a DER value in RX).

The syntactic label set of CorpoAfroAs, corresponding to the RX tier, is richer than the UD *upos* set of POS and the POS conversion was trivial for most of the labels. For instance, the labels V1, V2, LV, and IRG are all converted to the VERB *upos* tag. In order not to lose information, V1 and V2 receive a *VerbClass* feature with values 1 and 2 according to the original label. LV is provided with a VerbType=Light feature. In a similar way, the label DEM is converted into a DET *upos* with the PronType=Dem feature. The label REL for relativizers gives us an SCONJ *upos*, as well as the SUD relation *mod@relcl* (translated into UD *acl:relcl*). Moreover, due to the head-final behaviour of Beja, the relativizer can be linked to the verb preceding it (Figure 9).[5]



'a camel that I had let loose'

Figure 9. Relative clause (UD-style)

[5] This figure, as well as all the following figures, is extracted from the morph-based UD version of the treebank, revised for the paper, and which will be distributed on May 1st, 2022, for the UD2.10 release. All the SUD versions of the treebanks are available on the SUD website https://surfacesyntacticud.github.io/ and all versions can be requested on the Grew-match website (the latest versions of UD treebanks converted from SUD treebanks are at the end of the list of UD treebanks).

3.3 Manual SUD annotation and conversion from SUD to UD

The manual annotation was carried out in the SUD annotation scheme by a linguist specializing in Beja (Martine Vanhove), with the help of a specialist of treebank annotation (Sylvain Kahane) and a master student in NLP (Rayan Ziane), as well as some feedback from two native speakers (Ahmed Mohamed-Tahir Hamid and Mohamed-Tahir Hamid Ahmed). It was not possible to have a double annotation for this language. Some problems of analysis we faced during the annotation process are discussed in the next section. For the conversion from SUD to UD, we had to customize the conversion of the relations introduced for the affixes. The fact that UD forces the coordination relation *conj* to be head-initial was also a problem and SUD head-final *conj* relations were converted into an ad hoc *dep:conj* UD relation (see Section 4.2). The different conversions were mastered by Rayan Ziane and Bruno Guillaume.

4 Some constructions of Beja

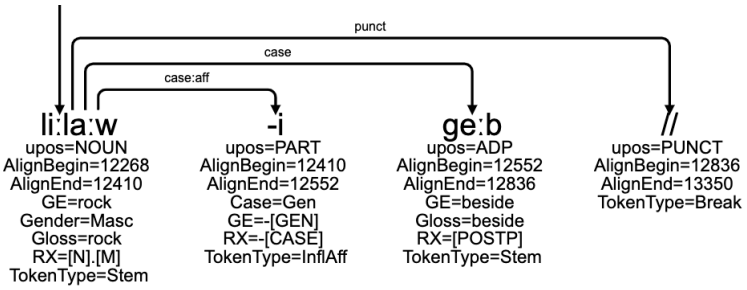
Below, we discuss four features of Beja syntax: affixes and word order (4.1), coordination (4.2), relative clauses (4.3) and serial verb constructions (4.4).

4.1 Affix and word order

The Beja treebank contains 684 words if we count both stems and clitics; 39% of words are clitics. The treebank contains 244 affixes for 418 stems, or a proportion of 58%. 59% of them are suffixes and 41% prefixes. 88% of the affixes are on verbs, 7% on nouns, and 5% on auxiliaries. All affixes on nouns are suffixes. Not all inflectional morphemes are affixes: 44% of the stems are in fact inflected forms, containing an inseparable inflectional morpheme, which increases the proportion of inflectional morphemes to 102% (102 inflectional morphemes for 100 stems).

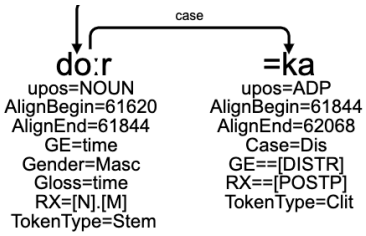
Beja is a head-final language: only 11% of the dependencies between two stems have the governor before the dependent in the SUD version of the treebank. Among the 31 dependencies concerned, 11 are for modifiers, 6 for discourse markers, 4 for dislocated objects, 2 for objects in canonical position, and 2 for determiners. Clitics occur on both sides: 47% are proclitics and 53% are enclitics. Clitics are mainly on verbs (56%) and nouns (38%). Clitics on nouns are determiners (70%), possessives (15%), postpositions (11%), and coordinating conjunctions (3%). Clitics on verbs are subordinating conjunctions (35%), object pronouns (25%), coordinating conjunctions (14%), an optative particle (2%), and, on nominalized forms, determiners (15%) and copulas (8%).

Beja adpositions are postpositions, either as independent words (Figure 10) or as enclitics (Figure 11):



‘next to a rock’

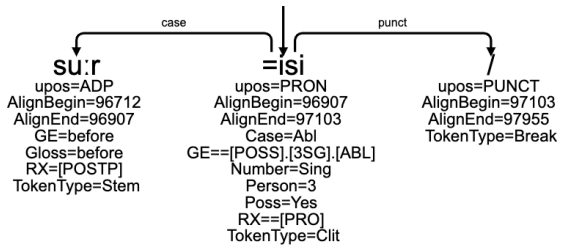
Figure 10. A Beja independent postposition (UD-style)



'in turn'

Figure 11. A Beja enclitic postposition (UD-style)

When the postposition complement is a pronoun, it is an enclitic and the postposition precedes its complement (Figure 12):

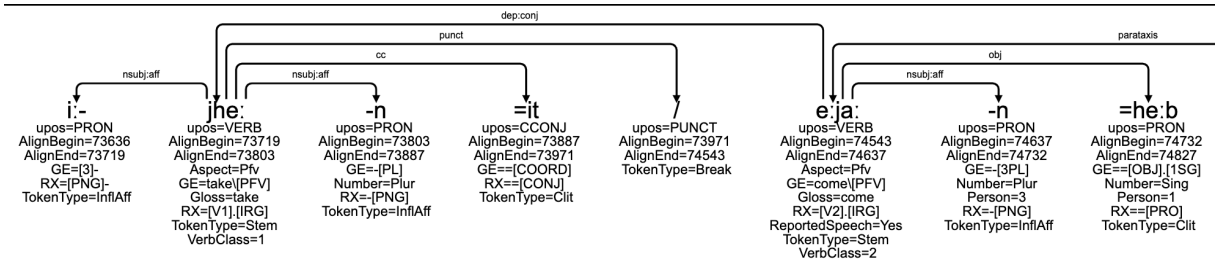


'in front of her'

Figure 12. Postposition with an enclitic pronoun (UD-style)

4.2 Coordination in a head-final language

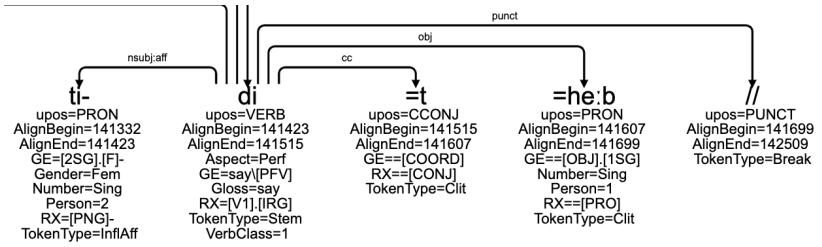
In Beja, verbal and nominal coordination are expressed with different enclitic morphemes. The texts contain 20 tokens of the verbal coordinating conjunction *=t* (and its allomorphs *=it* and *=ajt*) (Figure 13). For half of the tokens, the conjunctions occur at the end of a prosodic unit, be it a major or a minor prosodic break, or a sentence. For this reason, we attach the coordinating conjunctions to the first conjunct and we consider that the second conjunct is the head of the coordination. See Kanayama et al. (2018) for a similar analysis in two other head-final languages, Japanese and Korean. As the *conj* relation is forbidden from right to left in UD, we introduced a *dep:conj* relation.



'they took them and they came to me'

Figure 13. Verbal coordination (UD-style)

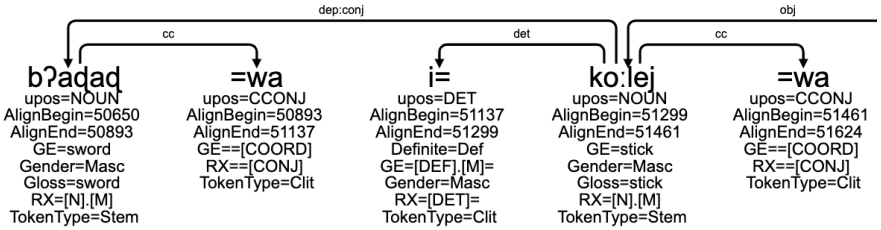
The verbal coordinating conjunction is tightly linked to the right of the verb. It even occurs before enclitic object pronouns (6 tokens), as in Figure 14.



‘you told me and …’

Figure 14. Position of the verbal coordinating conjunction (UD-style)

The nominal coordinating conjunction is the enclitic *wa*. It is expressed on each conjunct as shown in Figure 15.



‘a sword and the stick’

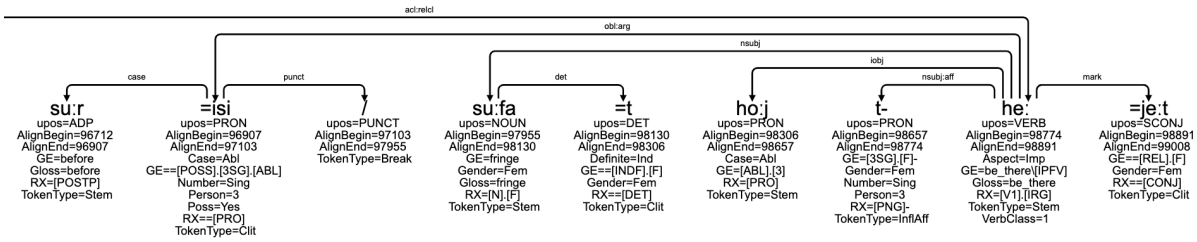
Figure 15. Nominal coordination (UD-style)

4.3 Relative clauses

In the texts, only object relative clauses occur, for which the number of tokens amounts to 18. They are marked by several clitics, either proclitics (*ji=, j=, wi=, w=*), or enclitics (*=e:b, =e:t, =t, =b, =e:*). There are also instances of a zero morph. 7 tokens were found with a preposed antecedent and 11 tokens with a postposed antecedent.

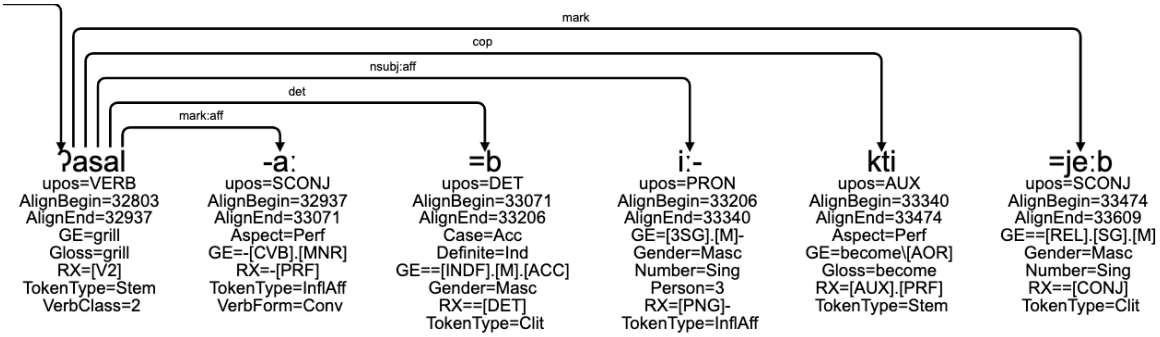
The anteposition of an antecedent is an unusual word order in verb-final languages. The SUD annotation revealed that this construction occurs in two contexts linked to information structuring:

1. when the object of the transitive verb of the matrix clause is topicalized (in Figure 16 two relative clauses precede the verb of the matrix clause)
2. when the relative comes as an afterthought (Figure 17).



‘(a girl) who has a fringe in front of her’

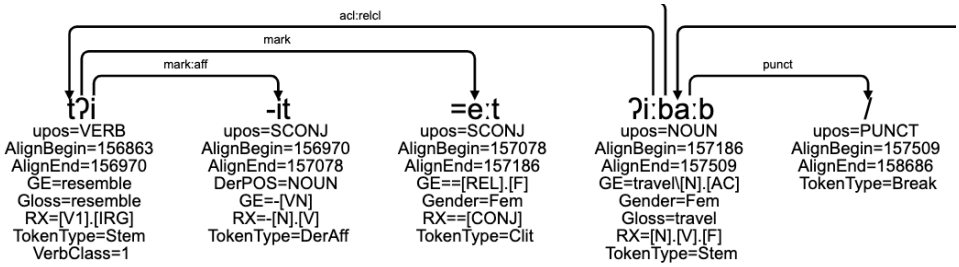
Figure 16. Topicalized preposed antecedent (UD-style)



‘(The man was carrying on his shoulder a lamb.) That he had grilled.’

Figure 17. Preposed antecedent in an afterthought (UD-style)

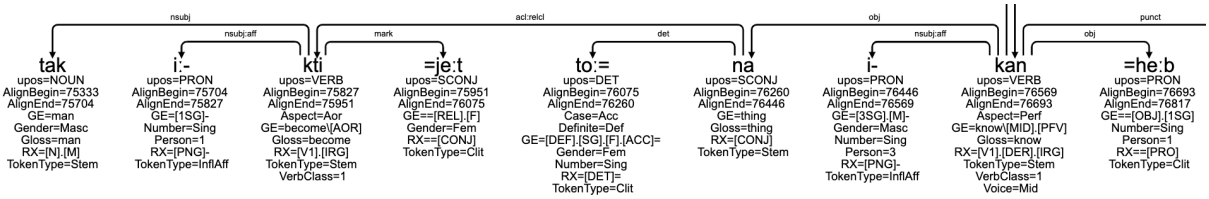
Otherwise the canonical constituent order is used: relative clause – object antecedent – main verb (Figure 18).



‘a story like that (happened to me)’, lit. a travel that resembles

Figure 18. Antecedent in canonical word order (UD-style)

Complement clauses may also be formed on the basis of a relative clause. In such cases, the antecedent, which is the dummy noun *na* ‘thing’, is always placed after the relative clause, i.e. the canonical constituent order (Figure 19).

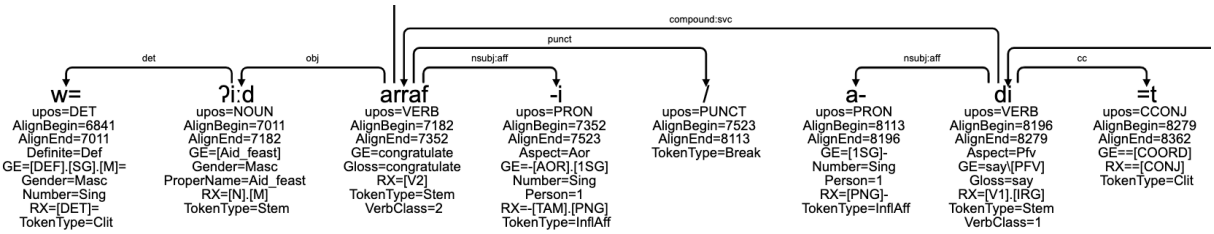


‘he realized that I was a man’

Figure 19. Relative-based complement clause (UD-style)

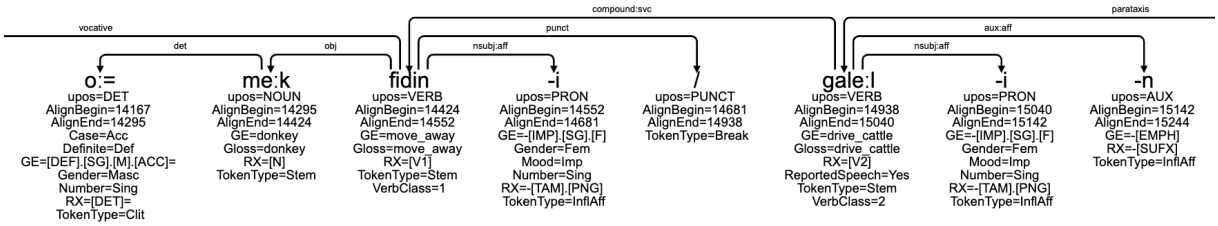
4.4 Non canonical Serial Verb Constructions

We chose to label as SVCs any series of two finite verbs of the same semantic domain which are not coordinated and do not have a predicate–argument relation. This characterization only partly complies with e.g. Haspelmath’s (2016: 296) narrow definition of serial verb constructions as a comparative concept: “A serial verb construction is a monoclausal construction consisting of multiple independent verbs with no element linking them and with no predicate–argument relation between the verbs.” Such a definition does not impose any semantic restriction on the semantic domains of SVCs (apart from expressing a dynamic event, Cleary-Kemp 2015: §4.2.1.3; Haspelmath 2016: 302), as is the case in Beja. Moreover, SVCs seem to be unproductive, limited to very few semantic domains, and restricted to series of two verbs. There are 3 occurrences of SVCs in our data, 1 with a verb of saying (Figure 20), and 2 with motion verbs (Figures 21) conjugated at various TAM.



‘(I went) to wish him a blessed Aid’

Figure 20. SVC with verbs of saying (aorist + perfective) (UD-style)



‘chase the donkey away!’

Figure 21. SVC with motion verbs (both imperfective) (UD-style)

5 Conclusion

Beja belongs to a sub-family of Afroasiatic languages, the Cushitic languages, for which there were no treebanks. It is a language with a rich morphology and which, unlike its cousins, the Semitic languages, is a head-final language with non-canonical serial verb constructions.

The Beja treebank that we present is a very small treebank, but richly annotated, with a segmentation into morphs, glosses, and an alignment to the sound file.

While developing a morph-based treebank for Beja we have been led to bring some enrichments to the SUD and UD annotation schemes. We introduced the *TokenType* feature which takes five possible values (*Stem, InflAff, DerAff, Clitic, Break*) and an *aff* extension for syntactic relations to indicate more explicitly the internal relations of words. We also introduced the feature *DerPos* on derivational morphs for indicating the POS of the derived form.

We have also seen that the SUD version of the morph-based treebank makes it easier to compute the word-based version of the treebank, since it explicitly indicates the internal structure of the word and the order in which the affixes combine with the stem.

As it is possible to convert the morph-based annotation into a word-based annotation, we think it is better to distribute the morph-based annotation, which contains more information and is closer to the format that field linguists use. This format allows us to extract qualitative and quantitative information about the inflectional morphology of the language, which is extremely useful for typological studies (Greenberg 1960).

The Universal Dependency project was initially developed to unify treebank annotation schemes in order to have a common format for the development of NLP tools. The UD annotation scheme is heavily based on the output format developed for the Stanford parser for English (de Marneffe et al. 2006). The 33 languages of the UD1.2 (Nivre et al. 2016) were all languages with long-standing writing traditions, and all corpora were written corpora following well-established orthographic conventions, most of them with a segmentation into words.

UD is now integrating a wide range of new languages coming from different families. Many field linguists having data that are already analyzed in IGT are ready to enrich their corpus with a syntactic annotation. It is necessary that UD offer the possibility of a morpheme-based view of annotation, which allows them to keep the IGT structure. This paper is a first step in this direction by setting up a processing chain to convert an IGT into a morph-based treebank, then a word-based treebank.

Acknowledgements

We would like to thank our reviewers for their valuable remarks. The master internship of Rayan Ziane has been funded by the ANR project NaijaSynCor (2017-2021), directed by Bernard Caron.

References

Cleary-Kemp, J. 2015. *Serial Verb Constructions Revisited: A Case Study from Koro.* PhD dissertation, University of California at Berkeley.

Comrie, B. 2015. From the Leipzig Glossing Rules to the GE and RX lines. In A. Mettouchi, M. Vanhove & D. Caubet (eds.), *Corpus-based Studies of Lesser-described Languages*. John Benjamins, 207-219.

Comrie, B., Haspelmath, M., & Bickel, B. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology & the Department of Linguistics of the University of Leipzig. Retrieved January, 28, 2010. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

CorpAfroAs: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages. http://dx.doi.org/10.1075/scl.68.website.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*.

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. 2021. Universal dependencies. *Computational Linguistics*, 47(2), 255-308.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Universal Dependencies Workshop (UDW)*.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. 2021. Starting a new treebank? Go SUD!. In *Proceeding of the 6th conference on Dependency Linguistics (Depling)*.

Gerdes, K., Kahane, S., & Chen, X. (2021). Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*, 6(1).

Greenberg, J. H. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3), 178-194.

Guillaume, B. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*, 168–175.

Haspelmath, M. 2016. The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics* 17(3), 291–319.

Haspelmath, M. 2020. The morph as a minimal linguistic form. *Morphology* 30: 117–134. https://doi.org/10.1007/s11525-020-09355-5.

Kahane, S., Caron, B., Gerdes, K., & Strickland, E. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Kanayama, H., Han, N. R., Asahara, M., Hwang, J. D., Miyao, Y., Choi, J. D., & Matsumoto, Y. 2018. Coordinate structures in universal dependencies for head-final languages. In *Proceedings of the Second Universal Dependencies Workshop (UDW)*, 75-84.

Mettouchi, A., & Chanard, C. 2010. From Fieldwork to Annotated Corpora: The CorpAfroAs Project. *Faits de Langue-Les Cahiers n°2*, 255-265.

Park, J. 2017. Segmentation granularity in dependency representations for Korean. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, 187-196.

Pietrandrea, P., Kahane, S., Lacheret-Dujour, A., & Sabio, F. (2014). The notion of sentence and other discourse units in corpus annotation. In T. Raso, H. Mello, M. Pettorino (eds.), *Spoken Corpora and Linguistic Studies,* John Benjamins, Amsterdam, 331-364.

Tsarfaty, R., & Goldberg, Y. 2008. Word-Based or Morpheme-Based? Annotation Strategies for Modern Hebrew Clitics. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*.

Vanhove, M. 2006. The Beja language today in Sudan: The state of the art in linguistics. *Proceedings of the 7th International Sudan Studies Conference*. Bergen: University of Bergen, CD Rom.

Vanhove, M. 2014. The Beja Corpus. In Mettouchi, A. and C. Chanard (eds.). *The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*. http://dx.doi.org/ 10.1075/scl.68.website.

Vanhove, M. 2017. *Le Beja*. Leuven, Paris: Peeters (coll. Les Langues du Monde 9).

Vincze, V., Szauter, D., Almási, A., Móra, G., Alexin, Z., & Csirik, J. 2010. Hungarian dependency treebank. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*.

Zhang, M., Zhang, Y., Che, W., & Liu, T. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1326-1336.

Towards Building a Modern Written Tamil Treebank

Parameswari Krishnamurthy
Centre for Applied Linguistics and Translation Studies
University of Hyderabad, India
pksh@uohyd.ac.in

Kengatharaiyer Sarveswaran
University of Moratuwa, Sri Lanka.
Department of Computer Science,
University of Jaffna, Sri Lanka.
sarves@univ.jfn.ac.lk

Abstract

In this paper, we describe the creation of a morphosyntactically annotated treebank for modern written Tamil following the Universal Dependencies (UD) framework to support the implementation and evaluation of Tamil dependency parsers. At present, this treebank consists of 534 sentences. This paper discusses unique constructions found in Tamil and explains sub-relations and language-specific relations introduced, apart from outlining the methodology. This carefully annotated treebank can also serve as the benchmark dataset to evaluate Tamil Natural Language Processing (NLP) tools. The treebank will be extended further to cover more complex constructions in Tamil, and annotations will be enriched by incorporating the Enhanced Universal Dependencies scheme.

1 Introduction

The paper presents a treebank for modern written Tamil following the Universal Dependencies (UD) framework called Modern Written Tamil Treebank (MWTT). The sentences in MWTT are extracted from Lehmann's *A Grammar of Modern Tamil* (Lehmann, 1989), which consists of various well-formed sentences from modern written Tamil covering different sentence structures. The first and the current release of the treebank has 534 sentences containing 2536 tokens.

Tamil is a Dravidian language spoken natively by more than 78 million people worldwide,[1] including in India, Sri Lanka, Malaysia, Singapore, and Mauritius. Despite its significant speaker population and historical time depth, Tamil is low-resourced from the perspective of Natural Language Processing (NLP) (Bhattacharyya et al., 2019). Although there have been enormous efforts in creating resources and building NLP applications for Tamil, most of them are not available for public use or obsolete/not maintained.

Neural-based approaches are the state-of-the-art when it comes to the development of NLP applications, including syntactic parsing. These approaches require a significant amount of annotated data for training. However, there are no morphosyntactically annotated data with acceptable quality and with a wide syntactic structural coverage available publicly to develop and evaluate applications. In this context, we have created the UD-based treebank carefully. This paper also gives an account of unique syntactic constructions in Tamil, which we encountered when analysing and tagging simple sentences.

2 Review of Literature

2.1 Tamil Treebanks

There are many syntactic annotation schemes that are being used to create treebanks, including PennTreebank (Marcus et al., 1993), Prague Dependency Treebank (Böhmová et al., 2001), AnnCorra (Bharati et al., 2006), and the UD (Nivre et al., 2016). There have been few attempts to create treebanks for Tamil using these schemes. However, except Loganathan's Tamil PDT (Ramasamy and Žabokrtský, 2012) and UD_Tamil_TTB,[2] others are not available for public use.

[1] http://www.languagesgulper.com/eng/Tamil.html
[2] https://github.com/UniversalDependencies/UD_Tamil-TTB/tree/master

Tamil_TTB is mapped from Tamil PDT using a script. We noticed several inconsistencies and errors in tokenisation and annotation in the Tamil_TTB treebank. For instance, some words are segmented incorrectly. Although in Tamil, nouns with dative case marking can be used to mark subjects, obliques and indirect objects, it is incorrectly used to mark objects at least in 37 instances, and indirect objects are wrongly marked as objects at least in eight instances. Further, there are inconsistencies with the usage of tags `nmod` and `obl` that are found widely in the treebank.

There is also a need to create an error-free gold standard treebank for Tamil, which can be used as a benchmark dataset, as so far, different researchers have used different datasets to validate the system they developed. The current attempt is to build such annotated dataset covering different sentence structures for modern written Tamil.

2.2 Universal Dependencies Framework

UD (Nivre et al., 2016) is a dependency framework, which proposes a morphosyntactic annotation scheme. This cross-linguistically consistent scheme has been developed by deriving existing standards on POS, morphology, and dependency annotations to facilitate multilingual research studies and parser development. The dependency relations are created between syntactic words; words that have more than one syntactic information are broken into separate tokens before the relations are established.

We identified that most of the newly created treebanks, even low-resource and morphosyntactically-rich languages, are annotated using the UD scheme; hence adopting it for Tamil is also beneficial. This allows us to create cross-lingual mapping with other languages, and to make use of tools and resources which are already built around UD.

2.3 Syntactic Parsers for Tamil

Tamil is morphologically rich and relatively free-order in nature. It has an (S)OV word order with left-branching. Several attempts have been made to develop syntactic parsers for Tamil using various formalisms. However, apart from *Thamizhi*UDp (Sarveswaran and Dias, 2020), and the other off-the-shelf parsers such as Stanza (Qi et al., 2020), UDPipe (Straka and Straková, 2017), and TranKit (Nguyen et al., 2021) others are not available publicly to use and build upon. All available parsers are implemented using state-of-the-art neural-based approaches. These approaches require more and more annotated data to improve the parsing accuracy. Further, it is also noticed that there is no well-curated benchmark dataset to evaluate and compare the accuracy of these parsers.

3 Data selection

We aim to create the UD annotated treebank consisting of different syntactic types of sentences to implement and evaluate syntactic parsers. Dataset without much noise and covering widely acceptable sentence structures in the modern written Tamil would be beneficial for such tasks. Though we initially put efforts in compiling datasets from real-time occurrences from news-papers, blogs and online platforms, they did not comprehensively cover all grammatical constructions which can be used as a representative dataset to implement and evaluate syntactic parsers. Hence, we have chosen sentences from Lehmann's *A Grammar of Modern Tamil* (Lehmann, 1989) to start with, as they cover different linguistic structures with exceptions and these sentences represent written Tamil which are even today widely accepted.

4 Methodology

The annotations of the treebank consist of POS, lemma, morphological, and dependency information in accordance with UD. We have used a step by step process given below to annotate the treebank:

1. The dataset is pre-processed and tokenised.
2. Multi-word tokens are identified and expanded. The multi-word expansion is an essential feature in UD, using which tokens are divided into multiple syntactic units; this is discussed in detail in Section 4.2.
3. The processed text are POS tagged using *Thamizhi*POSt (Sarveswaran and Dias, 2021).

4. Morphological tags are added to each token using Apertium Tamil morphological analyser (Parameswari, 2011).

5. Dependency relations are marked manually on the top of POS and morphological information.

4.1 Preprossessing and Tokenisation

We chose 534 simple sentences with mostly one clause as the first step. As the next step, we plan to extend the treebank by including complex sentences that comprise embedded clauses. We extracted the sentences and cleaned them manually. Unicode normalisation is done using the script we developed.[3] The sentences were then tokenised to separate symbols and special characters and converted to CoNLL-U annotation format.

4.2 Multi-word Token Expansion

In UD, the basic unit of annotation is syntactic words, not phonological or orthographic words.[4] When language is morphologically rich, it tends to add multiple grammatical pieces of information within a word that are morpho-syntactically relevant. There are instances where languages do add syntactic elements such as clitics, conjunctions, particles, compound verbs *etc.* within a word, which need to be split and provided with the token status for further processing. In MWTT, multi-word tokens are identified and given token status. For instance, Figure 1 shows the clitic *-um*, which expresses coordinating conjunction and it is identified as multi-word. Similarly, Figure 2 explicates the clitic *-ō* that functions as a complementiser; hence it is tokenised for its syntactic relation though morphologically manifested.

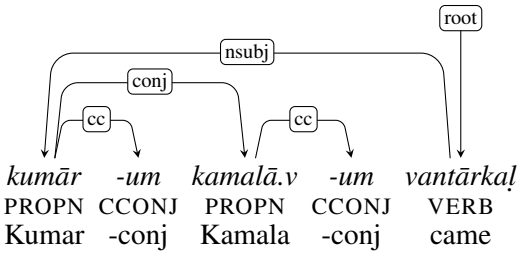


Figure 1: 'Kumar and Kamala came'

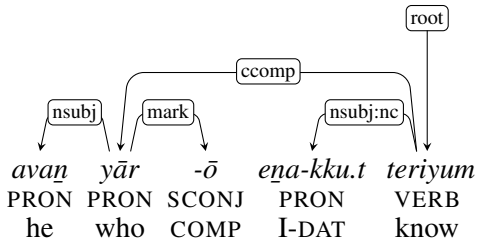


Figure 2: 'I know who is he'

Similarly, in compound verb (Verb (V1) + Verb (V2)) constructions, when V2s express aspect, mood, passive, causation and polarity, they are identified as multi-word tokens as in Figure 7. However, it is not split when the V2 is semantically bleached for its meaning and functions as an explicative verb. In the compound verb *ōṭip-pō* run.PART-go 'elope', the V2 *-pō* '*lit.* go' has partially lost its original meaning; hence it has not been split.

There are 43 multi-word tokens found in our treebank. On average, one multi-word token consists of 2.12 syntactic words.

4.3 POS Tagging

We used *Thamizhi*POSt (Sarveswaran and Dias, 2020), a contextual POS tagger to tag data with POS information. This POS tagger is trained on the text, which is not multi-word expanded. Therefore, most of the multi-word tokenised elements *i.e.* clitics, particles, compound verbs *etc* were not annotated correctly. Therefore, after the POS tagger output, we manually reviewed the POS tags.

Our treebank uses 14 POS tags out of 17 tags given in the UD POS scheme, see Table 1 for the POS tag inventory.

Table 1: POS tag frequencies of MWTT

ADJ	36	ADP	70	ADV	161	AUX	86	CCONJ	10	DET	57	NOUN	534
NUM	105	PART	2	PRON	171	PROPN	315	PUNCT	534	VERB	512	SCONJ	2

[3] https://github.com/sarves/thamizhi-preprocessor

[4] https://universaldependencies.org/u/overview/tokenization.html

Since the present version of MWTT consists only the simple sentences, the treebank does not contain any interjections, symbols and unknowns; therefore, INTJ, SYM and X were not utilised. PUNCT, NOUN, VERB and PROPN are the most frequent POS tags found with the frequency of 534(21%), 524(21%), 512(20%), and 315(12%), respectively.

4.4 Morphological Analysis

Tamil is known for its agglutinating morphology, where words are loaded with rich linguistic information. Words are morphologically analysed using Apertium Tamil Morphological Analyser (Parameswari, 2011), and then the output is mapped to UD features.[5] Nouns and pronouns are mainly analysed for their gender, number, person, case, politeness and rationality features. Adjectives are looked over for their gender, number and person details. Verbs and auxiliaries are analysed for gender, number, person, tense, aspect, mood, polarity, voice and verb form. However, since the morphological analysis is not contextual in nature, we have also reviewed the annotations.

4.5 Dependency Annotation

We annotated dependency relations according to UD schema. Around 22 relations are utilized to annotate the simple sentences out of 37 relations that are documented in UD. Apart from these main relations, although the treebank covers a simple and very limited number of syntactic constructions, it uses 17 sub-relation types that provide language-specific syntactic information separating by a colon (:) (see Table 2 for the top 5 sub-relations). All these annotations were done manually. Some of these syntactic analyses require in-depth linguistic inquiry. We have given an initial account of these constructions and issues in Section 5. This dataset is evaluated with Tamil UDPipe parser[6] and *Thamizhi*UDp. While UDPipe which is trained on Tamil_TTB treebank provides a Labeled Attachment Score (LAS) of 27.19, *Thamizhi*UDp which is trained using news data provides a LAS of 83.31.

5 Dependency Relations in Tamil

This section presents the discussion on predicates, subjects, oblique relations, compounds, coordinations, and other Tamil specific relations.

5.1 Predicates in Tamil

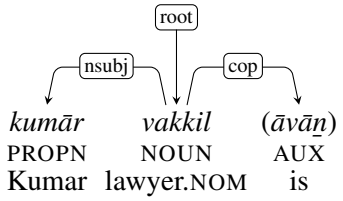
Sentences in contemporary written Tamil is constructed most commonly with verbal predicates. Complex predicates in Tamil consist of (i) verb+verb construction where a series of verbs can be added periphrastically to the first verb which is either in the form of verbal participle or infinitive forms to express aspect, mood, passive, causation, negative polarity and attitude (see Figure 7), (ii) noun+verb construction where a noun functions as the head and a verb as the light verb (see Figure 12 (cf. (Lehmann, 1989; Rajendran, 2004; Sarveswaran and Butt, 2019)).

However, in copular constructions, nominal predicates are identified as `root`. The copula verb *āku* '*lit.* to become' optionally occurs in nominal predicates. Figure 3 is an example of nominal predicate where the nominative case marked noun is identified as `root` and the copula verb is attached to it with the relation `cop`. The negative copula verb *illai* appears obligatorily to express constitution negation as in Figure 4. When the verb *illai* 'not' occurs as an existential negation, it is considered as *root* as seen in Figure 5. There are instances where the nominal predicate is dative-case marked to express benefaction as in Figure 6 and the copula is absent.

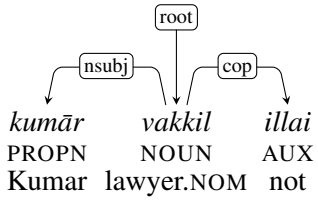
[5] https://universaldependencies.org/u/feat/index.html
[6] http://lindat.mff.cuni.cz/services/udpipe/

Table 2: Sub-relations or Language-specific relations used in MWTT

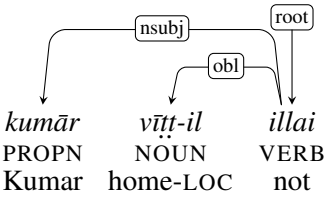
Relation	Description	Count
`nsubj:nc`	Non-canonical nominal subject	47
`obl:tmod`	Temporal modifier – oblique	45
`nmod:poss`	Nominal modifier – possessive	28
`compound:lvc`	Light verb	18
`obl:lmod`	Locative modifier – oblique	16



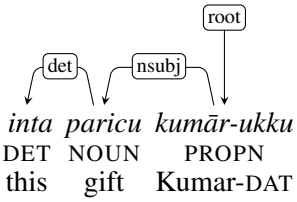
'Kumar is a lawyer'
Figure 3: Nominal Predicate



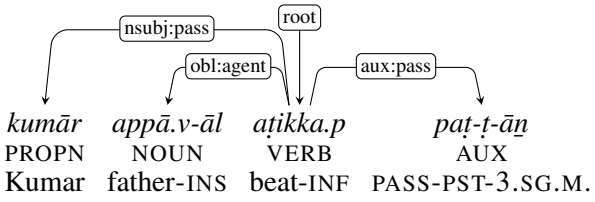
'Kumar is not a lawyer'
Figure 4: *illai* as Copula



'Kumar is not at home'
Figure 5: *illai* as Predicate



'This gift is for Kumar'
Figure 6: Dative as Predicate



'Kumar was beaten by (his) father'
Figure 7: Subject in Passive Construction

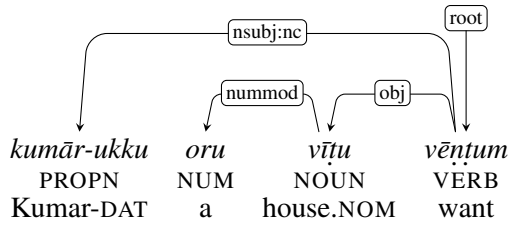
5.2 Subjects in Tamil

In Tamil, most commonly, a nominative case marked noun phrase functions as a subject and controls verb agreement. However, in the passive construction, the nominal subject with the nominative case marker (not a proto-agent) is identified as *nsubj:pass* which controls the verb agreement and the agent as *obl:agent* (see Figure 7).

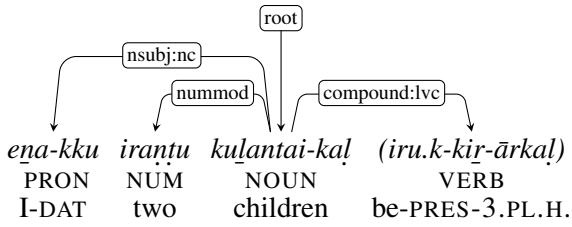
Subjects in Tamil are also realised with non-nominative markers such as the dative, the instrumental and the locative case markers. There are several discussions and diagnostic tests of Non-Nominative Subjects (NNS) (Siguresson, 2004; Subbārāo, 2012), which include NNSs that can occur as antecedents to anaphors and NNSs as controllers of PRO.

In NNS, the dative subject construction is the most widespread in Dravidian languages (Subbārāo, 2012) and are called experiencer subjects. Verma and Mohanan (1990) describes "in the so-called experiencer subject constructions in South Asian languages, the thematically prominent argument, which we expect to be a grammatical subject, is quite often an experiencer, and is marked with the case otherwise associated with indirect objects". In Tamil, stative predicates expressing the notion of mental, emotional and physical experience require the case-marking pattern of DAT-ACC (Lehmann, 1989; Pappuswamy, 2005). The tag *nsubj:nc* is used to mark non-canonical subjects. Dative subjects can also occur to express need or necessity, but the object is not explicitly marked for the accusative marker in Tamil as in Figure 8.

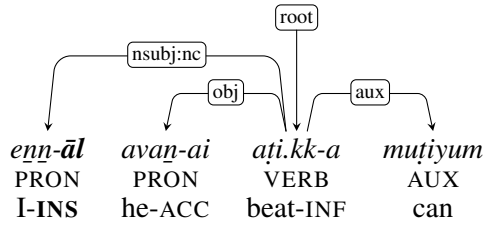
While expressing inalienable possession and kinship, the subject is marked with the dative in Tamil. The verb *iru* 'to be' is used as a possessive verb ('to have') and occurs optionally. In such constructions, the possessed noun is marked as `root` and the verb *iru* as a light verb i.e. `compound:lvc` as seen in Figure 9. The subject is marked with the locative case marker to show the temporary or alienable possession, and the existential verb *iru* 'to be' is used as the possessive verb (see Figure 11). When the predicate expresses capabilitative mood in Tamil, the subject is marked for the instrumental case marker, see Figure 10. It is also seen that the theme is marked for the accusative though the subject is NNS.



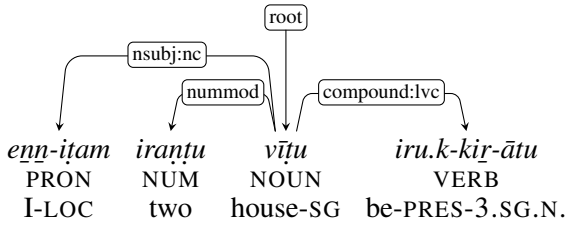
'Kumar wants a house'
Figure 8: Dative Subject-1



'I have two children'
Figure 9: Dative Subject-2



'I can beat him'
Figure 10: Instrumental Subject



'I have two houses'
Figure 11: Locative Subject

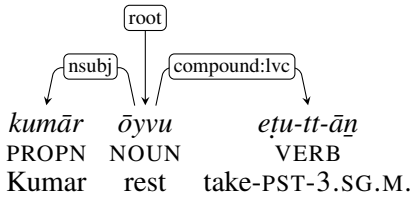
5.3 Oblique Cases

Non-core arguments are grouped under oblique cases in UD (Nivre et al., 2016). Language-specific oblique tags to differentiate locative (`obl:loc`), instrumental (`obl:inst`), ablative (`obl:abl`), sociative (`obl:soc`), place (`obl:pmod`) and temporal modifiers (`obl:tmod`), comparatives (`obl:cmpr`), agents in passive (`obl:agent`) are introduced in our treebank.

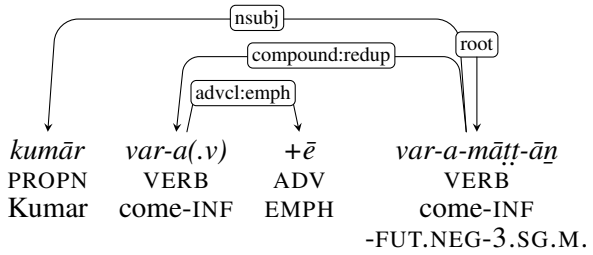
5.4 Compound

The relation `compound` is majorly marked for noun-noun compounds in UD. In Tamil, a language-specific tag `compound:lvc` is adopted for light-verb constructions where the noun occurs in juxtaposition to verb and carries the semantic content. The same tag is also used in Telugu treebank (Rama and Vajjala, 2018), however in our case, the noun is marked as `root`, and the light-verb is marked as `compound:lvc` as in Figure 12 following the practice in UD standard, whereas they are seen in the reverse direction in Telugu.

Reduplication and echo-word formation are other linguistic processes that are commonly found in many South-Asian languages (Subbārāo, 2012). They provide emphasis or distributive meaning. In Tamil, verbs (see Figure 13), nouns, determiners, adjectives and adverbs can be reduplicated. They are marked with the relation `compound:redup`.



'Kumar took rest'
Figure 12: `compound:lvc`



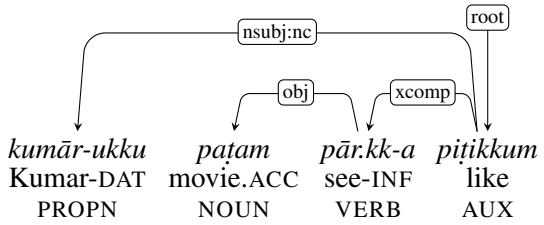
'Kumar won't come (with emphasis)'
Figure 13: `compound:redup`

5.5 Coordination

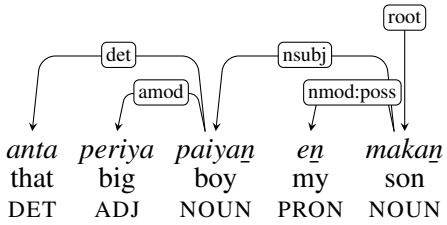
In Tamil, most commonly the clitic *-um* 'and' (see Figure 1), *-ō* 'or', *-āvatu* 'either.. or' are added as suffixes to each word or phrase or clause which are coordinated. The free morphemes *maṛṛum* 'and', *allatu* 'or', and *āṇāl* 'but' are also used, similar to English coordinators. The relation `conj` is used to conjoin them with the head-first approach in compliance with UD guidelines.

5.6 Other Relations

An open clausal complement (`xcomp` in UD) is found in Tamil as explicated in Figure 14, where the subject in the higher clause behaves as a subject to the subordinate predicate *pār* 'to see'. The relation `nmod:poss` is marked on possessive noun which is either realised in genitive case marker or oblique stem as in Figure 15. Auxiliaries are distinguished with relations `aux`, `aux:neg` and `aux:pass` in Tamil as negative and passive information are encoded as auxiliaries. Similarly, other relations such as `acl:relcl` is marked for relative clauses, `advcl` is marked for adverbial clauses and `advcl:cond` is for conditional adverbial clauses. Modifier relations such as `advmod`, `nummod`, `nmod` and `amod` are utilised in the treebank. To capture the emphasis to the meaning of any constituent, either the emphatic clitic *-ē* as a bound morpheme or the free morpheme *tāṉ* is used and identified as `advmod:emph`.



'Kumar likes to see movies'
Figure 14: The relation `xcomp`



'That big boy is my son'
Figure 15: The relation *nmod:poss*

6 Conclusion

In this paper, we have reported the creation of a Modern Written Tamil Treebank (MWTT) according to the Universal Dependencies framework.[7] We followed a hybrid approach and used an existing POS tagger and a morphological analyser to reduce manual annotation. We have also highlighted different syntactic constructions of simple sentences found in our corpus and how those are captured using the Universal Dependencies formalism. This treebank is useful as a benchmark dataset to evaluate syntactic parsers and other NLP tools such as POS taggers and morphological analysers.

As part of the future work, we will extend the resource by adding other complex syntactic constructions found in Lehmann's grammar book and other Tamil grammar books that we can access. Further, the Enhanced Universal Dependencies (EUD) scheme will also be incorporated to capture deep syntactic information.

Acknowledgements

We want to thank unknown reviewers for the valuable input, which shaped the final paper well. We extend our thanks to Keerthama B for involving in the annotation process. Further, we would like to thank Dan Zeaman for his continuous technical support in publishing MWTT.

References

Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma, and Lakshmi Bai. 2006. Anncorra: Annotating corpora guidelines for POS and Chunk annotation for Indian languages. *LTRC-TR31*, pages 1–38.

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga, and Ranjiva Munasinghe. 2019. Indic language computing. *Communications of the ACM*, 62(11):70–75.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2001. The Prague Dependency Treebank: Three-Level Annotation Scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Thomas Lehmann. 1989. *A Grammar of Modern Tamil*. Pondicherry Institute of Linguistics and Culture, India.

[7] https://github.com/UniversalDependencies/UD_Tamil-MWTT/tree/master

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, and Natalia Silveira. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association.

Umarani Pappuswamy. 2005. Dative Subjects in Tamil: A Computational Analysis. *South Asian Language Review*, XV(2):40–62.

K Parameswari. 2011. An implementation of APERTIUM morphological analyzer and generator for Tamil. *Parsing in Indian Languages*, pages 41–44.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

S Rajendran. 2004. Strategies in the formation of compound nouns in Tamil. *Languages of India*, 4.

Taraka Rama and Sowmya Vajjala. 2018. A Dependency Treebank for Telugu. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 119–128, Prague, Czech Republic. Association for Computational Linguistics.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague Dependency Style Treebank for Tamil. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1888–1894, Istanbul, Turkey. European Language Resources Association (ELRA) .

Kengatharaiyer Sarveswaran and Miriam Butt. 2019. Computational Challenges with Tamil Complex Predicates. In Miriam Butt, Tracy Holloway King, and Ida Toivonen, editors, *Proceedings of the LFG19 Conference, Australian National University*, pages 272–292, Stanford. CSLI Publications.

Kengatharaiyer Sarveswaran and Gihan Dias. 2020. ThamizhiUDp: A Dependency Parser for Tamil. In *Proceedings of the 17th International Conference on Natural Language Processing*, pages 200–207, Indian Institute of Technology Patna, India. NLP Association of India.

Kengatharaiyer Sarveswaran and Gihan Dias. 2021. Building a Part of Speech tagger for the Tamil Language. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, Singapore. IEEE.

HA Sigurdsson. 2004. Icelandic non-nominative subjects. In Bhaskararao, P. and Subbarao, K.V., editor, *Typological Studies in Language*, chapter 7, pages 137–159. John Benjamins Publishing Company.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Kārumūri V Subbārāo. 2012. *South Asian Languages: A Syntactic Typology*. Cambridge University Press.

Manindra K Verma and KP Mohanan. 1990. Introduction to the Experiencer Subject Construction. In Manindra K Verma and KP Mohanan, editors, *Experiencer subjects in South Asian Languages*, chapter 1, pages 1–12. Center for the Study of Language and Information (CSLI), Stanford, CA.

How Universal is Genre in Universal Dependencies?

Max Müller-Eberstein and **Rob van der Goot** and **Barbara Plank**
Department of Computer Science
IT University of Copenhagen, Denmark
mamy@itu.dk, robv@itu.dk, bapl@itu.dk

Abstract

This work provides the first in-depth analysis of genre in Universal Dependencies (UD). In contrast to prior work on genre identification which uses small sets of well-defined labels in mono-/bilingual setups, UD contains 18 genres with varying degrees of specificity spread across 114 languages. As most treebanks are labeled with multiple genres while lacking annotations about which instances belong to which genre, we propose four methods for predicting instance-level genre using weak supervision from treebank metadata. The proposed methods recover instance-level genre better than competitive baselines as measured on a subset of UD with labeled instances and adhere better to the global expected distribution. Our analysis sheds light on prior work using UD genre metadata for treebank selection, finding that metadata alone are a noisy signal and must be disentangled within treebanks before it can be universally applied.

1 Introduction

Identifying document genre automatically has long been of interest to the NLP community due to its immediate applications both in document grouping (Petrenz, 2012) as well as task-specific data selection (Ruder and Plank, 2017; Sato et al., 2017).

Cross-lingual genre identification has however remained a challenge, mainly due to the lack of stable cross-lingual representations (Petrenz, 2012). Recent work has shown that pre-trained masked language models (MLMs) capture monolingual genre (Aharoni and Goldberg, 2020). Do such distinctions manifest in highly multilingual spaces as well? In this work, we investigate whether this property holds for the genre distribution in the 114 language Universal Dependencies corpus (UD version 2.8; Zeman et al., 2021) using the multilingual mBERT MLM (Devlin et al., 2019).

In absence of an exact definition of textual genre (Kessler et al., 1997; Webber, 2009; Plank, 2016), this work will focus on the information specifically denoted by the `genres` metadata tag in UD. We hope that an in-depth, cross-lingual analysis of what this label represents will enable practitioners to better control for the effects of domain shift in their experiments. Previous work using these UD metadata for proxy training data selection have produced mixed results (Stymne, 2020). We investigate possible reasons and identify inconsistencies in genre annotation. The fact that genre labels are only available at the level of treebanks makes it difficult to gather a clear picture of the *sentence-level* genre distribution — especially with some treebanks having up to 10 genre labels. We therefore investigate the degree to which instance-level genre is recoverable using only the treebank-level metadata as weak supervision.

Our contributions entail the, to our knowledge, first detailed definition of all UD metadata genre labels (Section 3), four weakly supervised methods for extracting instance-level genre across 114 languages (Section 4) as well as genre identification experiments which show that our proposed two-step procedure allows for effective genre recovery in multilingual setups where language relatedness typically outweighs genre similarities (Section 5).[1]

[1]Code available at `https://personads.me/x/syntaxfest-2021-code`.

2 Related Work

The largest hurdle for cross-lingual genre classification is the lack of shared representational spaces. Sharoff (2007) use shared POS n-grams in order to jointly classify the genre of English and Russian documents. Petrenz (2012) similarly seek out features which are stable across languages in order to classify English and Chinese documents into four shared genres. A recent data-driven approach finds that monolingual MLM embeddings can be clustered into five groups closely representing the data sources of the original corpus (Aharoni and Goldberg, 2020). In this work, we investigate whether this holds for multilingual settings as well.

Being able to identify textual genre has been crucial for domain-specific fine-tuning (Dai et al., 2020; Gururangan et al., 2020) including dependency parsing. For parser training, in-genre data is typically selected by proxy of the data source (Plank and van Noord, 2011; Rehbein and Bildhauer, 2017; Sato et al., 2017). Data-driven approaches which include automatically inferred topics based on word and embedding distributions (Ruder and Plank, 2017) as well as POS-based approaches (Søgaard, 2011; Rosa, 2015; Vania et al., 2019) have also been found effective.

Universal Dependencies (Nivre et al., 2020) aims to consolidate syntactic annotations for a wide variety of languages and genres under a single scheme. The latest release contains 114 languages — many with fewer than 100 sentences. In order for languages at all resource levels to benefit from domain adaptation, it will continue to be important to identify cross-lingually stable signals for genre. While language labels are generally agreed upon, differences in genre are more subtle. Metadata at the treebank level provides some insights into genres of original data sources, however these are "neither mutually exclusive nor based on homogeneous criteria, but [are] currently the best documentation that can be obtained" (Nivre et al., 2020).

Stymne (2020) performs an initial study on using these treebank metadata labels for the selection of spoken and Twitter data. Results show that training on out-of-language/in-genre data is superior to out-of-language/out-of-genre data. However the best results are obtained using in-language data regardless of genre-adherence. This holds across multiple methods of proxy dataset selection (e.g. treebank embeddings; Smith et al., 2018).

Recently, Müller-Eberstein et al. (2021) have shown that combining UD genre metadata and MLM embeddings can improve proxy training data selection for zero-shot parsing of low-resource languages. The use of genre in their work is more implicit as it is mainly driven by the genre of the target data. In contrast, this work takes a holistic view and explicitly examines the classification of instance-level genre for all sentences in UD.

As genre appears to be a valuable signal, we set out to investigate how it is defined and distributed within UD. Due to the coarse, treebank-level nature of current genre annotations, we hypothesize that a clearer picture can only be obtained by moving to the sentence level. We therefore transition from prior supervised document genre prediction to weakly supervised *instance* genre prediction. Additionally, we expand the linguistic scope from mono- or bilingual corpora to all 114 languages currently in UD.

More generally, this task can be viewed as predicting genre labels for all sentences in all corpora of a collection while only being given the set of labels said to be contained in each corpus.

3 UD-level Genre

We analyze genre as currently used in the `genres` metadata of 200 treebanks from Universal Dependencies version 2.8 (Zeman et al., 2021). Section 3.1 provides an overview of all UD genre types and Section 3.2 analyzes how these global labels relate to the subset of treebanks which do provide treebank-specific, instance genre annotations.

3.1 Available Metadata

UD 2.8 (Zeman et al., 2021) contains 18 genres which are denoted in each treebank's accompanying metadata. Around 36% of treebanks contain a single genre while the remaining majority can contain between 2–10 which are not further labeled at the instance level. There is no official description of each genre label, however they can be roughly categorized as follows:

academic Collections of scientific articles covering multiple disciplines. Note that this label may subsume others such as *medical*.

bible Passages from the bible, frequently from older languages (e.g. Old Church Slavonic-PROIEL by Haug and Jøhndal, 2008). Largely non-overlapping passages are used across treebanks.

blog Internet documents on various topics which may overlap with other genres such as *news*. They are typically more informal in register. Some treebanks group social media content and reviews under this category (e.g. Russian-Taiga by Shavrina and Shapovalova, 2017).

email Formal, written communication. This includes English-EWT's (Silveira et al., 2014) subsection based on the Enronsent Corpus (Styler, 2011) as well as letters attributed to Dante Alighieri as part of Latin-UDante (Cecchini et al., 2020).

fiction Mostly paragraphs from diverse sets of fiction books and magazines.

government The least represented genre, mainly denoting texts from governmental sources. These include political speeches (English-GUM by Zeldes, 2017) as well as inscriptions from Neo-Assyrian kings from around 900 BCE (Akkadian-RIAO by Luukko et al., 2020).

grammar-examples Sentences from teaching or grammatical reference books which are typically short, but cover a wide range of dependency relations (e.g. Tagalog-TRG by Samson and Cöltekin, 2020).

learner-essays Small genre occurring in three single-genre treebanks. Sentences were written by second-language learners and either contain original errors (English-ESL by Berzak et al., 2016), manual corrections (IT-Valico by Di Nuovo et al., 2019) or both (Chinese-CFL by Lee et al., 2017).

legal Relatively frequent genre based mostly on laws and legal corpora within the public domain.

medical Scientific articles/books in the field of medicine (e.g. cardiology, diabetes, endocrinology for Romanian-SiMoNERo by Mitrofan et al., 2019). It is subsumed by *academic* for some treebanks (e.g. Czech-CAC by Hladká et al., 2008).

news The highest-resource genre by a large margin corresponding to news-wire texts as well as online newspapers on specific topics (e.g. IT-news in German-HDT by Borges Völker et al., 2019).

nonfiction Second most frequent genre with a high degree of variance, subsuming e.g. *academic* and *legal*. German-LIT (Salomoni, 2019) contains three philosophical books from the 18th century. Other *non-fiction* treebanks can originate from multiple sources (e.g. books and internet) and time spans.

poetry Smaller, yet distinct genre covering mostly older texts and language variations (e.g. Old French-SRCMF by Stein and Prévost, 2013).

reviews Medium-resource genre covering informal online reviews with unnormalized orthography (e.g. English-EWT) as well as formal reviews (e.g. newspaper film reviews in Czech-CAC).

social Encompasses social media data such as tweets (e.g. Italian-TWITTIRÒ by Cignarella et al., 2019) as well as newsgroups (e.g. English-EWT). Some *spoken* data is co-labeled with this genre when it refers to colloquial speech (e.g. South Levantine Arabic-MADAR by Zahra, 2020).

spoken Distinct genre which typically consists of spoken language transcriptions. Sentences contain filler words and may have abrupt boundaries. Sources range from elicited speech of native speakers (Komi Zyrian-IKDP by Partanen et al., 2018) to radio program transcriptions (Frisian Dutch-Fame by Braggaar and van der Goot, 2021).

web Similarly ambiguous genre as *non-fiction*. It occurs in conjunction with specific genres such as *blog* and *social* and never appears alone (e.g. Persian-PerDT by Sadegh Rasooli et al., 2020).

wiki Denotes data from Wikipedia for which cross-lingual authoring guidelines exist.

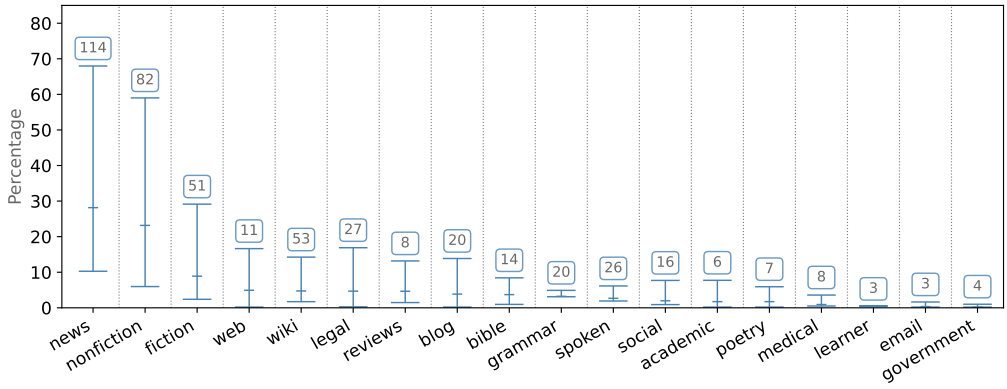


Figure 1: **Genre Distribution in UD Version 2.8.** Ranges indicate upper/lower bounds for sentences per genre inferred from UD metadata. Center marker reflects the distribution under the assumption that genres within treebanks are uniformly distributed. Labels above the bars indicate the number of treebanks which contain each genre.

Figure 1 shows the approximated distribution of these genres in UD. Maximum/minimum sentence counts are inferred from the size of single-genre treebanks plus the size of all treebanks in which a genre is said to occur. The center line denotes the distribution under the assumption that genres are uniformly distributed within each treebank.

It is clear that *news* and *non-fiction* constitute more than half of the entire dataset. Specialized genres such as *medical* are less represented. For broader genres such as *web*, which frequently co-occurs with others, the exact number of sentences is hard to estimate, but must lie between 0–20%. Considering these large variances, access to instance-level genre will likely be crucial for effective proxy data selection and downstream domain adaptation.

3.2 Instance-level Annotations

In addition to the aforementioned 18 treebank-level genre labels, some treebanks provide instance-level genre annotations in the comment-metadata before each sentence. We find such annotations in 26 out of 200 treebanks in UD 2.8 amounting to 124k or 8.25% of all sentences.

Out of this set, 20 treebanks belong to the Parallel Universal Dependencies (PUD; Nivre et al., 2017). They are split 500/500 between *news* and *wiki*, as denoted by sentence IDs beginning with `n` and `w` respectively. The parallel nature of PUD makes it interesting for analyzing cross-lingual genre identification performance. However these two genres only represent a small fraction of non-fiction texts and furthermore, each PUD-treebank is test-split-only. Note also that Polish-PUD as an exception has the metadata labels *news* and *non-fiction*.

The remaining six treebanks for which we were able to identify instance-level genre annotations are Belarusian-HSE (Lyashevskaya et al., 2017), Czech-CAC (Hladká et al., 2008), English-EWT (Silveira et al., 2014), German-LIT (Salomoni, 2019), Polish-LFG (Patejuk and Przepiórkowski, 2018) and Russian-Taiga (Shavrina and Shapovalova, 2017). They cover a wider set of 12 genres. Annotation schema vary across treebanks and are neither fully compatible amongst each other nor with the 18 UD labels. Approximate mappings can however be drawn thanks to source data documentation by the respective authors (Section 4.2).

Further comment-metadata which may guide genre separation within treebanks includes document, paragraph and source identifiers. Again, these are unfortunately not available for all sentences (although coverage of these metadata reaches up to 45%) and their values do not provide further indications about genre adherence.

4 Instance Genre from Treebank Labels

From the previous analysis, it is evident that finer-grained genre labels are needed before domain adaptation can be successful across all languages.

Formally, the task of predicting instance-level UD genre can be defined as assigning a set of labels $\mathcal{L} = \{l_0, l_1, \ldots, l_K\}$ (i.e. genres) to all instances x_n of a corpus $\mathcal{X}$ (i.e. UD). The corpus consists of S distinct subsets $\mathcal{X} = \{\mathcal{X}_0 \cup \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_S\}$ (i.e. treebanks) each with a subset of labels $\mathcal{L}_s \subseteq \mathcal{L}$. As no instance-level labels $x_n \rightarrow l$ are available, models must learn this mapping based solely on the subset of labels said to be contained in each data subset $\mathcal{X}_s \rightarrow \mathcal{L}_s$.

4.1 Genre Prediction Methods

As instance-level labels are noisy and sparse, we investigate two classification-based and two clustering-based approaches for inferring instance genre labels from the treebank metadata $\mathcal{L}_s$ alone. Building on Müller-Eberstein et al. (2021), our proposed methods leverage latent genre information in the pre-trained mBERT language model (Devlin et al., 2019).

BOOT In order to select proxy training data which matches the genre of an unseen target, Müller-Eberstein et al. (2021) propose a bootstrapping-based approach to genre classification (BOOT). An mBERT-based classifier (Devlin et al., 2019) is initially trained on sentences from single-genre treebanks, corresponding to standard supervised classification. Above a confidence threshold (i.e. softmax probability of 0.99), sentences from treebanks containing a known genre in mixture are bootstrapped as single-genre training data for the next round. After bootstrapping sentences from all known genres, the remaining unclassified instances of any treebank containing a single unknown genre are inferred to be of that last genre. While this method was previously used for targeted data selection, we investigate the degree to which it actually recovers instance-level genre.

CLASS With approximate classification (CLASS), we simplify BOOT to naively learn instance genre labels from weak supervision. It fine-tunes the same mBERT MLM with a 18-genre classification layer on the `[CLS]`-token. For single-genre treebanks it is possible to measure the exact cross entropy between the predicted probability and the target (i.e. $x_n \rightarrow l$ with $l \in \mathcal{L}_s$ and $|\mathcal{L}_s| = 1$). For multi-genre treebanks with $|\mathcal{L}_s| > 1$, this is not possible as the gold label is unknown. For the CLASS approach, each sentence from a k-genre treebank is therefore classified k times — once for each class in $\mathcal{L}_s$.

GMM In addition to classification, we also evaluate two common clustering algorithms. First we investigate whether clusters formed by untuned MLM sentence embeddings (mean over sentence subwords) represent genre to such a degree that Gaussian Mixture Models can recover the 18 UD genre groups. For monolingual data from five genres, such clusters were shown to be recoverable (Aharoni and Goldberg, 2020). We extend this approach to the 114 language setting of UD.

LDA As all methods so far are to some degree dependent on the pre-trained MLM representations, we also evaluate the recoverability of genre using Latent Dirichlet Allocation (Blei et al., 2003) with lexical features. Feature vectors are constructed using the frequency of character 3–6-grams.

Cluster Labeling Both clustering methods produce 18 groups of sentences from UD, however these will not carry meaningful labels as with classification. While labels could be assigned manually post-hoc by matching representative sentences in each cluster to one of the 18 global UD genres, this process is bound to be subjective and also depends on the annotator to be fluent in most of the 114 languages.

In order to automate this procedure, we propose **GMM+L** and **LDA+L** which combine clustering and classification. Both methods start by clustering each treebank $\mathcal{X}_s$ into the number of genres specified by its metadata (note that standard GMM and LDA cluster all of UD at once, i.e. $\mathcal{X}$).

Next, the mean embedding of each cluster is computed such that they can be compared in a single representational space. Note that this would not be possible using monolingual models as their latent spaces are not as cross-lingually aligned. Analogous to BOOT, single-genre treebanks can then be used as a single-label signal such that the closest cluster from each treebank containing the respective genre

can be extracted. Newly identified clusters are added to the pool of single-genre clusters. This process need only be repeated for three rounds before all sentences in UD can be assigned a single label.

Using these four methods, we aim to assign a single genre label to each sentence in UD. By comparing model ablations, we further depart from prior work and explicitly quantify the genre information in MLM embeddings as well as how it manifests within and across treebanks in UD.

4.2 Supervised Evaluation

For the 26 treebanks with instance genre labels, we are able to measure standard F1 after applying a mapping from the treebank-specific labels to the 18 global UD genre labels. The mapping was created according to the following criteria.

First, we only allowed treebank-specific genre labels to be mapped to the set of UD genre labels specified in each treebank's metadata.

Second, if possible treebank labels are mapped to UD labels of the same name (e.g. `fiction` $\rightarrow$ *fiction*) or to the closest subsuming category (e.g. `spoken (prepared)` $\rightarrow$ *spoken*).

Third, decisions involving subjective uncertainty were based on the label which covers the majority of data sources. E.g., Czech-CAC has the metadata label set {*legal, medical, news, non-fiction, reviews*} and only three types of instance labels (`aw`, `nw`, `sw`). The `sw` (scientific-written) label is attached to many medical articles, but also to articles on philosophy or music. While *academic* may be the most fitting label, it is not in the metadata. As such we chose the broader *non-fiction* as the target label.

The full mapping is in Appendix A and we hope future work will be able to expand upon it.

4.3 Unsupervised Evaluation

For the remaining 174 treebanks without sentence-level gold labels it is difficult to measure the exact quality of the predicted genre distributions. Nonetheless, treebank annotations provide enough information for approximate, global comparisons.

Based on label/cluster assignments, it is possible to compute the standard cluster purity measure (PUR; Schütze et al., 2008). Across treebanks of the same genre, the majority of sentences should belong to the same label/cluster. We measure this using the ratio of cross-treebank label agreement (AGR). As in prior work (Aharoni and Goldberg, 2020) it is important to note that the aforementioned metrics can be misleading when taken on their own: A perfect score can for example be achieved by simply assigning all instances to the same genre.

To mitigate this issue we turn to the expected overlap of inter-treebank genre distributions. For multi-genre treebanks, it is known which genres are present, but not how they are distributed. Since treebanks are expected to have a certain amount of overlap, we can however estimate a global error. A {*fiction, spoken, wiki*} treebank should for example have no clusters in common with a {*news*} treebank, but should have many sentences in the same clusters as a {*fiction, medical, spoken*} one. Assuming that genres are uniformly distributed within each treebank, the first pair would share 0 mass between distributions while the second pair would share $\frac{2}{3}$. Intuitively, a good prediction would produce a global genre distribution that falls precisely between the metadata range bars of Figure 1, close to the center markers.

To quantify the overlap between two treebank genre distributions p and q over the genres in $\mathcal{L}_s$, we use the discrete Bhattacharyya coefficient:

$$BC(p, q) = \sum_{l \in \mathcal{L}_s} \sqrt{p(l)q(l)} \tag{1}$$

which has often been applied to distributional comparisons (Choi and Lee, 2003; Ruder and Plank, 2017). It is computed for all pairs of treebanks such that the overlap error $\Delta\text{BC} \in [0, 100]$ is the mean absolute difference between the expected distributional overlap of each treebank pair and the predicted one (i.e. lower is better).

While none of these metrics can individually provide an exact measure of a prediction method's fit to the UD-specified distribution, they complement each other as to allow for global comparisons in absence of any sentence-level annotations.

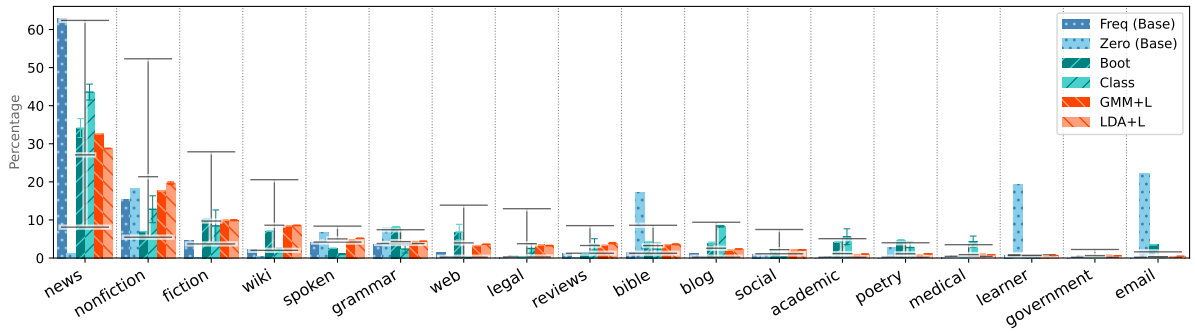


Figure 2: **Genre Predictions on UD (Test).** Ranges indicate upper/lower bounds inferred from UD metadata and the distribution under treebank-level uniformity at the center marker. Bars show averaged distribution predictions with standard deviations by FREQ, ZERO, BOOT, CLASS, GMM+L and LDA+L.

5 Experiments

5.1 Setup

Data From the 1.5 million sentences in UD, we construct global training, development and testing splits. All original test splits are left unchanged and gathered into one global test split containing 204k sentences. Note that test-only treebanks and languages are thereby never seen during training or tuning. For instance-level, supervised evaluation, this means that all PUD treebanks and German-LIT are excluded, leaving five treebanks for tuning.

Next, all original training and development splits are concatenated and split 10/90 into a global training and development split with 102k and 915k sentences respectively. The reason for this small "training" split is that it is only required for training CLASS and BOOT. Within it, we again split the data 70/30 (71k and 31k sentences) for classifier training and held-out data for early stopping. All exact splits are provided in Appendix A.

Baselines For our comparisons, we use a maximum frequency baseline (FREQ) which labels all sentences within a treebank with the metadata genre label that is most frequent overall. For example, in any treebank containing *news*, all instances are labeled as such.

In order to measure the untuned classification performance of mBERT, we propose an additional zero-shot classification baseline (ZERO). Prior research has found that classifying sentences based solely on their cosine similarity to genre label strings in MLM embedding space can be remarkably effective (Veeranna et al., 2016; Yin et al., 2019; Davison, 2020). For example, a sentence is labeled as *academic* if this is the closest embedded label out of all 18 genre strings.

Training Every method from Section 4.1 is run with three initializations. CLASS and BOOT are trained for a maximum of 30 epochs with an early stopping patience of 3. ZERO, GMM+L and LDA+L (by extension GMM, LDA) do not require training and can be directly applied to the target data. Implementation details and development results are reported in Appendices B and C.

5.2 Results

Using the 8% subset of annotated instances (Section 4.2) in addition to the unsupervised metrics from Section 4.3, we can gather an estimate of each method's performance in Table 1. UD-level genre predictions in addition to instance-level confusions are further visualized in Figures 2 and 3.

Baselines The FREQ baseline highlights the issue of using individual unsupervised metrics for estimating performance. As it assigns all sentences per treebank to the same genre, it automatically achieves 100% single-genre treebank purity and agreement. Considering that the instance-level F1 covers 12 genres, a baseline score of 47 is also competitive. Note that this is mostly due to the data imbalance towards *news*. This unlikely distribution predicted by FREQ is also reflected in Figure 2.

METHOD	PUR	AGR	ΔBC	F1
FREQ	100±0.0	100±0.0	21±0.0	47±0.0
ZERO	46±0.0	56±0.0	47±0.0	12±0.0
CLASS	83±1.4	63±3.9	34±1.1	32±0.9
BOOT	86±0.4	70±0.7	29±0.3	38±1.2
GMM	90±0.5	45±2.6	31±0.3	—
+LABELS	100±0.0	100±0.0	4±0.2	54±2.1
LDA	77±0.8	34±2.6	31±0.2	—
+LABELS	100±0.0	100±0.0	2±0.1	51±1.5

Table 1: **Results of Genre Prediction on UD (Test).** Purity (PUR ↑), agreement (AGR ↑), overlap error (ΔBC ↓) and micro-F1 over instance-labeled TBs (F1 ↑) for FREQ, ZERO, CLASS, BOOT and GMM, LDA with/without cluster label predictions (+LABELS). Standard deviation denoted ±.

ZERO-shot classification is not fine-tuned on UD-specific signals and as such predicts a genre distribution that does not adhere to the metadata at all (see Figure 2). It severely underpredicts high-frequency genres such as *news* and overpredicts less frequent genres such as *email*. This reflects in our metrics, with ZERO obtaining the lowest PUR, AGR and F1 while having the highest ΔBC of 47.

Classification With regard to explicit genre fine-tuning, CLASS increases purity by 38 points compared to ZERO. Agreement across treebanks also improves, while overlap error decreases. These differences are also reflected in Figure 2 in that the predicted distribution is more within the range that would be expected given the metadata.

BOOT fits the UD genre distribution more closely, resulting in a purity that is 4 points higher and agreement that is 11 points higher than CLASS. F1 also increases by 6 points while overlap error decreases by 4 points, indicating that these improvements are not merely due to e.g. assigning all sentences to the same genre. While instance-level F1 is below the FREQ baseline, both methods improve upon the untuned ZERO by a factor of 3.

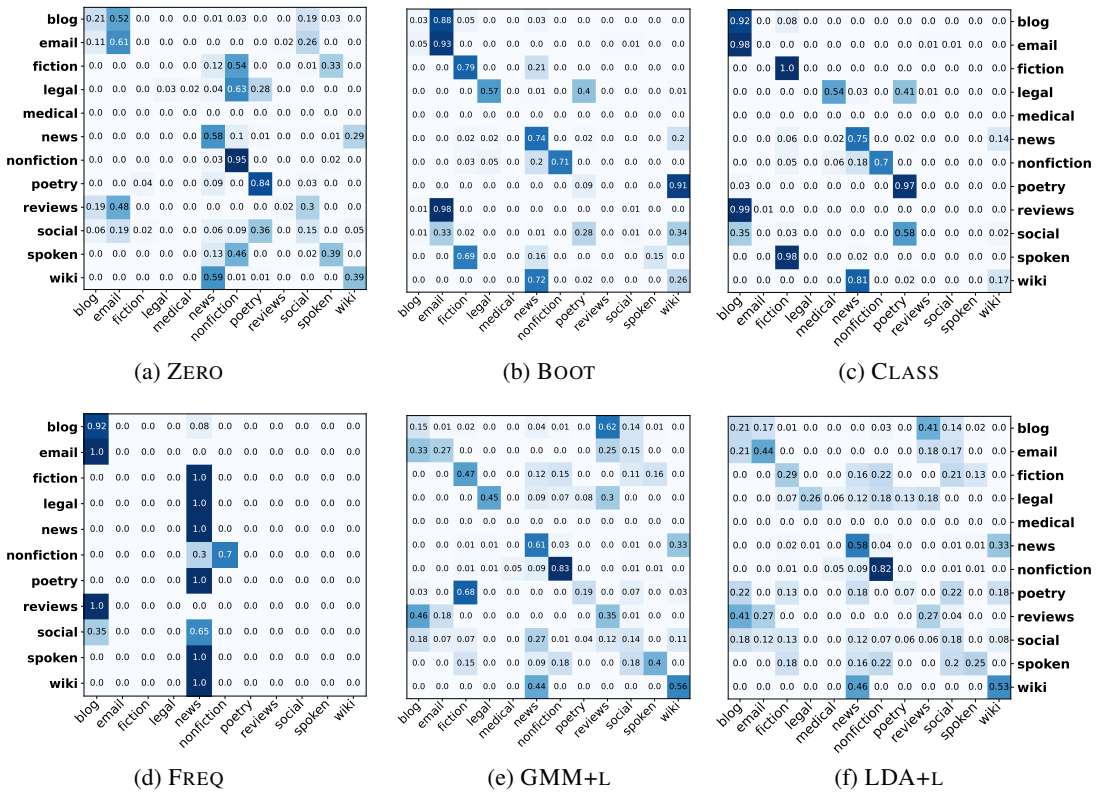
The benefits of the less noisy training signal are visible in Figure 2: Compared to CLASS, BOOT predicts labels in a way that more closely resembles the expected distribution even when the label only occurs in multi-genre treebanks and is ambiguous (e.g. *web*). While BOOT agrees upon the same genre-label across languages (e.g. all *social* treebanks are labeled as such), CLASS tends to overassign the globally most frequent labels (e.g. half of *social* treebanks are labeled *wiki*) and has a larger variance in its assignments across initializations.

Clustering GMM clusters from untuned mBERT embeddings follow the distribution specified by UD metadata more than the LDA clusters produced from lexical information. Although sentence representations are gathered using a naive mean-pooling approach, the resulting clusters reach 90% PUR compared to 77% for LDA. AGR follows a similar pattern and ΔBC is equivalent.

Turning to our cluster labelling approaches, both GMM+L and LDA+L obtain the highest overall F1 scores, outperforming both baselines. They achieve 100% PUR and AGR by the same process as the FREQ-baseline while their overlap error is significantly lower at 4 and 2 points respectively. Figure 2 reflects this, as GMM+L and LDA+L are always closest to the expected genre distribution, regardless of overall genre frequency. This shows how focusing on treebank-internal differences before applying a global labelling procedure combines the benefits of local clustering with the benefits of bootstrapped classification, resulting in an effective overall method.

5.3 Analysis

From the F1 scores in Table 1 it is clear that predicting instance genre based on treebank metadata alone — while accounting for its skewed distribution and inter-treebank shifts of genre definitions — is a difficult task. In the following we analyze the performance characteristics of each method.



(a) ZERO (b) BOOT (c) CLASS

(d) FREQ (e) GMM+L (f) LDA+L

Figure 3: **Confusions of Instance-level Genre.** Ratios of predicted labels (columns) per target (row) for ZERO, BOOT, CLASS, FREQ, GMM+L, LDA+L on test splits of 26 instance-annotated treebanks.

Overall, trends of the unsupervised metrics follow the supervised F1, leading us to believe that the methods would behave comparatively should labels for all instances in UD be available. The confusion matrices with prediction ratios per gold label in Figure 3 reflect our previous observations.

Baselines The FREQ baseline's predictions are clearly dominated by the most frequent *news* genre, followed by the similarly high frequency *non-fiction* and *blog* (see Figure 3d).

ZERO appears to follow a pattern similar to BOOT (e.g. *blog* and *email*), however it also makes more predictions away from the diagonal (see Figure 3a).

Classification Both CLASS (Figure 3c) and BOOT (Figure 3b) assign most instances of a genre to a single prediction label, often strongly aligning with the target diagonal. CLASS more often assigns a single label per target instead of spreading out predictions across multiple labels as in BOOT. Nonetheless, both methods make some unintuitive errors such as BOOT classifying parts of *poetry* as *wiki*. For these 68 samples from Russian-Taiga, BOOT likely overfits the language signal from Russian-GSD (McDonald et al., 2013; *wiki*).

Compared to ZERO which approximates the predictions of an untuned mBERT model, BOOT and CLASS fine-tuning appears to amplify existing patterns and shifts some predictions to better align with genres as defined in UD (e.g. *fiction* and *legal* in BOOT).

Clustering Grouping all 1.5 million sentences of UD into 18 unlabeled clusters using GMM and LDA results in purity and ΔBC comparable to CLASS and BOOT. However, looking into the cluster contents of the former reveals that they are oversaturated with large treebanks such as German-HDT. Cosine similarities of cluster centroids from the mBERT-based GMM further indicate that proximity corresponds foremost to language similarity.

Some clusters predominantly contain *news*, *wiki* or *social*. This corresponds to cases such as the Italian Twitter treebank TWITTIRÒ in which specific tokens (e.g. "@user") are distinct enough to override the

language signal. Overall, most UD-level clusters do not have clear genre distinctions and are influenced more strongly by language than genre, resulting in high treebank purity while having low intra-treebank agreement. Attempting to cross-lingually cluster all sentences in UD directly is therefore not as effective for recovering instance-level genre as it was in the monolingual setting (Aharoni and Goldberg, 2020).

Initially constructing clusters within each treebank as in the GMM+L and LDA+L methods appears to restore the benefits observed in the monolingual setting. A qualitative analysis of the treebank-level LDA clusters reveals that *wiki* clusters often contain lexical indicators for the genre, such as brackets, while *news* features often contain n-grams which may be related to spoken quotes such as "said", "Ik␣" (first person pronoun).

Attaching labels to these clusters using the globally shared mBERT space yields confusion plots for GMM+L and LDA+L which most closely follow the diagonal (see Figures 3e and 3f). Overall, their predictions follow a similar pattern indicating that clustering at the treebank-level using either mBERT embeddings or lexical features results in similar sentence groups.

Within the instance-labeled subset, all models share confusions between *news* and *wiki* (mainly from PUD). While *wiki* is often predicted as *news*, both GMM+L and LDA+L substantially improve upon this "*news*-bias" with a confusion ratio that is 13%–56% lower compared to all other methods. The sentence-bounded context from which all models must make their genre predictions nonetheless limits the amount of improvement possible. For example, using the aforementioned LDA features the algorithm would very likely be unable to distinguish between *news* and *wiki* (both non-fiction, edited texts describing facts) for cases such as, *"Weiss was honored with the literature prizes from the cities of Cologne and Bremen."*

6 Discussion and Conclusion

This work provided an in-depth analysis of the 18 genres in Universal Dependencies (UD) and identified challenges for projecting this treebank metadata to the instance level. As these genre labels were not part of the first UD releases, but were added in later versions, we identified large variations in the way they are interpreted and applied — resulting in far less universal definitions of genre than for syntactic dependencies. Most treebanks furthermore contain multiple genres while not providing finer-grained instance-level annotations thereof. This also sheds light on prior work which used UD metadata for training data selection, where treebank-level genre improved in-language parsing performance (Stymne, 2020) and where moving to instance-level genre signals lead to additional increases even across languages (Müller-Eberstein et al., 2021).

Building on the latent genre information stored in MLM embeddings, we investigated four methods for projecting treebank-level labels to the instance level. In contrast to prior monolingual work, immediately clustering multilingual embeddings yielded clusters dominated by language similarity instead of genre (Section 5.3). Similarly, zero-shot labelling using the untuned mBERT latent space proved to be insufficient for producing a genre distribution which adheres to the UD metadata. The classification-based CLASS and BOOT methods are able to extract a stronger genre signal from mBERT than ZERO.

Our proposed GMM+L and LDA+L methods which combine local treebank clusters with the global, cross-lingual representation space reach the best overall performance, outperforming both baselines as well as both classification methods at a much lower computational cost (Section 5.2; Appendix B). This highlights how the current genre annotations are far from universal, yet can still guide our local-to-global instance-level genre predictors in identifying cross-lingually consistent, data-driven notions of genre.

Future work may be able to improve instance genre prediction by using a more consistent label set or human annotations. The definition of genre macro-classes or a broader taxonomy covering existing annotations could also guide further investigations into cross-lingual language variation. Nonetheless, we expect the task of predicting sentence genre to remain difficult due to the short context within which both annotators and models must make their predictions.

Within the complex scenario of highly cross-lingual, instance-level genre classification, our methods have nonetheless demonstrated that genre is recoverable across the 114 languages in UD — shedding light on prior genre-driven work as well as enabling future research to more deliberately control for additional dimensions of language variation in their data.

Acknowledgements

We would like to thank the NLPnorth group for insightful discussions on this work — in particular Elisa Bassignana and Mike Zhang. Thanks to Héctor Martínez Alonso for feedback on an early draft as well as ITU's High-performance Computing Cluster team. Finally, we thank the anonymous reviewers for their helpful feedback. This research is supported by the Independent Research Fund Denmark (DFF) grant 9063-00077B and an Amazon Faculty Research Award (ARA).

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised Domain Clusters in Pretrained Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany, August. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France, August. Association for Computational Linguistics.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine, April. Association for Computational Linguistics.

Flavio M Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7. CEUR-WS. org.

Euisun Choi and Chulhee Lee. 2003. Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, 36:1703–1709, 08.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France, August. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. Cost-effective selection of pretraining data: A case study of pretraining BERT on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1675–1681, Online, November. Association for Computational Linguistics.

Joe Davison. 2020. Zero-Shot Learning in Modern NLP, May. Accessed December 4th, 2020.

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, June. Association for Computational Linguistics.

Elisa Di Nuovo, Cristina Bosco, Alessandro Mazzei, and Manuela Sanguinetti. 2019. Towards an italian learner treebank in universal dependencies. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR-WS.

Fonticons. 2021. Font Awesome Icons. CC-BY 4.0 License.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July. Association for Computational Linguistics.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovskỳ, and Jan Raab. 2008. The Czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89(1):41–96.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Madrid, Spain, July. Association for Computational Linguistics.

John Lee, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden, May. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *Computing Research Repository*, arXiv: 1711.05101. version 3.

Mikko Luukko, Aleksi Sahala, Sam Hardwick, and Krister Lindén. 2020. Akkadian treebank for early neo-assyrian royal inscriptions. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 124–134, Düsseldorf, Germany, October. Association for Computational Linguistics.

Olga Lyashevskaya, Angelika Peljak-Łapińska, and Daria Petrova. 2017. UD_Belarusian-HSE. `https://github.com/UniversalDependencies/UD_Belarusian-HSE`.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Maria Mitrofan, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2019. MoNERo: a biomedical gold standard corpus for the Romanian language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 71–79, Florence, Italy, August. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Nikola Ljubešić, Olga Loginova, 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, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Davide Rovati, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan

Seraji, Lena Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0 – CoNLL 2017 shared task development and test data. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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, May. 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, November. Association for Computational Linguistics.

Agnieszka Patejuk and Adam Przepiórkowski. 2018. *From Lexical Functional Grammar to Enhanced Universal Dependencies: Linguistically informed treebanks of Polish*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philipp Petrenz. 2012. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–21, Avignon, France, April. 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, Portland, Oregon, USA, June. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS*, Bochum, Germany, September.

Ines Rehbein and Felix Bildhauer. 2017. Data point selection for genre-aware parsing. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 95–105, Prague, Czech Republic.

Rudolf Rosa. 2015. Parsing natural language sentences by semi-supervised methods. *Computing Research Repository*, arXiv: 1506.04897. version 1.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Pegah Safari, Amirsaeid Moloodi, and Alireza Nourian. 2020. The Persian dependency treebank made universal. *arXiv e-prints*, pages arXiv–2009.

Alessio Salomoni. 2019. UD_German-LIT. `https://github.com/UniversalDependencies/UD_German-LIT`.

Stephanie Samson and Cagrı Cöltekin. 2020. UD_Tagalog-TRG. `https://github.com/UniversalDependencies/UD_Tagalog-TRG`.

Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada. Association for Computational Linguistics.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Serge Sharoff. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*, pages 83–94.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:"Taiga" syntax tree corpus and parser. In *Proceedings of "CORPORA-2017" International Conference*, pages 78–84.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Anders Søgaard. 2011. Data point selection for cross-language adaptation of dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 682–686, Portland, Oregon, USA, June. Association for Computational Linguistics.

Achim Stein and Sophie Prévost. 2013. Syntactic annotation of medieval texts. *New methods in historical corpora*, 3:275.

Will Styler. 2011. The Enronsent Corpus.

Sara Stymne. 2020. Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany, October. Association for Computational Linguistics.

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, Hong Kong, China, November. Association for Computational Linguistics.

Sappadla Prateek Veeranna, Jinseok Nam, Eneldo Loza Mencıa, and Johannes Fürnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In *Proceeding of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: Elsevier*, pages 423–428.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682, Suntec, Singapore, August. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Computing Research Repository*, arXiv: 1609.08144. version 2.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923, Hong Kong, China, November. Association for Computational Linguistics.

Shorouq Zahra. 2020. Parsing low-resource Levantine Arabic: Annotation projection versus small-sized annotated data.

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

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, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, 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, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, 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, Gustavo Godoy, 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, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, 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, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, 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, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn 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, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, 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, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, 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, Marinella Testori, 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, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, 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, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix

A Universal Dependencies Setup

All experiments make use of Universal Dependencies v2.8 (Zeman et al., 2021). From the total set of 202 treebanks, we use all except for the following two (due to licensing restrictions): *Arabic-NYUAD* and *Japanese-BCCWJ*. In total 1.51 million sentences are used in our experiments.

Data Splits The experiments in Section 5 use the 204k global test split. Initial comparisons were performed on the 915k dev set. The 102k training split was used to fine-tune CLASS and BOOT. For early stopping, 31k sentences from the latter split were used as a held-out set. The exact instances are available in the associated code repository for future reproducibility.

Genre Mapping For 26 treebanks with instance-level genre labels in the metadata comments before each sentence, we created mappings from the treebank genre labels to the UD genre label set according to the guidelines described in Section 4.2. The genre metadata typically either follow the format `genre = X` or are implied by the document source specified in the sentence ID (e.g. `sent_id = genre-...`). There are a total of 91 mappings which will be made available with the codebase upon publication.

B Model and Training Details

The following describes architecture and training details for all methods. When not further defined, default hyperparameters are used. Implementations and predictions are available in the code repository at `https://personads.me/x/syntaxfest-2021-code`.

Infrastructure Neural models are trained on an NVIDIA A100 GPU with 40 GB of VRAM.

Language Model This work uses mBERT (Devlin et al., 2019) as implemented in the Transformers library (Wolf et al., 2020) as `bert-base-multilingual-cased`. Embeddings are of size $d_{\text{emb}} = 768$ and the model has 178 million parameters. To create sentence embeddings, we use the mean-pooled WordPiece embeddings (Wu et al., 2016) of the final layer.

Classification CLASS and BOOT build on the standard mBERT architecture as follows: mBERT $\rightarrow$ `CLS`-token $\rightarrow$ linear layer ($d_{\text{emb}} \times 18$) $\rightarrow$ softmax. The training has an epoch limit of 30 with early stopping after 3 iterations without improvements on the development set. Backpropagation is performed using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 10^{-7} on batches of size 16. The fine-tuning procedure requires GPU hardware which can host mBERT, corresponding to 10 GB of VRAM. Training on the 71k relevant instances takes approximately 10 hours.

Clustering Both *Gaussian Mixture Models* (GMM) and *Latent Dirichlet Allocation* (Blei et al., 2003; LDA) use scikit-learn v0.23 (Pedregosa et al., 2011). LDA uses bags of character 3–6-grams which occur in at least 2 and in at most 30% of sentences. GMMs use the mBERT sentence embeddings as input. Both methods are CPU-bound and cluster all treebanks in UD in under 45 minutes.

Random Initializations Each experiment is run thrice using the seeds 41, 42 and 43.

C Additional Results

Table 2 shows results on the 915k development split of UD. Performance patterns are similar to those on the test split: the labeled clustering methods GMM+L and LDA+L perform best out of our proposed methods and outperform the baselines on the majority of metrics. With respect to classification, BOOT outperforms both the noisier CLASS and ZERO. Note that the frequency baseline FREQ performs especially well on the dev set, since only 5 of 26 instance labeled treebanks are included and 4 of these have the majority genre *news*.

METHOD	PUR	AGR	ΔBC	F1
FREQ	100±0.0	100±0.0	23±0.0	27±0.0
ZERO	43±0.0	66±0.0	50±0.0	5±0.0
CLASS	87±1.2	77±3.9	29±1.9	9±4.5
BOOT	95±0.2	100±0.0	24±0.3	16±1.0
GMM	92±0.1	55±5.5	30±0.7	—
+LABELS	100±0.0	100±0.0	5±0.1	17±1.6
LDA	88±1.0	42±2.2	30±0.2	—
+LABELS	100±0.0	100±0.0	5±0.0	15±0.9

Table 2: **Results of Genre Prediction on UD (Dev).** Purity (PUR ↑), agreement (AGR ↑), overlap error (ΔBC ↓) and micro-F1 over instance-labeled TBs (F1 ↑) for FREQ, ZERO, CLASS, BOOT and GMM, LDA with/without labels. Standard deviation denoted ±.

Asia Minor Greek in Contact (AMGiC): Towards a dialectal treebank comprising contact-induced grammatical changes.

Konstantinos Sampanis
Boğaziçi University

Prokopis Prokopidis
Institute for Language and Speech Processing/Athena RC

Abstract

In this contribution we briefly present methodological and theoretical aspects of the "Asia Minor Greek in Contact" (AMGiC) treebank. AMGiC is a project in preparation that comprises annotated sentences of contact-induced morphosyntactic change observed in Greek varieties spoken in the region of Cappadocia in Anatolia until the beginning of the 20th century. The treebank is being compiled in accordance with the Universal Dependency annotation scheme and incorporates a geodemographic and a sociolinguistic component in its metadata so that it serves as a tool for comprehensive research in the domain of language contact.

1 An Asia Minor Greek treebank focusing on language contact

"Asia Minor Greek in Contact" (AMGiC) is a treebank in preparation which follows the Universal Dependencies (UD) annotation scheme (Nivre et al. 2020, Marneffe et al. 2021). The treebank, which we present herein, is characterized by two "peculiarities":

a) AMGiC consists of material from Inner Asia Minor Greek (AMG). Inner AMG comprises several interrelated but clearly distinct Cappadocian subdialects as well as the varieties of Silliot and Pharasiot (cf. Manolessou, 2019). Cappadocian Greek (CG), Silliot and Pharasiot are in fact classified as distinct dialects (cf. Janse, 2020: 203). Nevertheless, there are several arguments in favor of examining these dialects together: the dialects share several lexical and grammatical similarities, in terms of geography they were all located in central Anatolia[1] and they were all subject to considerable influence of Turkish varieties. Given that the ISO 639-3 code we utilize for AMGiC is *cpg*, i.e. "Cappadocian Greek", we sometimes employ CG as a *pars pro toto* designation for all Inner AMG varieties. Thus, the terms 'CG' and 'Inner AMG' are interchangeable in our text unless we specify the (sub)dialect within Inner AMG.

b) The treebank chiefly gleans instantiations of sentences which exhibit cases of Contact-Induced MorphoSyntactic Phenomena (CIMSP) triggered by a century-long contact between Greek and Turkish varieties in Central Anatolia. The impact of Turkish on CG is regarded in the relevant literature as a par excellence case of intensive Language Contact (LC) which led the Greek (sub)dialects to significant grammatical changes (cf. Thomason and Kaufman, 1988; Thomason, 2001; Johanson, 2002; Winford, 2003). Several CIMSP attested in CG have been thoroughly examined and analyzed (cf. e.g. Janse, 2009a; Kappler, 2011) yet there is neither an annotated CG treebank nor a detailed list of these phenomena. AMGiC aims to offer both.

[1] The term 'Anatolia' is usually identified with the Asian part of Turkey although, geographically, it is more correct to distinguish the Mesopotamian area of southeastern modern Turkey from the Anatolian one. The terms 'Anatolia' and 'Asia Minor' are used as synonymous here.

AMGiC aligns with BOUN Laz treebank in offering a UD compliant treebank of a linguistically understudied Anatolian variety (cf. Türk et al, 2020). It also resembles Turkish–German code-switching treebank (Çetinoğlu and Çöltekin, 2019) inasmuch as AMGiC is similarly a treebank having a "special focus" on contact-induced grammatical phenomena occurring within the boundaries of a sentence. On the other hand, AMGiC's architecture entails certain novelties which we briefly present below.

The structure of this paper is as follows: in section 2 we provide some information regarding the linguistic profile of the varieties we will examine (§2.1), we argue in favor of the relevance of a UD approach to Inner AMG/CG in the light of a "peculiar" syntactic structure of the variety (§2.2) and we present the methodology we follow towards the compilation of AMGiC (§2.3). In section 3 we deal with certain particularities of the Inner AMG/CG treebank and we refer to some challenges that emerge while working on AMGiC. Section 4 sums up our discussion and highlights the importance of the preparation of an annotated treebank of these Greek varieties.

2 Grammatical Features of Inner AMG/CG and the AMGiC treebank

2.1 Introductory remarks

Inner AMG/CG (and related varieties) were spoken in Anatolia until 1923, the year when a population exchange between Greece and Turkey obliged CG speakers to abandon their homeland. The diachronic development of CG varieties is shrouded in mystery due to the absence of any written records until the end of the 19th century when certain writers present some short text collections which are however very unsystematically collected. It was a single work published in 1916 by Dawkins, a British scholar, who conducted in situ research in central Anatolia and compiled a grammar of Inner AMG (sub)dialects, that shed light on the linguistic situation in Cappadocia at that time. Along with reporting and recording the Greek dialects in a region in which Turkish was expected to be dominant, Dawkins also emphasized that CG was shaped under the intensive influence of Turkish varieties to the extent that Dawkins aphoristically stated that "the Turkish has replaced the Greek spirit; the body has remained Greek, but the soul has become Turkish" (Dawkins, 1916: 198). The dramatic undertones of Dawkins are suggestive of the fact that CG has undergone a substantial grammatical restructuring that differentiates it from the rest of the Greek dialects, even from Pontic or Aegean AMG. Some of the contact-induced grammatical features that CG developed under the Turkish influence is the borrowing of numerous free grammatical elements (cf. Melissaropoulou and Ralli, 2020), development of agglutinating-like declension and conjugation (cf. Janse, 2009b, 2019; Karatsareas, 2016; Revithiadou et al., 2017), encliticization of the copula verb, left-branching/head-final syntactic structures etc. It is due to all these contact-induced features that Janse (2009a: 37) described CG as a "mixed language" (similarly Winford, 2003 referred to CG as "a Greek Turkish mixture").

While Dawkins had already provided us with a first-detailed list of the contact-induced phenomena (Dawkins, 1916: 209), there are only few attempts to revise this list in the light of state-of-the-art LC research (cf. Theodoridi, 2017 and Karantzola et al., forthc.) and, crucially, there is no annotated treebank of Cappadocian[2]. AMGiC aims to offer a treebank of that sort with a focus on contact-induced phenomena which attract the interest of LC scholars as well as with a sociolinguistic metadata component to which we will refer in section 3 below.

2.2 Dealing with syntactic issues

An obvious advantage of a UD analysis for CG is the fact that this facilitates an immediate typological comparison between - for instance - CG and Standard Modern Greek (SMG) or Standard Turkish. Notwithstanding their Greek provenance, several Cappadocian varieties' syntactic structures considerably differ from respective SMG ones. Consider example (1) below:

[2] A CG Dialectal Atlas as well as a Dialectal eDictionary have been announced within the framework of the DiCaDLand (Digitizing the Cappadocian Dialectal Landscape) project, cf. `http://cappadocian.upatras.gr/en`.

(1)[3] Inner AMG/CG: Settlement of Ulaghátsh[4]

írte	*'na devjú*	*manajú*	*t'*	*to spit*
come.AOR.3SG.	*a-giant*.GEN.	*mother*.GEN.	POSS.	*the-house*.ACC.

"(S/he) came to the house of the mother of a giant."

In CG sentence in (1) the genitive complements of the noun phrases are preposed. In simple words, the head of the noun phrase (NP) *spit* (N^o_1) has a possessive complement in genitive, namely *manajú* (N^o_2) which in turn has another noun in genitive as a complement, namely *devjú*. Schematically, this can be written down as follows: [[[NP] ← N^o_2] ←N^o_1].

In Standard MG such an array of complements is ungrammatical. The default syntactic order of a "genitive chain" of possessive constructions would be as in (2):

(2) Standard Modern Greek

írthe	*s-to spíti*	*tis mánas*	*enós ghíghanda*
come.AOR.3SG.	*to-the-house*.ACC.	*the-mother*.GEN.	*a-giant*.GEN.

"(S/he) came to the house of the mother of a giant."

So, the structure of the phrase in (1) is as follows: [N^o_1 → [N^o_2 → [NP]]]. It is similarly grammatical to prepose the entire embedded phrase for topicalization/focalization yielding a phrase like this: *tis mánas enós ghíghanda to spíti* (*the-mother*.GEN. *a-giant*.GEN. *the-house*.NOM./ACC.). In that case the structural analysis involves a phrasal movement, not a directionality shift: [[N^o_2 → [NP]]$_i$ N^o_1 → [N^o_2 → [NP]]$_i$]. As expected, a phrase comprising head bidirectionality or extraction from Complex NP[5] is ungrammatical: **enós ghíghanda tis mánas to spíti* (*the-giant*.GEN. *the-mother*.GEN. *the-house*.NOM./ACC.).

Now, after applying the UD annotation scheme on (1) and (2) the respective sentences can be visualized as follows:

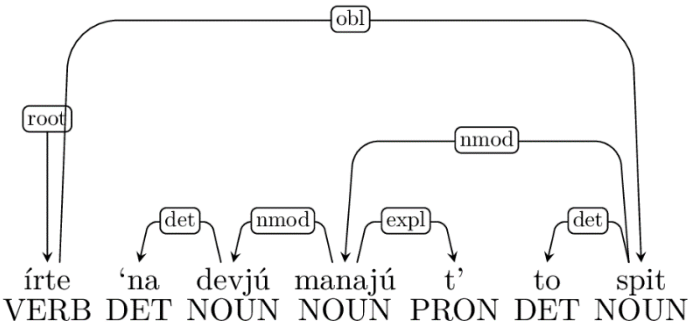


Figure 1: UD annotation scheme visualization of CG sentence (1)

[3] Cf. Dawkins, 1916: 378. Cf. also discussion in Theodoridi, 2017: 489.

[4] Ulaghátsh (Turkish spelling: *Ul(u)ağaç*) was one of the approximately twenty Cappadocian villages in which CG was spoken. The varieties spoken in each village differed from each other with respect to the degree of the Turkish influence they exhibited (cf. e.g. Karatsareas, 2011: 11ff).Thus, CG should be understood as a cover term of interrelated yet distinct dialectal varieties. Due to this extended variation within CG, we regularly refer to "CG (sub)varieties". The CG variety of Ulaghátsh was one of the most heavily influenced by Turkish (Dawkins, 1916: 209; Janse, 2020: 203f) and therefore also one of the most interesting for observing CIMSP.

[5] On the Complex NP Constraint -which stems from a generative theoretical framework- cf. Bošković, 2015.

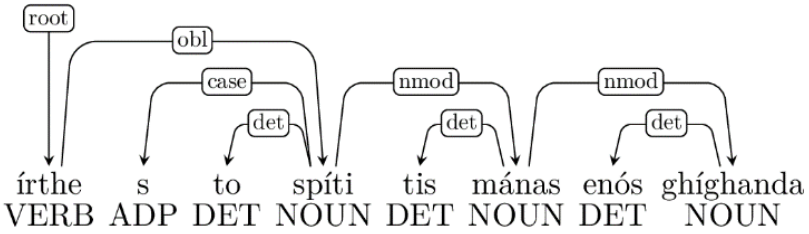


Figure 2: UD annotation scheme visualization of SMG sentence (2)

The contrastive presentation of CG and SMG does not only serve to illustrate the difference in the directionality of head dependencies but also presents an explicit analysis of these structures which is typologically useful. Although we do not want to attempt a thorough comparison between dependency grammar and other models of syntactic analysis, it is tantalizing to think of proposals within the generative grammar framework which may turn to be less clear-cut in the description of typological differentiation and change - consider e.g. the "Linear Correspondence Axiom" proposed by Kayne (1994), a theory that postulates a universal head-complement syntactic linearization, opposes the head directionality parameter and accounts for typological variation by means of constituent movements which cannot be easily justified by default word orders. On these grounds, a UD approach can be deemed more appropriate for a straightforward typological analysis.

Another advantage of a UD analysis is the fact that the annotation scheme "obliges" the annotator to make a decision about the exact description of an observed phenomenon. This is especially relevant in cases of varieties such as AMG/CG which lack a linguistically based descriptive grammar. For instance, in (1) the CG sentence entails the phrase *manajú t'* that can be roughly translated as 'mother of him' with *t'* referring to the noun *devjú*. The structure is unknown to non-CG Greek dialects and Dawkins (1916: 201) was right in indicating the Turkish 3rd person possessive ending *-(s)I* (Göksel and Kerslake 2005: 45) as the trigger for the formation in the Cappadocian variety.[6] In our annotation we analyze *t'* as a pleonastic nominal and we assign it the UD `expl`[7] label. The `expl` label has also been used for the annotation of constructions involving clitic doubling in Modern Greek[8], constructions that exist in AMG/CG as well. In doing so, we draw a distinction between the influence of Turkish and the replication of a phenomenon in Greek: The Turkish structure is adopted yet by means of existing syntactic features of the Greek variety.[9] Accordingly, the cross-linguistic typological uniformity of the UD annotation scheme discourages *ad hoc* analyses for phenomena noted in less studied linguistic varieties.

2.3 Inner AMG textual sources and the compilation of AMGiC

As afore-mentioned, Dawkins was the sole researcher to collect texts of Inner AMG *in situ*, i.e. in Anatolia before 1923's population exchange and therefore his opus is a principal textual source for AMGiC. After the population exchange, the Centre for Asia Minor Studies[10] published a number of grammars on AMG (sub)dialects: on the dialect of the settlement of Ulaghátsh (Kesisoglou, 1951), on Aravaní (Phosteris and Kesisoglou, 1960), on Anakú (Costakis, 1964), on Sílli (Costakis, 1968). These works along with Dawkins 1916 constitute the pool for the "mining" of sentences that comprise CIMSP which are in turn annotated for AMGiC. The treebank is not exhaustive, in the sense that not all sentences featuring contact-induced phenomena are included. It is however representative of all phenom-

[6] Compare the Standard Turkish equivalent of (1): *devin anne-**si**-nin evine geldi* (= *giant*.GEN. *mother*.***his***.GEN. *house*.DAT *come*.3SG.PAST.

[7] https://universaldependencies.org/u/dep/expl.html

[8] Cf. `https://universaldependencies.org/el/dep/expl.html`

[9] Joseph (2000: 22) argued that "the syntactic similarities found in Sprachbünde and other contact situations tend to be superficial in nature and are really a matter of a convergence in surface structure, rather than in deep structure". This can also mean that languages tend to get grammatically similar by generalizing existing structures of each language.

[10] `http://en.kms.org.gr/`

ena of that sort. Hence, upon completion, AMGiC will comprise instantiations of every single morpho-syntactic phenomenon that emerged due to LC. An indicative list of these phenomena and their predefined tags is as follows:

1.3. FrGrEl = Free Grammatical Elements
1.3.1. AdvMod = Modal Adverb
1.3.2. AdvTime = Time Adverb
1.3.3. AdvSp = Space Adverbs
1.3.4. ConjCo = Conjunction/Coordinator(s)
1.3.5. ConjSub = Conjunction/Subordinator
1.3.6. Det = Determiners
1.3.7. EmphPart = Emphatic Particle
1.3.8. NegQ = Negation Quantifier
1.3.9. Num = Numerals
1.3.10. Post = Postposition
1.3.11. Quant = Quantifier
1.3.12. QPart = Question Particle
1.3.13. SentPart = Sentential Particles
1.3.14. WhW = "Wh"-Words

1.6. SynIn = Pattern Replication/Syntactic Interference
1.6.1. FunV = Functional Verbs
1.6.2. HFin = Head Final
1.6.3. HFinNC = Head Final/Nominal Complements
1.6.4. HFinPost = Head Final/Postposition
1.6.5. HFinVFin = Head Final/Verb Final
1.6.6. IdEx = Idiomatic Expressions
1.6.7. Red = Reduplication
1.6.8. RelCl = Relative Clauses

Apart from the obvious utility of AMGiC as a tool for LC researchers, our treebank offers a concrete categorization of CIMSP which can eventually be applied to other analogous cases of LC. CIMSP are provided both as comments at the initial metadata section or as an annotation component at the CoNLL-U `MISC` field. In particular, the exact incorporation of CIMSP is as follows: a sentence is gleaned by the afore-mentioned textual pool because of a contact-induced phenomenon it contains. In AMGiC the sentence is manually annotated and once the contact-induced phenomenon is located this is initially marked with `LC=YES` (LC = Language Contact) at `MISC`. Subsequently, AMGiC provides the morphosyntactic category (`MorphSynC`) and subcategory (`MorphSynSC`) of the phenomenon in case. CIMSP categories and subcategories are codified in the annotation as predefined tags.

(3) Inner AMG: Settlement of Sílli[11]
...írtis *ro* *m'* *ki?*
...come.2sg.aor. *here* QPART EMPHPART
"…did you really come here?" (Kostakis 1968:116)

```
16      m'    mi    AUX   _      _      14    aux:q _
  LC=YES|MorphSynC=FrGrM|MorphSynSC=QPart
17      ki    ki    ADV   _      _      14    advmod:emph _
  LC=YES|MorphSynC=FrGrM|MorphSynSC=EmphPart
```

[11] Sílli was the only Greek-speaking enclave in the region close to the city of Iconium (Modern Turkish: *Konya*).

Example (3) is part of a wider sentence annotated in AMGiC. The underlined free grammatical elements *m'* and *ki* are borrowed from Turkish (`LC=YES`). The former free element is a Question Particle (QPart) used in yes/no questions and is tagged as a 'Free Grammatical Morpheme' for the broad morphosyntactic category (`MorphSynC=FrGrM`) and a Question Particle for the morphosyntactic subcategory (`MorphSynSC=QPart`). In the same vein, the latter free element is an Emphatic Particle (EmphPart) that expresses surprise and is tagged as a 'Free Grammatical Morpheme' for the broad morphosyntactic category (`MorphSynC=FrGrM`) and an 'Emphatic Particle' for the morphosyntactic subcategory (`MorphSynSC=EmphPart`). Hence, AMGiC provides a fine-grained categorization of CIMSP which is easily searchable and is open to statistical approaches.

3 Structural particularities of AMGiC

AMGiC tackles mainly oral, dialectal, non-standardized material of language mixing which often entails highly "idiosyncratic" constructions. An interesting case is illustrated in (4) in which the AMG variety employs the grammaticalized-converb/subordinator *deyí* from Turkish (< Ottoman *deyü*, Standard Modern Turkish: *diye*):

(4)) Inner AMG: Settlement of Sílli

Vavás	čis	éršiti,	náftši	ta ksíla	op'	čin	iréan
father	*her*	*comes*	*lights*	*the-wood*.ACC.PL.N.	*from*	*the*	*idea*.ACC.SG.F.

óči	kóri	apés'	tun	éni	deyí
that	*daughter*	*inside*	*these*	*is*	SUBORD

"Her father comes, he sets light to the wood, thinking that his daughter is inside."
(Dawkins 1916:284)

Following Göksel and Kerslake (2005: 354) on Modern Turkish *diye* we designate the CG borrowed form *deyí* as a subordinator although this grammatical element is syntactically so "versatile" that this designation may be somehow restrictive (cf. Gündoğdu 2017). In AMGiC almost every borrowed grammatical element is regarded as integrated part of the Inner AMG variety, not as a case of code-switching, since the large extent of Turkish influence and the incorporation of Turkish lexical and grammatical features is an essential - not a coincidental - aspect of Inner AMG/CG. Nevertheless, the cooccurrence of structures of both Greek and Turkish provenance gives rise to grammatical configurations that can be challenging for the annotators, at least initially.

Figure 3a indicates our first annotation approach, according to which *deyí* has the same dependency relation with *óči* 'that' (SMG: *óti*), and therefore the Turkish element can be seen at face value as "pleonastic". This analysis is not paradoxical, given that similar "pleonastic" constructions in which a "genuine" Greek and a borrowed Turkish grammatical element cooccur are attested, cf. e.g. Kesisoglou (1951: 60) on coexisting conditional subordinators and Bağrıaçık (2018: 295ff) for a similar phenomenon in Pharasiot.

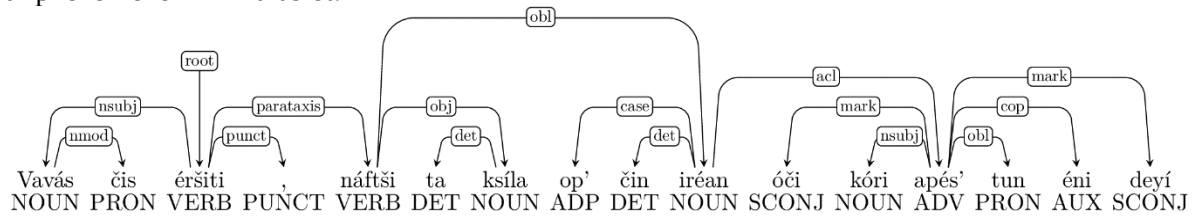


Figure 3a: UD annotation scheme visualization of the first analysis of (3)

While the first analysis could be valid in the light of germane phenomena in the Inner AMG varieties, we decided to revise the syntactic analysis as demonstrated in Figure 3b: *deyí* is now dependent on the noun *iréan* 'idea' so it introduces the cause why "the father sets light to the woods". This analysis seems

to be both more elegant and accurate, yet it is not one not causing problems. In this case, *deyí* introduces the cause for father's action but so also does the prepositional phrase *op'čin iréan*. What is more, tagging *deyí* as a marker may be seen as erroneous since "marker is the word marking a clause as subordinate to another clause"[12] and in this case there is no subordinated clause. However, the prepositional phrase functions semantically as a kind of adverbial clause or converb (compare the translation of the prepositional phrase as "thinking"). Should we assign a case dependency relation to *deyí* (as it is the case sometimes with the Standard Turkish equivalent *diye*) we miss the subordinating function of the grammatical element and we would again face the problem of having "pleonastic" dependents on the noun of the prepositional phrase *iréan*, namely both *op'* and *deyí*. Clearly, this is a tricky point which reveals the challenges of working with contact-induced phenomena in dialectal varieties.

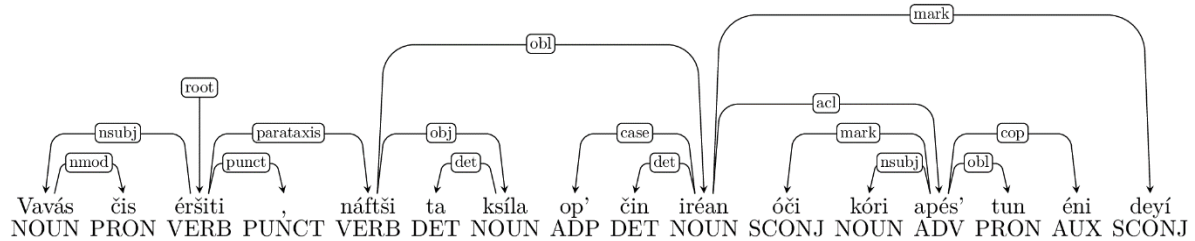


Figure 3a: UD annotation scheme visualization of the updated analysis of (3)

Another interesting aspect of AMGiC is the contribution to the analysis of Greek grammatical phenomena in general. Although Inner AMG deviates considerably from Standard Modern Greek, there are several grammatical structures shared with most Greek varieties, one of which is clitic doubling,[13] a phenomenon we referred to in the previous section. Standard Modern Greek has already been analyzed within the UD framework through the Greek UD treebank (UD_Greek-GDT, cf. Prokopidis and Papageorgiou, 2017). A crucial distinction between AMGiC and UD_Greek-GDT is the afore-mentioned orality and dialectal character of the former. Due to these features, AMGiC is expected to comprise more structures that are found in the spoken language. Indeed, clitic doubling is recurrent in the AMG sources whereas it is attested only once in UD_Greek-GDT due to the fact that this was compiled on the basis of written sources or parliamentary sessions the register of which is more formal. Consider example (5) below:

(5) Inner AMG: Settlement of Sílli

ke	tus gjavúriri		re	se
and	*the infidels*		*not*	FUT
tus	eleísis	xets,	se	su páru.
them	*harm.*2SG.	*at all*	FUT	*you.*ACC. *take.*1SG.

"And you will not harm the infidels (i.e. the Christians), (then) I will marry you."

The occurrence of the clitic doubling structure is realized by the usage of the weak pronominal *tus* which semantically refers to the noun *gjavúriri*. In terms of UD annotation, *tus* is assigned an expletive dependency relation. This is observable in Figure 4 below.

[12] `https://universaldependencies.org/u/dep/mark.html`
[13] On clitic doubling in AMG cf. Janse, 2008. Cf. also Condoravdi and Kiparsky, 2002 on clitics in the diachrony of Greek.

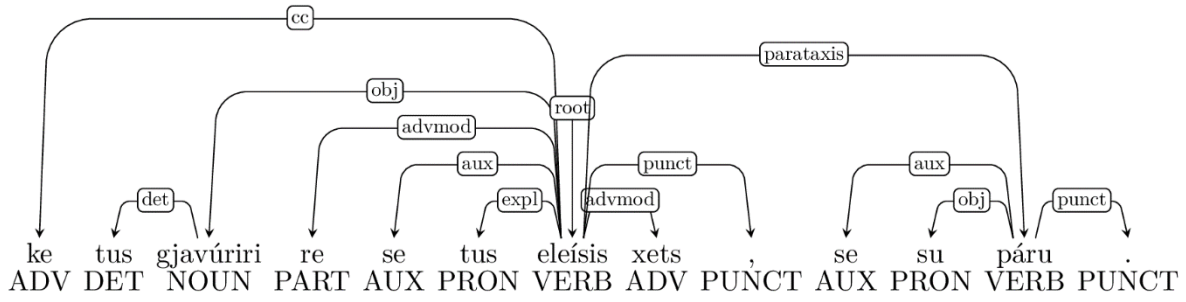


Figure 4: UD annotation scheme of a clitic doubling instantiation. The annotation for the *tus* pronoun includes the `Case=Acc|Clitic=Yes|Gender=Masc|Number=Plur|Person=3|PronType=Prs` feature/value pairs.

4 Conclusions and Further Desiderata

In this contribution we provided a brief presentation of AMGiC, a treebank in preparation focusing on Inner Asia Minor Greek dialects. The treebank, which comprises instantiations of contact-induced morphosyntactic phenomena that emerged through the longstanding contact between Greek and Turkish varieties in central Anatolia. As stated, the compilation of AMGiC posits several challenges that has to do with the examination of the understudied Asia Minor Greek varieties, since the incorporation of Turkish elements gave rise to a typologically "peculiar" grammatical architecture. AMGiC does not only provide a digital annotation of these varieties, but it also puts forward proposals of syntactic analyses that are of interest for researchers of language change, syntax and typology.

The preparation of AMGiC aligns with a wider research project of correlating contact-induced morphosyntactic phenomena with sociocultural and geodemographic parameters (cf. acknowledgements). The treebank entails a sociolinguistic component in its metadata that can be statistically related to the respective description of contact-induced phenomena of each annotated sentence. On these grounds, AMGiC is unique in its design and objectives.

Acknowledgements

This publication/paper has been produced benefiting from the 2236 Co-Funded Brain Circulation Scheme2 (CoCirculation2) of TÜBİTAK (Project No: 120C061). However, the entire responsibility of the publication/paper belongs to the owner of the publication/paper. The financial support received from TÜBİTAK does not mean that the content of the publication is approved in a scientific sense by TÜBİTAK.

References

Bağrıaçık, Metin. 2018. *Pharasiot Greek : word order and clause structure*. Unpublished Dissertation. Ghent University, Ghent, Belgium.

Bošković, Željko. 2015. From the Complex NP Constraint to everything: On deep extractions across categories. *The Linguistic Review*, 32: 603 - 669.

Condoravdi, Cleo and Paul Kiparsky. 2002. Clitics and clause structure. *Journal of Greek Linguistics, 2*(1): 1-39.

Costakis, Athanasios P. 1964. *Le parler grec d'Anakou*. Centre for Asia Minor Studies, Athens, Greece.

Costakis, Athanasios P. 1968. *To ghlosikó idhíoma tis Síllis*. [In Greek: Τὸ γλωσσικὸ ἰδίωμα τῆς Σίλλης. - 'The dialect of Silli']. Centre for Asia Minor Studies, Athens, Greece.

Dawkins, Richard M. 1916. Modern Greek in Asia Minor: A Study of the Dialects of Sílli, Cappadocia and Phárasa with Grammar, Texts, Translations and Glossary. Cambridge: Cambridge University Press.

Janse, Mark. 2008. Clitic doubling from Ancient to Asia Minor Greek. In Dalina Kallulli and Liliane Tasmowski (eds.), *Clitic doubling in the Balkan languages*, pages 165–202. John Benjamins, Amsterdam / Philadelphia.

Janse, Mark. 2009a. Greek-Turkish language contact in Asia Minor. *Études Helléniques/Hellenic Studies* 17: 37–54.

Janse, Mark. 2009b. Watkins' Law and the development of agglutinative inflections in Asia Minor Greek. *Journal of Greek Linguistics* 5.3-26: 93–109.

Janse, Mark. 2019. Agglutinative Noun Inflection in Cappadocian. In Angela Ralli (ed.), *The Morphology of Asia Minor Greek*. pages 66–115, Brill, Leiden/Boston.

Janse, Mark. 2020. Back to the future: Akritic light on diachronic variation in Cappadocian (Asia Minor Greek). In Klaas Bentein and Mark Janse (eds.), *Varieties of Post-classical and Byzantine Greek*, pages 201 - 239. Mouton De Gruyter, Berlin/Boston.

Johanson, Lars. 2002. *Structural factors in Turkic language contacts.* Richmond, Curzon.

Joseph, Brian. 2000. Is Balkan comparative syntax possible? In María Luisa Rivero and Angela Ralli (eds.), *Comparative Syntax of Balkan Languages*, pages 17–43, Oxford University Press, New York.

Göksel, Aslı and Celia Kerslake. 2005. *Turkish: A comprehensive grammar*. Routledge, London and New York.

Gündoğdu, Hilal Yıldırım. 2017. *The Structure of Diye Clauses in Turkish.* Unpublished doctoral thesis. Boğaziçi University, Istanbul, Turkey.

Kappler, Matthias. 2011. A Tale of two Languages: Tracing the History of Turkish-Greek Language Contacts. *Türk Dilleri Araştırmaları*, 21.1: 95-130, Ankara, Turkey.

Karantzola, Eleni; Anatoli Theodoridi and Konstantinos Sampanis. forthc. The Interplay of External and Sociolinguistic Factors in contact-induced language change: Cappadocian Greek as a case study. *Mediterranean Language Review*, (paper accepted for publication).

Karatsareas, Petros. 2011. *A study of Cappadocian Greek nominal morphology from a diachronic and dialectological perspective*. Doctoral Dissertation, University of Cambridge, Cambridge, UK.

Karatsareas, Petros. 2016. Convergence in word structure: Revisiting agglutinative noun inflection in Cappadocian Greek. *Diachronica* 33: 31-66.

Kayne, S. Richard. 1994. *The Antisymmetry of Syntax*. MIT Press, Cambridge/Massachusetts.

Kesisoglou, Ioannis I. 1951. *To ghlosikó idíoma tu Ulaghátsh*. [In Greek: Τὸ γλωσσικὸ ἰδίωμα τοῦ Οὐλαγὰτς - French Cover title: Le dialecte d'Oulagatch). Institut Français d'Athènes, Athens, Greece.

Kostakis, Thanasis. 1968. *To ghlosikó idíoma tis Sílis* [In Greek: Τὸ γλωσσικὸ ἰδίωμα τῆς Σίλλης -'The dialect of Silli']. Center of Asia Minor Studies and French Institute of Athens, Athens, Greece.

Mavrochalyvidis, G. & I. I. Kessissoglou. 1960. *To ghlosikó idíoma tis Axú*. [In Greek: Τὸ γλωσσικὸ ἰδίωμα τῆς Ἀξοῦ - 'The Dialect of Axos']. Center of Asia Minor Studies and French Institute of Athens, Athens, Greece.

Manolessou, Io. 2019. The historical background of the Asia Minor dialects. In Angela Ralli (ed.), *Morphology of the Asia Minor Greek dialects*, pages 20–65, Brill, Leiden, the Netherlands.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47, no. 2: 255–308. https://doi.org/10.1162/coli_a_00402.

Melissaropoulou, Dimitra and Angela Ralli. 2020. Revisiting the Borrowability Scale(s) of Free Grammatical Elements: Evidence from Modern Greek Contact induced Varieties. *Journal of Language Contact* 12(3): 707–736.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. arXiv:2004.10643. Marseille, France.

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.

Phosteris, Dimitrios and Ioannis I. Kesisoglou (1960). *Lexilóghio tu Aravaní.* [In Greek: *Λεξιλόγιο τοῦ Ἀραβανὶ* 'Dictionary of Aravani', Cover French title: *Vocabulaire d'Aravani*], Institut Français d'Athènes, Athens, Greece.

Prokopidis, Prokopis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies* (UDW 2017), pages 102 – 106.

Revithiadou, Anthi, Vassilios Spyropoulos and Giorgios Markopoulos. 2017. From fusion to agglutination: The case of Asia Minor Greek. *Transactions of the Philological Society* 115, pages 297-335. London, United Kingdom.

Thomason, Sarah Grey. 2001. *Language Contact*. Edinburgh University Press, Edinburgh.

Thomason, Sarah Grey and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. University of California Press, Berkeley.

Theodoridi, Anatoli. 2017. Cappadocians dialects and Pharasiotika : sociolinguistic and structural aspects of their contact with Turkish (In Greek: Καππαδοκικές διάλεκτοι και φαρασιωτική: κοινωνιογλωσσικά και δομικά στοιχεία της γλωσσικής επαφής τους με την τουρκική). Unpublished doctoral dissertation, University of the Aegean, Rhodes, Greece.

Utku Türk, Kaan Bayar, Ayşegül Dilara Özercan, Görkem Yiğit Öztürk, Şaziye Betül Özateş. 2020. First Steps towards Universal Dependencies for Laz. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 189–194, Barcelona, Spain.

Parsing with Pretrained Language Models, Multiple Datasets, and Dataset Embeddings

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

Miryam de Lhoneux
Uppsala University
KU Leuven
University of Copenhagen
ml@di.ku.dk

Abstract

With an increase of dataset availability, the potential for learning from a variety of data sources has increased. One particular method to improve learning from multiple data sources is to embed the data source during training. This allows the model to learn generalizable features as well as distinguishing features between datasets. However, these dataset embeddings have mostly been used before contextualized transformer-based embeddings were introduced in the field of Natural Language Processing. In this work, we compare two methods to embed datasets in a transformer-based multilingual dependency parser, and perform an extensive evaluation. We show that: 1) embedding the dataset is still beneficial with these models 2) performance increases are highest when embedding the dataset at the encoder level 3) unsurprisingly, we confirm that performance increases are highest for small datasets and datasets with a low baseline score. 4) we show that training on the combination of all datasets performs similarly to designing smaller clusters based on language-relatedness.[1]

1 Introduction

Many studies have shown the benefits of training dependency parsers jointly for multiple treebanks, either within the same language (Stymne et al., 2018) or across different languages (multilingual models; Ammar et al., 2016; Vilares et al., 2016; Smith et al., 2018), which makes it possible to transfer knowledge between treebanks. This has been enabled by the development of treebanks in multiple languages annotated according to the same guidelines which have been released by the Universal Dependencies (UD; Nivre et al., 2020) project. In this context, a method which has been shown to be effective is the use of embeddings that represent each individual treebank. This is referred to as a *language embedding* (Ammar et al., 2016) or a *treebank embedding* (Smith et al., 2018), we opt for the more general term: *dataset embedding*. The main intuition behind these dataset embeddings is that they allow the model to learn useful commonalities between different datasets, while still encoding dataset specific knowledge.

In the last few years, large multilingual pretrained language models (LMs) pretrained on the task of language modelling, such as mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020), have made it possible to train high performing multilingual models for many different NLP tasks, including dependency parsing (Kondratyuk and Straka, 2019). These models have shown surprising cross-lingual abilities in spite of not getting any cross-lingual supervision. Wu and Dredze (2019), for example, fine-tuned mBERT on English data for five different tasks and applied it to different languages with high accuracy, relative to the state-of-the-art for these tasks. These large pretrained multilingual LMs are now widely used in multilingual NLP.

Dataset embeddings have so far only been evaluated in the context of multilingual dependency parsing without such large pretrained multilingual LMs. Given the large gains in accuracy that have been obtained by using them, and given that they seem to be learning to do cross-lingual transfer without cross-lingual supervision, it is unclear whether or not datataset embeddings can still be useful in this

[1]Code available at: https://bitbucket.org/robvanderg/dataembs2, our implementations will also be included in the next MaChAMp release (v0.3)

context. This is the question we ask in this paper. Our main research question can be formulated as follows:

RQ Is the information learned by dataset embeddings complementary to the information learned by large multilingual LM-based parsers?

2 Background

Vilares et al. (2016) were among the first to exploit UD treebanks to train models for multiple languages. They trained parsing models on 100 pairs of languages by simply concatenating the treebank pairs. They observed that most models obtained comparable accuracy to the monolingual baseline and some outperformed it. This shows that even in its simplest form, multilingual training can be beneficial. Ammar et al. (2016) build a more complex model to train a parser for eight languages. They introduce a vector representation of each language which is used as a feature, concatenated to the word representations of each word in the sentence.

Smith et al. (2018) used this idea in the context of the CoNLL 2018 shared task (Zeman et al., 2018) where the task was to parse test sets in 82 languages. Smith et al. (2018) found that training models on clusters of related languages using dataset embeddings led to a substantial accuracy gain over training individual models for each language.

Stymne et al. (2018) further exploited this dataset embedding method in the monolingual context, using heterogeneous treebanks. They compared this method to several methods including simply concatenating the treebanks to learn a unique parser for all treebanks and found the dataset embedding method to be superior to all other methods.

Wagner et al. (2020) investigated whether dataset embeddings can be useful in an out-of-domain scenario where the treebank of the target data is unknown. They found that it is possible to predict dataset embeddings for such target treebanks, making the method useful in this scenario.

Kondratyuk and Straka (2019) trained a single model for the 75 languages available in UD at the time by concatenating all treebanks and using a large pretrained LM. They found that this did not hurt accuracy compared to training monolingual models, and it even improved accuracy for some languages. This type of model has become standard and has been used in many studies. They did not make use of dataset embeddings in this setup.

Dataset embeddings have mostly been used with BiLSTM parsers. This may partially be due to the fact that they were concatenated to the word embedding before passing it into the encoder, which is non-trivial in LM-based setups where the word embedding and encoding size is fixed. To the best of our knowledge, the only attempt to use dataset embeddings in combination with a LM-based parser was from van der Goot et al. (2021). They use a large pretrained multilingual LM as encoder, and concatenate the dataset embeddings to the word embedding before decoding. They show that this leads to improved performance if the task is the same and the languages/domains of the datasets differ. For a setup where different tasks are combined via multi-task learning (i.e. GLUE), dataset embeddings helped to decrease the performance gap compared to single-task models.

Contributions In this work, 1) we introduce a method to incorporate dataset embeddings also in the encoder in LM-based parsers; 2) we test whether or not dataset embeddings are useful when used in combination with large pretrained multilingual LMs, using both our newly proposed method as well as the existing approach; 3) we compare the effectiveness of dataset embedding when training on small clusters of datasets as well as training on all considered datasets simultaneously.

3 Methodology

3.1 Methods

In most previous implementations of dataset embeddings, the dataset embedding is concatenated to the word embedding before it is passed into the encoder. When using the language model as encoder, as is now commonly done with BERT-like embeddings, this is impossible, as the word embedding is expected

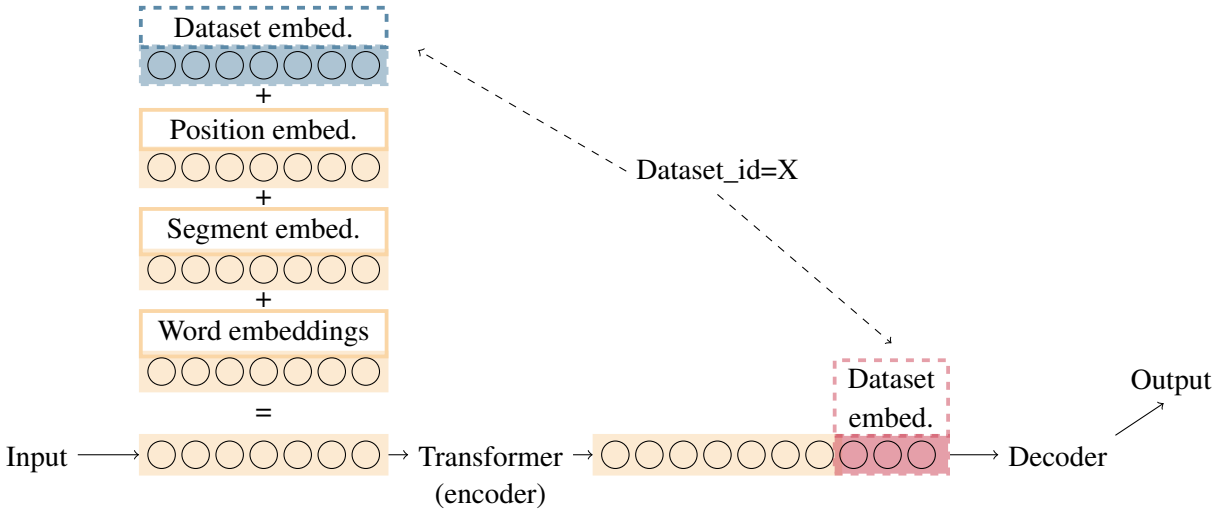


Figure 1: Visualization of the two models to integrate dataset embeddings into transformer based parsers. The dataset embeddings for DECODER is displayed with the red box (dashed on the right), and ENCODER is displayed in the blue box (dashed on the left).

to be of a fixed size. For this reason, we experiment with two alternative setups: concatenating the embedding after the encoding, and summing it to the word embedding. These two approaches are illustrated in Figure 1, and are explained in detail in the following two paragraphs. Note that these approaches can also be used simultaneously, which constitutes our third setup (BOTH).

We use the deep biaffine parser (Dozat and Manning, 2017) implementation of MaChAmp (van der Goot et al., 2021) as a framework for evaluating our models. MaChAmp already includes an implementation of dataset embeddings on the decoder level. In this implementation the output of the LM for each wordpiece is concatenated to the dataset embedding before it is passed on to the decoder. We refer to this approach as DECODER. In this setup, we choose to use dataset embeddings of size 12, based on previous work (Ammar et al., 2016; Smith et al., 2018). This embedding is then concatenated to the output of the transformer, meaning that the input size of the decoder will be the size of the original output + 12.

The second approach incorporates the dataset embeddings in the LM parameters. In most transformer-based LM implementations, multiple embeddings are summed before the transformer layers to represent the input for each wordpiece. In the BERT model for example, these are the token embeddings, segment embeddings and position embeddings. We supplement these by also summing a dataset embedding to this input. In this way, the dataset embeddings can be taken into account throughout the transformer layers. We refer to this approach as ENCODER. In this setup, we choose to match the dimension size of the embeddings at the token level, to avoid having smaller embeddings summing only to an arbitrary subset of the weights.

3.2 Experimental Setup

We essentially reproduce the experiments from Smith et al. (2018) using large pretrained multilingual LMs but do a more large-scale evaluation of the method, by testing more settings, comparing to more baselines and doing more extensive analysis. More specifically, we use the clusters from Smith et al. (2018), but use the updated versions of the treebanks (UD v2.8), and implement all models using MaChAmp, the library by van der Goot et al. (2021). We use all default hyperparameters, including the mBERT embeddings (Devlin et al., 2019) which is used during the original tuning of MaChAmp van der Goot et al. (2021). All reported results are the average over 5 runs with different random seeds for MaChAmp.

To avoid overfitting on the development or test set, we follow van der Goot (2021), and compare our models on the development split while using a tune split for model picking. We will confirm our main findings on the test data in Section 4.3. We use the updated splitting strategy proposed by van der Goot

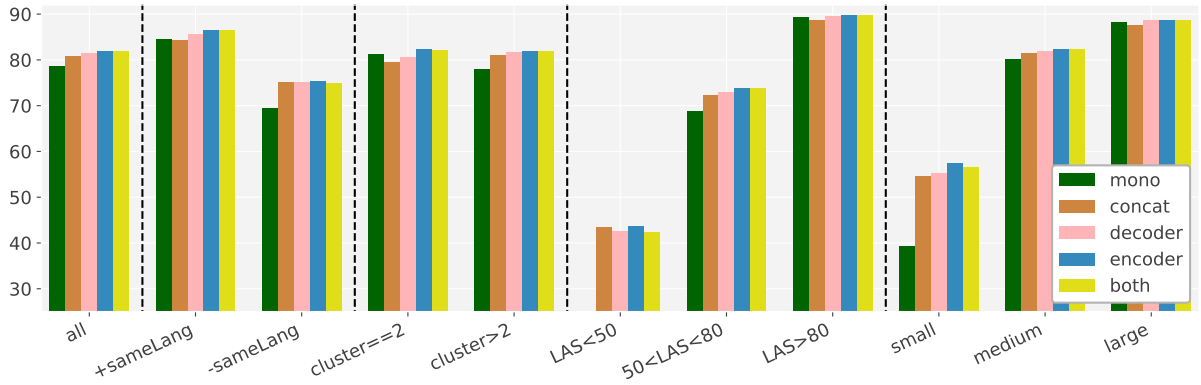


Figure 2: Average LAS scores (dev) over different subsets of the data. `All`: all data, `+sameLang`: datasets for which another in-language treebank exists, `cluster==2`: clusters of size 2, `LAS<50`: treebanks for which the 'mono´ baseline scores <50 LAS, `small, medium, large`: datasets with a maximum size of respectively: 1,000, 10,000, 20,000 sentences.

(2021): for datasets with less then 3,000 sentences, we use 50% for training, 25% for tune, and 25% for dev, for larger datasets we use 750 sentences for dev and tune, and the rest for train. We limit the size of each dataset to 20,000.

We compare models with dataset embeddings to baselines where we use the exact same parser (Figure 1), but without enabling any dataset emebddings. We use two training setups for our baselines: 1) monolingual models (MONO) and 2) models where all treebanks are concatenated (CONCAT). This is in contrast to Smith et al. (2018) who only compared to a monolingual baseline. We test this on the 59 test sets that are part of a cluster.[2] Furthermore, we also explore what happens if we train one parser on all the datasets simultaneously, similar to Udify (Kondratyuk and Straka, 2019), who do not use dataset embeddings in their setup.

4 Evaluation

4.1 Results

Full results can be found in Table 1. We can see that dataset embeddings still seem to be largely useful, outperforming the monolingual baseline (MONO) in almost all cases (56/59 test sets) and the concatenated baselines (CONCAT) in a majority of cases (40/59). The dataset embedding methods are on average 3 LAS points above the MONO baseline. These gains are even larger than the gains observed in Smith et al. (2018) which indicates that dataset embeddings are still relevant when using large pretrained multilingual LMs. It should be noted though, that a large portion of this gain is already present when using the CONCAT strategy. ENCODER scores highest overall, but for some clusters, DECODER is competetive (af-de-nl, es-ca, it, old, sw-sla). These clusters have in common that they are small and/or contain relatively high-resource languages (which are probably better represented in the mBERT embeddings).

For treebanks from languages not used during mBERT pretraining, scores are very low for the MONO baseline, but they gain a lot from training on other languages. Kazakh also has very low scores, this is probably because the training split is very small. It gains a lot already in CONCAT, likely because Turkish is a related language, but then loses accuracy when using dataset embeddings compared to CONCAT, likely because there is not enough in-language data to learn an accurate dataset embedding.

4.2 Results on Subsets

To find trends in our results, we report scores over different subsets of the data. In Figure 2, we report the results when training a parser for each cluster. The leftmost part of the graph shows the scores averaged over all datasets. `+sameLang` are the average scores for all datasets for which a dataset in the

[2]Smith et al. (2018) trained mono-treebank models for the remaining test sets

			Trained on clusters				Trained on all			
Cluster	Treebank	MONO	CONCAT	DECODER	ENCODER	BOTH	CONCAT	DECODER	ENCODER	BOTH
af-de-nl	af_afribooms	80.63	82.12	82.62	81.17	80.93	82.98	**83.53**	82.62	82.37
	nl_alpino	92.75	93.01	92.97	92.50	**93.15**	92.65	92.59	93.08	92.76
	nl_lassysmall	87.25	89.71	89.91	89.59	89.69	89.45	**90.29**	90.11	90.05
	de_gsd	87.02	87.60	87.52	87.09	87.31	87.63	**87.63**	87.25	87.53
e-sla	ru_syntagrus	**94.75**	94.74	94.56	94.60	94.67	94.50	94.45	94.51	94.52
	ru_taiga	74.93	74.54	75.11	75.75	76.21	75.43	75.95	**76.46**	76.45
	uk_iu	88.56	89.96	89.97	89.73	89.56	90.28	**90.67**	90.58	90.26
en	en_ewt	89.34	88.90	89.06	**89.67**	89.12	88.49	88.16	88.86	88.94
	en_gum	90.38	89.06	90.17	90.74	**90.91**	88.64	89.65	90.53	90.71
	en_lines	86.57	84.72	84.77	87.00	87.38	84.70	85.08	87.41	**87.42**
es-ca	ca_ancora	93.33	93.66	93.56	93.42	**93.76**	93.45	93.43	93.46	93.44
	es_ancora	92.99	93.08	**93.38**	93.30	93.30	93.35	93.28	93.24	93.16
finno	et_edt	**85.61**	84.66	85.29	85.11	85.33	85.17	84.98	85.58	85.58
	fi_ftb	89.03	82.22	89.53	**90.14**	89.87	80.15	89.01	89.06	89.38
	fi_tdt	88.55	82.59	88.99	88.90	89.35	84.24	89.13	**89.44**	89.39
	sme_giella*	49.86	62.35	62.74	62.96	64.71	59.55	59.55	63.95	**65.28**
fr	fr_gsd	94.45	**94.62**	94.40	94.57	94.46	94.49	94.42	94.44	94.54
	fr_sequoia	88.55	85.85	87.86	91.32	90.91	86.17	86.46	91.68	**91.73**
	fr_spoken	79.15	83.80	83.33	84.30	83.89	83.47	83.20	84.09	**84.50**
indic	hi_hdtb	92.54	92.47	92.56	**92.66**	92.37	92.49	92.37	92.51	92.34
	ur_udtb	81.02	81.78	81.93	**82.01**	81.74	81.66	81.41	81.77	81.42
iranian	kmr_mg*	21.43	14.29	16.67	**33.33**	30.95	21.43	19.05	28.57	14.29
	fa_seraji	**87.01**	86.96	86.84	86.37	86.28	86.32	86.30	86.54	86.27
it	it_isdt	92.92	93.33	93.22	92.96	93.09	**93.44**	93.26	93.24	93.29
	it_postwita	79.66	80.00	80.34	80.34	80.26	79.96	80.08	**80.72**	80.16
ko	ko_gsd	**82.80**	71.46	79.63	82.71	82.24	69.41	79.13	82.55	82.77
	ko_kaist	87.05	81.52	86.52	**87.66**	87.54	81.12	85.81	87.07	86.86
n-ger	da_ddt	86.24	86.37	86.32	**87.05**	86.50	85.90	85.30	86.41	86.31
	no_bokmaal	93.91	94.18	**94.77**	94.31	94.26	94.17	94.28	94.15	94.30
	no_nynorsk	92.57	92.85	93.21	**93.30**	92.93	92.73	92.47	92.93	93.15
	no_nynorsklia	74.23	76.72	77.18	76.94	77.26	76.76	76.85	**77.39**	77.18
	sv_lines	84.93	85.07	85.43	86.03	86.05	85.15	85.30	**86.19**	**86.19**
	sv_talbanken	83.85	84.30	85.32	85.85	85.90	85.18	85.41	**86.64**	86.14
old	grc_proiel	75.57	77.37	77.26	**77.46**	76.91	76.76	76.70	76.30	76.47
	grc_perseus*	60.48	**65.13**	64.86	64.17	64.35	64.52	64.41	63.93	64.29
	got_proiel*	72.74	80.36	**80.75**	79.87	79.91	78.97	78.72	78.72	79.50
	la_ittb	90.08	89.98	90.17	89.93	89.62	**90.30**	90.13	90.08	90.16
	la_proiel	77.59	79.56	79.76	**79.99**	78.84	79.07	78.96	79.41	79.13
	la_perseus	56.28	65.34	65.63	70.07	69.72	64.71	65.40	**71.56**	70.05
	cu_proiel*	61.24	64.48	**65.22**	64.88	63.91	64.11	64.42	64.17	64.93
pt-gl	gl_ctg	81.16	80.92	80.94	**81.70**	81.38	80.93	80.89	81.50	81.56
	gl_treegal	70.80	65.73	75.01	81.07	82.11	64.60	68.93	**82.49**	82.28
	pt_bosque	90.45	90.41	90.43	90.52	**90.62**	90.49	90.28	90.44	90.09
sw-sla	hr_set	88.20	88.33	88.43	88.39	88.17	88.35	88.73	**88.84**	88.74
	sr_set	87.20	87.78	88.89	88.94	88.88	88.31	89.16	89.47	**89.48**
	sl_ssj	93.22	**93.54**	93.29	93.39	93.44	93.37	93.30	93.40	93.29
	sl_sst	60.30	70.67	70.78	70.40	70.28	70.31	70.84	**71.22**	71.09
turkic	bxr_bdt*	19.51	26.83	31.71	**36.59**	21.95	26.83	21.95	29.27	31.71
	kk_ktb	14.02	**62.62**	58.88	48.60	49.53	59.81	61.68	51.40	49.53
	tr_imst	65.57	66.02	65.66	64.48	64.86	**66.29**	66.03	65.99	66.00
	ug_udt*	47.85	49.11	49.01	49.18	48.68	**50.39**	49.86	50.06	49.91
w-sla	cs_cac	91.86	92.24	92.19	92.20	92.18	92.12	92.22	92.09	**92.40**
	cs_fictree	92.97	94.01	94.23	94.23	94.41	94.11	94.35	**94.52**	94.10
	cs_pdt	89.57	90.39	90.57	90.45	90.42	90.32	90.34	90.89	**90.89**
	pl_lfg	95.41	93.30	96.17	**96.49**	96.40	92.87	96.19	96.43	96.43
	pl_pdb	91.64	91.29	91.82	91.72	91.72	91.74	91.85	91.72	**91.94**
	sk_snk	92.07	93.62	93.51	**93.77**	93.58	93.13	93.41	93.29	93.02
	hsb_ufal*	14.47	59.21	60.53	63.16	59.21	61.84	**65.79**	61.84	63.16
avg.		78.52	80.63	81.58	**82.16**	81.77	80.60	81.26	**82.10**	81.88

Table 1: LAS scores for each dataset (dev) for all of our settings, both when training a parser per cluster ("Trained on cluster"), as well as having one parser for all treebanks ("Trained on all"). Bold: highest score for this training setup, omitted if the MONO baseline performs best. * not used in mBERT pretraining.

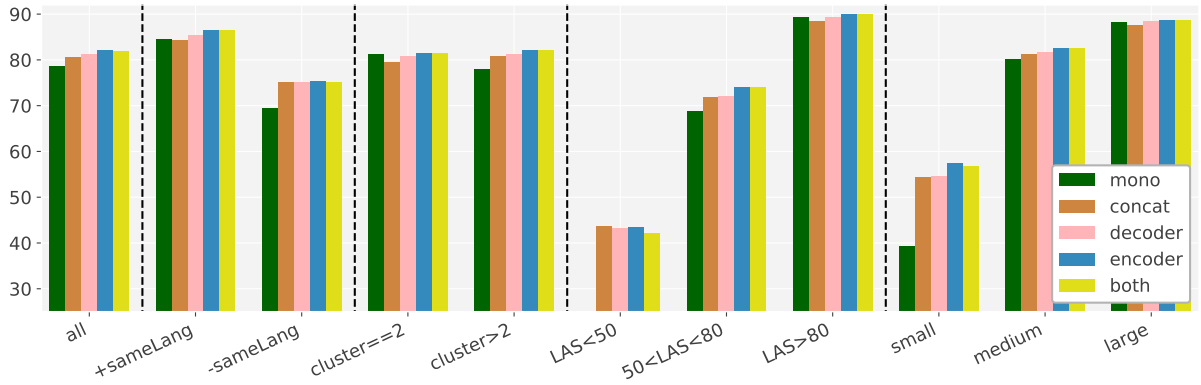


Figure 3: Scores for same subsets as Figure 2, but when training a single parser on all data at once, instead of separate parsers for each cluster.

same language is included in the cluster, and `-sameLang` for all datasets for which this is not the case. `cluster==2`, are scores for clusters consisting of 2 datasets, and `cluster>2` for larger clusters. We also divide the datasets based on their performance with the MONO baseline (LAS<50, 50<LAS<80, LAS>80), and finally based on the size of the training data: small (<1,000 sentences), medium (<10,000) and large (>20,000).

Dataset embeddings are especially beneficial for datasets where performance of the MONO baseline is low (`LAS<50`) and for small datasets. They help moderately for medium sized datasets, and for datasets where performance of MONO is mediocre (`50<LAS<80`). For larger and high-performing datasets, performance increases diminish. The CONCAT baseline outperforms the MONO baseline in most setups, except for small clusters indicating that some of the gains from using the dataset embedding methods in these settings are due to the additional use of data.

Perhaps a bit counterintuively, the dataset embeddings methods improve results more for treebanks for which there is not a treebank of the same language (`-sameLang`). However, this may very well be due to a confounding factor: the scores are generally a lot higher in the `+sameLang` setting than in the `-sameLang` setting and the method seems to work better for treebanks for which the baseline scores are lower.

Overall, the ENCODER model performs best; it either outperforms all others, or performs on par with the best setup. The ENCODER model outperforms the other models mostly on datasets where the MONO baseline is low, and for small datasets. The DECODER strategy is only beneficial in some of the data subsets, and should be used with caution. Perhaps surprisingly, the BOTH strategy is not beneficial over the ENCODER, indicating that both strategies encode dataset information differently.

In Figure 3, we report the results when training one parser on all the treebanks for each setup. Results are very similar to results with the parsers trained per cluster (Figure 2). The main difference can be observed for the datasets with low performance of the MONO baseline (LAS<50), where the difference between the different dataset embeddings and CONCAT is smaller. The similarity to results obtained with smaller clusters indicates that 1) using dataset embeddings is also viable in a highly multilingual model and 2) a highly multilingual model might be able to pick up on dataset similarities and use the relevant data for individual languages. This has practical implications: it can be more practical to have one model that works on many languages (Kondratyuk and Straka, 2019) than multiple models and it removes the need to carefully construct clusters of related languages, which can be time-consuming without a guarantee for optimal clusters.

4.3 Test data

To avoid overusing the test data, we only confirm our main findings on the test data; we compare our best baseline to our best dataset embedding setup for both the setup trained in clusters and when training on all datasets simultaneously. Altough van der Goot (2021) suggests to concatenate the tune set to the

Clusters		All	
CONCAT	ENCODER	CONCAT	ENCODER
81.08	**82.05**	80.57	**82.28**

Table 2: Average LAS scores on the test data from our best baseline (CONCAT), and the best setup with dataset embeddings (ENCODER) for both the cluster trained-parser and the parser trained on all data.

training data for the final comparison on test, we only use the training split to save compute and because we are not aiming for a new state-of-the-art.

Results (Table 2) show that for parsers trained on clusters the CONCAT baseline performs a bit higher compared to the development data (Table 1), and the gain is thus smaller, but still substantial. When training on all datasets, the results are similar. Overall, the results on the test data confirm the findings on the development data: 1) dataset embeddings are useful in both setups, 2) one large multilingual model trained on all treebanks performs on par with multiple parsers trained on clusters of related treebanks.

5 Conclusion

We evaluated the usefulness of dataset embeddings for multilingual parsing using large pretrained multilingual LMs and found them to be useful in this context. Using this method improves over both a monolingual baseline and a baseline where training treebanks are concatenated, and across many settings. This method helps mostly for small treebanks. Using dataset embeddings in the encoder showed overall slightly better results than our other embedding strategies, even better than combining the two approaches to use dataset embeddings. Finally, we found that using dataset embeddings in a multilingual parser that uses training data from all available treebanks works just as well as using them with clusters of treebanks from related languages.

Acknowledgements

We would like to thank the anonymous reviewers, Ahmet Üstün, Max Müller-Eberstein and Daniel Varab for their discussions about dataset embeddings and evaluation. Miryam de Lhoneux was funded by the Swedish Research Council (Grant 2020-00437).

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Timothy Dozat and Christopher Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *Proceedings of the 5th International Conference on Learning Representations.*

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Process-*

ing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, 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.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A. Alonso. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–431, Berlin, Germany. Association for Computational Linguistics.

Joachim Wagner, James Barry, and Jennifer Foster. 2020. Treebank embedding vectors for out-of-domain dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8812–8818, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

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.

A Exact Scores

filter	mono	concat	decoder	encoder	both	ALLconc.	ALLdec.	ALLenc.	ALLboth
Avg.-58	78.50	80.71	81.45	81.96	81.82	80.57	81.20	**82.03**	81.94
hasSameLang-35	84.44	84.36	85.62	86.36	86.34	84.17	85.26	**86.48**	86.48
noSameLang-23	69.46	75.17	75.11	75.26	74.95	75.09	75.03	**75.26**	75.03
cluster=2-10	81.15	79.55	80.56	**82.31**	82.11	79.52	80.71	81.51	81.33
cluster>2-48	77.94	80.96	81.64	81.89	81.76	80.79	81.30	**82.14**	82.07
low-5	22.66	43.36	42.57	43.54	42.23	**43.69**	43.14	43.38	42.20
medium-14	68.68	72.26	72.97	73.87	73.84	71.77	72.04	74.00	**74.04**
high-39	89.18	88.54	89.48	89.79	89.77	88.46	89.37	89.87	**89.87**
small-7	39.26	54.45	55.15	57.35	56.45	54.37	54.55	**57.48**	56.71
medium-27	80.15	81.37	81.90	82.36	82.32	81.18	81.70	82.49	**82.50**
large-24	88.09	87.64	88.62	**88.68**	88.66	87.52	88.42	88.67	88.67

Table 3: Exact numbers for results in Figure 2. Average LAS scores over different subsets of the data. All: all data, +sameLang: datasets for which another in-language treebank exists, cluster==2: clusters of size 2, LAS<50: treebanks for which the 'mono´ baseline scores <50 LAS, small, medium, large: datasets with a maximum size of respectively: 1,000, 10,000, 20,000 sentences.

Discourse Tree Structure and Dependency Distance in EFL Writing

Jingting Yuan, Qiuhan Lin, John S. Y. Lee
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
{jingtyuan2-c,qiuhanlin2-c}@my.cityu.edu.hk
jsylee@cityu.edu.hk

Abstract

Quantitative research on learner writing has traditionally focused on lexical and syntactic features, but there has been increasing interest in incorporating discourse-level properties. This paper evaluates discourse complexity measures on learner texts in the framework of Rhetorical Structure Theory (RST). Specifically, we investigate whether discourse dependency distance and embedded structures in RST trees are correlated to learner proficiency level. In an analysis of manually annotated English essays, we found that more proficient learners tend to use longer dependency distance and more embedded structures. Further, an evaluation based on automatic discourse parsing suggests that dependency distance can potentially contribute to automatic assessment of learner texts.

1 Introduction

Text complexity depends on linguistic characteristics of the text at various levels, including lexical, syntactic and discourse features. Earlier research on text complexity mostly focused on surface features such as word length and sentence length (Kincaid et al., 1975) and n-grams (Schwarm and Ostendorf, 2005). Corpus development for discourse structure (Carlson et al., 2002; Prasad et al., 2008) has facilitated investigation into a variety of features related to discourse organization, including text cohesion, coherence and distribution of discourse relations (Lee et al., 2006; Pitler and Nenkova, 2008; Sun and Xiong, 2019). Some studies have found coherence features to be more highly correlated to text complexity than other types of features (Davoodi and Kosseim, 2016).

Text complexity is also closely related to research on language acquisition and interlanguage. Lexical and syntactic features in learner writing have been extensively studied (Jiang et al., 2019; Lu, 2011). Overuse and underuse of rhetorical relations in learner texts, and distinctive discourse structures, have also been identified (Skoufaki, 2009; Brown, 2019). Further, text complexity models can offer feedback for essay revision (Burstein et al., 2003) and support automatic assessment in language learning (Lyashevskaya et al., 2021). For example, discourse connectives and relations can help predict the level of learner proficiency (Rysová et al., 2016), and RST tree patterns can improve the performance of a speech scoring system for proficiency assessment (Wang et al., 2019).

This paper studies two types of discourse features derived from RST dependency trees. The first, *dependency distance*, refers to the linear distance between a discourse unit and its head in the dependency tree (Sun and Xiong, 2019). Second, we examine the usage of an *embedded structure* in which an elementary discourse unit (EDU) governs both its left and right neighbors, and also serves as the dependent of another discourse unit. To the best of our knowledge, these features have not yet been evaluated on their correlation to learner proficiency level. Using a corpus of texts written by learners of English as a Foreign Language (EFL), we investigate whether texts written by more proficient learners exhibit longer discourse dependency distance, and contain more embedded structures.

The rest of the paper is organized as follows. After summarizing previous work (Section 2), we define the proposed discourse complexity measures (Section 3), followed by a description of our dataset and its conversion to dependency trees (Section 4). We then present the results and analyze the correlation between these discourse complexity measures and learner proficiency level (Section 5). Finally, we conclude with a summary of our findings and suggestions for future work (Section 6).

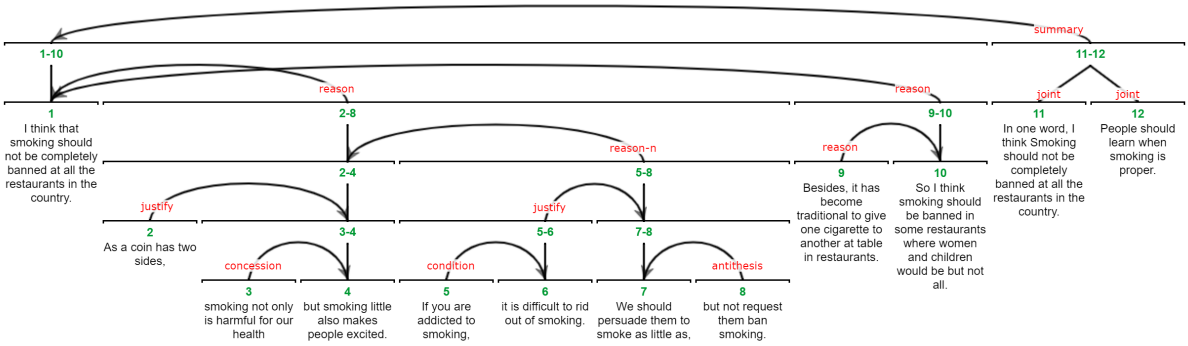


Figure 1: A manually annotated RST tree from our dataset (Section 4.2)

2 Previous work

All parts of a coherent text should be held together with appropriate discourse relations. Judicious use of discourse connectives, including their frequency and diversity, as well as the distribution of discourse relations, have been found useful in predicting essay quality, the proficiency of language learners, and language development of native speakers (Crossley et al., 2016; Rysová et al., 2016; Weiss and Meurers, 2019). Davoodi and Kosseim (2016) applied features based on discourse relations and their realizations to assess the text complexity in datasets derived from the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and Simple English Wikipedia. In both datasets, coherence features were shown to be more correlated to text complexity than lexical, syntactic and other surface features.

While the PDTB focuses on text spans that explicitly or implicitly serve as arguments of a discourse connective, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) models the full hierarchical structure of a text. According to RST, the discourse organization of a text can be represented by a constituent tree whose leaves are the elementary discourse units (EDUs). Two adjacent EDUs or text spans can be combined into a longer span with a rhetorical relation. Each EDU or span is either assigned to be the nucleus, which contains the more essential information; or the satellite, which provides supporting information. Figure 1 shows an example tree. The RST Discourse Treebank (RST-DT) (Carlson et al., 2002), which consists of 385 Wall Street Journal articles, has formed the basis of quantitative research in this framework.

Texts written by native and non-native speakers have been compared in terms of the use of RST discourse relations. In a corpus of English essays, relations concerned with ideation and/or content, called the "subject matter" relations (Brown, 2019), have been found to be used less frequently by native speakers of Japanese compared to native speakers of English. Since RST-style discourse parsing can better capture long-distance discourse dependencies, it has been found to outperform PDTB-style parsing on coherence assessment tasks such as essay scoring (Feng et al., 2014). RST tree-based features, such as tree depth, the ratio between the depth and the number of EDUs, and the frequency of different types of rhetorical relations, have been applied to proficiency assessment in the context of a speech scoring system (Wang et al., 2017; Wang et al., 2019).

To facilitate natural language processing tasks, algorithms have been proposed to convert RST constituent trees to dependency trees (Li et al., 2014). Adopting this methodology, Sun and Xiong (2019) constructed a dependency treebank from the RST-DT, examined text complexity in terms of dependency distance, and analyzed differences in discourse distance among various rhetorical relations.

ID	Elementary discourse unit (EDU)	Head	Relation	Dist.
1	We can't only focus on the working experience of a student	2	antithesis	1
2	Instead, the knowledge in professional field is more important	4	reason	2
3	when we are still students.	2	circumstance	1
4	So we should try to pay more attention to our study now.	n/a	root	0

Table 1: Dependency distance of the discourse relations in an extract from a learner essay

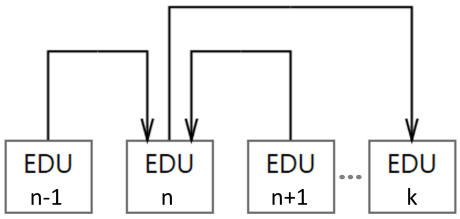


Figure 2: An embedded structure in an RST dependency tree, as defined in Section 3.2

3 Discourse complexity measures

3.1 Dependency distance

A syntactic dependency tree specifies grammatical relations between the words in a sentence. Each relation in the tree involves a governor (head) and a dependent, which modifies the head. In this context, *dependency distance* is defined as the number of words between the dependent and its head. Dependency distance can serve as a metric for syntactic complexity and language comprehension difficulty (Gibson, 1998; Liu, 2008), given its relation to cognitive load in linguistic processing (Fedorenko et al., 2013).

Following Sun and Xiong (2019), we apply the concept of dependency distance on RST trees. We define the *dependency distance* of an RST relation to be the number of EDUs between the dependent EDU and its head. Table 1 shows the dependency distance for each discourse relation in a sample text. The mean dependency distance (MDD) in an RST dependency tree can be calculated as:

$$MDD = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \tag{1}$$

where n is the total number of EDUs; s is the total number of texts; and DD_i is the dependency distance of the i^{th} dependency link of the text.[1] We will henceforth use the term MDD in the context of the discourse dependency tree, rather than its syntactic counterpart.

We hypothesize that MDD is indicative of learner proficiency. As will be shown in our manually annotated dataset, almost 60% of the discourse relations have a dependency distance of 1, and 15% have a distance of 2 (Figure 7). We will therefore compute the proportion of relations whose dependency distance is at least 3 (to be referred to as "**length-3 relations**") and at least 4 ("**length-4 relations**"), in addition to MDD.

3.2 Embedded structures

Two major types of dependency patterns have been identified in the PDTB: a pair of discourse relations may be "independent relations", or have "full embeddings" (Lee et al., 2006). A "fully embedded" structure consists of a discourse relation that is entirely realized as an argument of another discourse connective.

Similar structural patterns in RST dependency trees could potentially help characterize learner writing. As a preliminary investigation, we examine the subtree pattern illustrated in Figure 2.[2] The pattern contains a sequence of three EDUs where the middle unit (EDU n) governs both its left and right neighbor

[1]The MDD of the sample text in Table 1 is calculated as $(|2-1|+|4-2|+|2-3|)/(4-1) = 1.3$.

[2]Dependency tree diagrams in this paper are produced with Dependency Viewer developed by the Natural Language Processing Group at Nanjing University, China (http://nlp.nju.edu.cn/tanggc/tools/DependencyViewer_en.html).

Proficiency level	Essay Topic		# tokens	# tokens per text (SD)	# EDUs	# EDUs per text (SD)
	"Part-time job"	"Smoking"				
Native	9	9	4,045	224.7 (16.5)	352	19.6 (3.5)
B2+	9	9	4,221	234.5 (34.1)	368	20.4 (2.0)
B1	9	9	3,991	221.7 (23.2)	371	20.6 (3.4)
A2	9	9	3,820	212.2 (8.4)	388	21.6 (2.6)
Total	72		16,077	223.3 (23.5)	1,479	20.5 (3.0)

Table 2: Our dataset contains a total of 72 essays written by learners at three different proficiency levels and by native speakers.

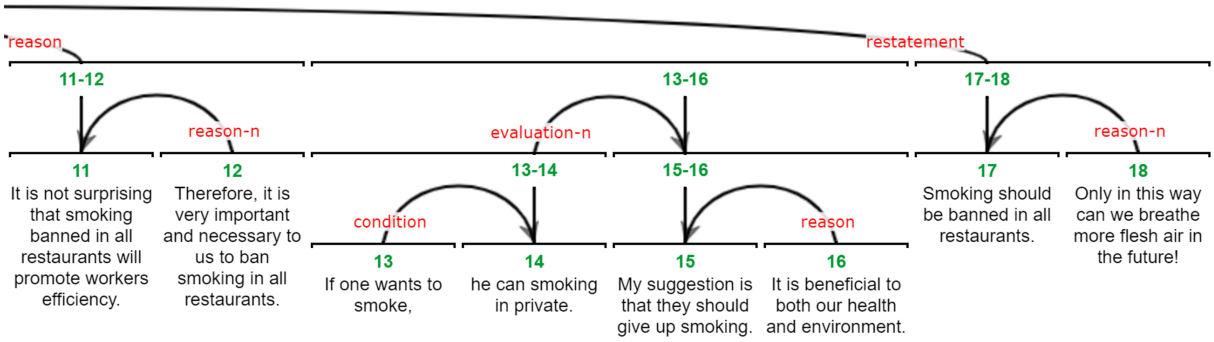


Figure 3: Text span 13-16 forms a dangling structure in the RST tree (Section 4.2)

(EDU $n-1$ and EDU $n+1$); in addition, the middle unit itself serves as a dependent of another unit (EDU k) that may precede or follow it. The text extract in Table 1 is an instance of such a structure. EDU 2 governs both EDU 1 and EDU 3, with the relations `antithesis` and `circumstance`, respectively. Further, EDU 2 is the dependent of EDU 4 in the `reason` relation, contributing to the writer's opinion expressed therein.

We will henceforth use the term "embedded structure" to refer to the specific pattern in Figure 2. We hypothesize that the complex discourse organization in this structure is indicative of learner proficiency, and will calculate the proportion of EDUs in a text that exhibit this embedded structure.

4 Data

We first present the textual material of our dataset (Section 4.1). We then describe the annotation guidelines (Section 4.2) and report inter-annotator agreement (Section 4.3).

4.1 Textual material

Our dataset consists of written essays drawn from the *International Corpus Network of Asian Learners of English* (ICNALE) (Ishikawa, 2013).[3] The corpus identifies the proficiency level of each writer according to the *Common European Framework of Reference for Languages* (CEFR, 2001). We randomly selected nine writers from the subcorpus of Chinese EFL learners at the A2, B1 and B2+ levels.[4] We also randomly selected nine native speakers of English to serve as control.

To facilitate a direct comparison, we selected one essay by each writer on the topic "Part-time job", and one essay on "Smoking", such that all essays had similar length. The final dataset contains 72 essays from 36 writers spanning four proficiency levels, with a total of 16,077 tokens (Table 2).

4.2 Annotation guidelines

We annotated the RST structure of each essay in our dataset according to the guidelines from Stede et al. (2017), which have been adopted by a variety of corpora (Das and Stede, 2018; Musi et al., 2018). The

[3]Downloaded from http://language.sakura.ne.jp/icnale/

[4]B2+ is defined as level B2 or above. The A1 level is not available.

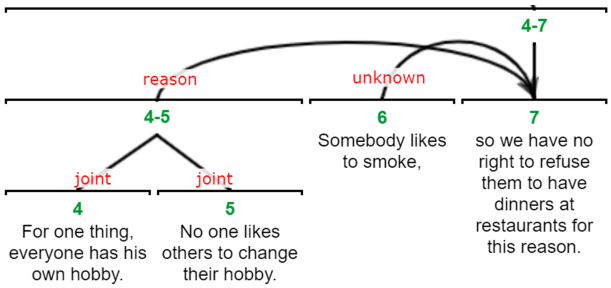
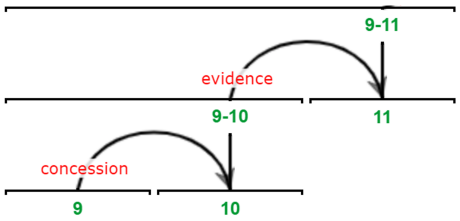
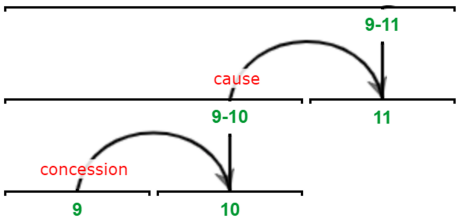


Figure 4: Example learner text annotated with an `unknown` relation



(a) Annotation with the `evidence` relation



(b) Annotation with the `cause` relation

Figure 5: Two interpretations for the discourse relation between text spans 9-10 and 11 in the following text: [No matter how hard they try,][9] [computer science textbook publishers cannot keep up with the rapid development of new computer hardware, software, and programming languages.][10] [C.S. student who wish to maintain a competitive edge in the job market should find part-time positions where they can learn state-of-the-art techniques.][11]

guidelines include 31 rhetorical relations, belonging to the pragmatic, semantic, textual and multinuclear categories. Figure 1 shows an annotated RST tree from our dataset.[5]

Since learner texts are not always grammatical, logical and well-organized, they may contain irrelevant text spans that form "dangling structures", without any link to the rest of the text (Skoufaki, 2009). Figure 3 provides such an example with the discourse span 13-16, which gives reasons for quitting smoking. The neighboring discourse spans 11-12 and 17-18, which argue why restaurants should ban smoking, have no apparent relation to this span.

The relation between two text spans is labeled as `unknown` when it is unclear or cannot be readily understood. Consider the text spans 6 and 7 in Figure 4. Although the connective "so" in span 7 signals a causality between these two spans, the sentence "Somebody likes to smoke" does not appear to be directly relevant to the opinion expressed in span 7, hence the `unknown` label.

4.3 Inter-annotator agreement

Two annotators, both with academic background in linguistics, performed the annotation using RSTTool version 3.0 (O'Donnell, 2000). To measure inter-annotator agreement, they double-annotated five texts written by learners and five written by native speakers in our dataset. Since paragraph boundaries are not marked in the raw text, the main topical units of a text can often be open to multiple interpretations. As an initial step in the annotation of each essay, the two annotators discussed its segmentation into elementary discourse units (EDUs) and larger text spans. After reaching an agreement, the annotators independently labeled the nuclearity of each span or text span, and assigned the rhetorical relations. There were a total of 94 EDUs in the native texts and 112 EDUs in the learner texts.

The two annotators achieved a Cohen's Kappa of 0.95 for nuclearity (i.e., assignment of nucleus vs. satellite) and 0.92 for relation (i.e., assignment of the rhetorical relation) on the native texts. The Kappa was slightly lower for the learner texts, at 0.94 for nuclearity and 0.86 for relation, likely reflecting the more ambiguous language therein. For relation assignment, a majority of the discrepancies (69%)

[5]RST tree diagrams in this paper are produced with rstWeb (Zeldes, 2016).

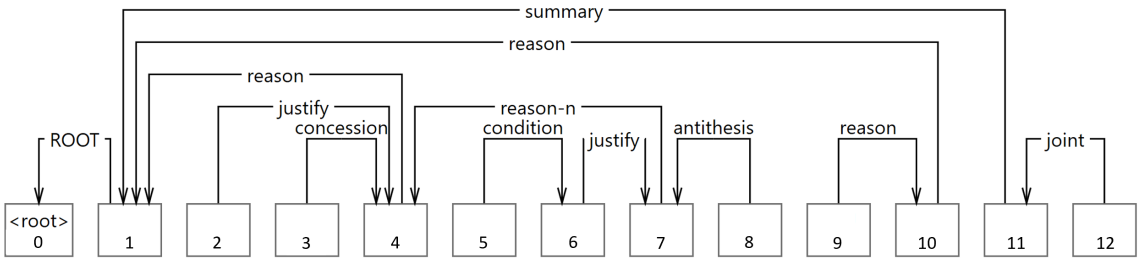


Figure 6: RST dependency tree automatically derived from the RST constituent tree in Figure 1

Complexity measure	Proficiency level			
	A2	B1	B2+	Native
Mean dependency distance	2.23	2.66*	2.97*	**3.12**
% length-3 relations	22.8	26.9	**27.8**	27.3
% length-4 relations	12.5	17.9*	20.1	**22.2**
% embedded structures	8.2	7.5	8.1	**9.5**

Table 3: Discourse complexity measures on the manually annotated dataset at different levels of language proficiency. An asterisk means the difference with the figure to its left is statistically significant.

between the two annotators occurred in the non-terminal text spans of the tree. Figure 5 shows two possible interpretations of a text. Both annotators agreed on text span 11 as nucleus and text span 9-10 as a satellite. However, one annotator regarded span 9-10 as providing `evidence` to the subjective claim in span 11 (Figure 5a), while the other saw span 9-10 as describing the `cause` of the objective state in span 11 (Figure 5b).

The RST trees for the remaining essays in our corpus were derived by one of the annotators. After manual annotation, we automatically converted the RST constituent trees into dependency trees (Li et al., 2014).[6] The RST constituent tree in Figure 1, for example, was converted to the dependency tree in Figure 6.

5 Analysis

We analyze the extent to which learner proficiency is correlated to the discourse complexity measures proposed in Section 3. Tables 3 and 4 show the results for the manually annotated and automatically parsed datasets, respectively, with a breakdown into different proficiency levels.

5.1 Manually annotated dataset

Mean Dependency Distance (MDD). MDD appears correlated to proficiency: it increases from 2.23 for A2 writers to 2.66 for B1 writers[7], then 2.97 for B2+ writers[8], then finally reaches the highest value with native speakers, at 3.12 (Table 3). A possible explanation is that, since less proficient writers need to dedicate more cognitive resources for linguistic processing, including retrieval of unfamiliar words and grammar checking, they lack the resources to produce more complex discourse (Yan and Li, 2019). The native speakers' MDD in our corpus is slightly shorter than the 3.18 observed in RST-DT (Sun and Xiong, 2019), perhaps reflecting the more formal register of news material.

Long-distance discourse relations. Texts produced by more proficient learners tend to contain more long-distance relations. In terms of the proportion of length-4 relations (see definition in Section 3.1), there is an upward trend from the A2 writers (12.5%), B1 writers (17.9%), B2+ writers (20.1%)[9], and to

[6]We used the conversion tool provided at https://github.com/amir-zeldes/rst2dep

[7]The difference between B1 and A2 is statistically significant at $p = 0.011$ by t-test.

[8]The difference between B2 and B1 is statistically significant at $p = 0.043$ by t-test.

[9]The difference between B2+ and B1 is not statistically significant ($p = 0.481$ by chi-squared test), but it is statistically significant between B2+ and A2 ($p = 0.006$).

Complexity measure	Proficiency level			
	A2	B1	B2+	Native
Mean dependency distance	1.49	1.59*	1.58	**1.74***
% length-3 relations	13.2	14.0	15.7	**18.7**
% length-4 relations	4.1	5.8	6.9	**11.5***
% embedded structures	7.5	**9.5**	8.4	9.1

Table 4: Discourse complexity measures on the automatically annotated dataset at different levels of language proficiency. An asterisk means the difference with the figure to its left is statistically significant.

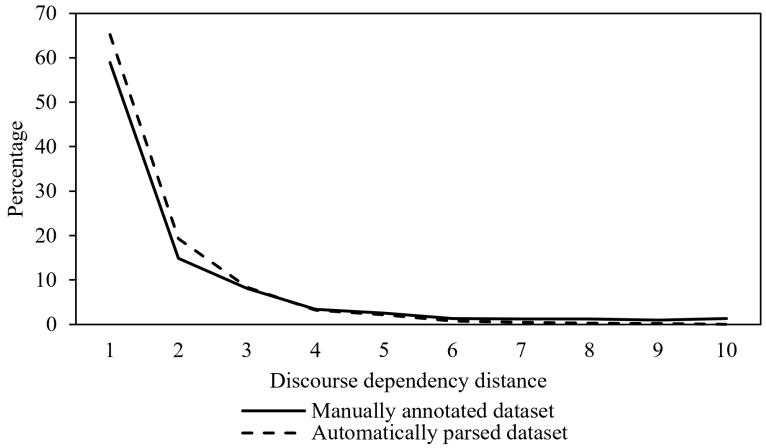


Figure 7: Distribution of discourse dependency distance in our dataset

the native speakers (22.2%) (Table 3).[10] A similar correlation is observed among the learners in terms of length-3 relations, whose proportion progresses from the A2 writers (22.8%), B1 writers (26.9%) to the B2 writers (27.8%). The differences between these groups are not statistically significant, however, perhaps in part due to the small size of the dataset. The native speakers use a higher proportion of length-3 relations (27.3%) than A2 and B1 but, contrary to expectation, their proportion is slightly lower than B2.

Embedded structures. We now turn to the embedded structures in RST dependency trees as defined in Section 3.2. The proportion of these structures shows an upward trend among the intermediate learners from B1 (7.5%) to B2+ (8.1%), and attains even higher frequency among the native speakers (9.5%). However, unexpectedly, these structures are more common in the A2 texts (8.2%) than their B1 and B2+ counterparts in our dataset. A larger dataset would be needed to clarify the correlation between the usage of this structure to language proficiency among learners.

5.2 Automatically parsed dataset

To gauge the effectiveness of the proposed features in text complexity assessment, we automatically parsed all essays in our dataset with the RST Extractor (Koto et al., 2019)[11], which implemented a state-of-the-art neural discourse parsing algorithm (Yu et al., 2018). We then converted the parser output to the dependency format, using the same procedure for the manually annotated dataset. Figure 7 compares the distribution of dependency distance between the dependency trees derived from the manually annotated and automatically parsed constituent trees.

Mean dependency distance (MDD). Across all proficiency levels, the MDD is lower in the automatically parsed dataset (Table 4) than in the manually annotated dataset (Table 3), suggesting difficulties for the automatic parser to identify long-distance relations. Table 5 shows the MDD of various discourse relations. The parser identified much fewer `evidence` and `summary` relations, for example, both of

[10]The difference between Native and A2 is statistically significant ($p = 0.001$ by chi-squared test).

[11]Downloaded from https://github.com/fajri91/RSTExtractor

Discourse Relation	Overall	Proficiency level				Frequency	
		A2	B1	B2+	Native	Auto	Manual
means	1.07	1.00	**1.25**	1.00	1.00	14	15
condition	1.08	1.00	1.00	1.05	**1.19**	65	79
cause	1.38	1.27	1.20	**1.85**	1.19	10	50
evaluation	2.00	1.50	2.00	**2.17**	1.83	13	36
contrast	2.82	1.83	3.33	**4.38**	2.11	19	29
evidence	3.34	2.78	3.12	2.37	**5.85**	4	67
summary	13.26	12.56	11.54	13.08	**15.83**	0	46

Table 5: Mean dependency distance (MDD) of selected RST discourse relations in the manually annotated corpus; and the frequency of these relations in the manually annotated ('Manual') and automatically produced ('Auto') datasets

which have relatively long MDD in manual annotation. The summary relation exhibits by far the longest MDD, at 13.26, which is comparable with its MDD of 9.34 in the RST-DT (Sun and Xiong, 2019).

MDD statistics in the automatically parsed dataset still largely demonstrate correlation to the proficiency level of EFL learners. MDD is shortest for the A2 writers (1.49), increases with the B-level writers (1.59 B1, 1.58 B2+) and reaches the largest value for native speakers (1.74).[12] However, the MDD of B1 and B2+ are indistinguishable. Hence, with automatic discourse parsing, MDD appears robust in discriminating between the beginner, intermediate and native texts, but not between subcategories at the intermediate level.

Long-distance discourse relations. Given the lower MDD in the automatically parsed dataset, the proportion of long-distance discourse relations is also lower. The correlation to the proficiency level continues to hold: the A2 writers used only 4.1% of length-4 relations; the B1 writers, 5.8%; and the B2+ writers, 6.9%[13]; and the native speakers exceeded all learners by far, at 11.5%.[14] Length-3 relations are less effective in distinguishing between the proficiency levels. While the native speakers still yielded a higher proportion (18.7%) than the learners[15], there is no statistically significant difference between A2 (13.2%), B1 (14.0%) and B2+ (15.7%).

Embedded structures. The proportion of embedded structures in the automatically parsed dataset is comparable to the manual version. These structures appear at a rate of 9.1% in native texts, more frequent than in A2 (7.5%) and B2+ (8.4%) texts. Surprisingly, however, the proportion of embedded structures is even higher at the B1 level (9.5%). This unexpected result may reflect the sensitivity of this measure to parser errors. A larger dataset would help determine if discourse parsers are sufficiently robust in detecting embedded structures, and whether the correlation holds among proficiency level and the frequency of these structures.

6 Conclusion

This paper analyzed learner texts according to discourse-level features based on RST dependency trees. Specifically, we investigated whether the dependency distance of discourse relations and the frequency of an embedded structure are correlated to learner proficiency level. Our dataset consists of English essays on similar topics written by native speakers of English, and by native speakers of Chinese at three proficiency levels.

In an analysis of the manually annotated dataset, we found mean dependency distance (MDD) to be significantly higher in texts written by more proficient learners than less proficient ones. Our results also suggested correlation between proficiency level and the proportion of discourse relation of at least

[12]The difference between B1 and A2 is statistically significant ($p = 0.046$ by t-test); so is the difference between B2+ and A2 ($p = 0.033$) and between Native and B2+ ($p = 0.002$).

[13]The difference between B2+ and A2 is statistically significant ($p = 0.043$ by chi-squared test).

[14]The difference between Native and B2+ is statistically significant ($p = 0.010$ by chi-squared test).

[15]The difference is statistically significant between Native and A2 only ($p = 0.021$ by chi-squared test).

length 4, which is significantly higher among native speakers than beginners. Further, native speakers utilized embedded structures more frequently, although the difference with learners did not reach statistical significance. In the automatic setting, despite parsing errors, texts written by beginners, intermediate learners and native speakers could still be differentiated in terms of MDD, suggesting its potential use in automatic text assessment.

In future work, we plan to expand our dataset to verify the effectiveness of the proposed complexity measures. We would also like to explore a wider range of embedded structures, as well as other discourse-level features such as coreference (Kunz et al., 2016).

Acknowledgements

We thank the anonymous reviewers for their insightful comments.

References

Jonathan D. Brown. 2019. *Using Rhetorical Structure Theory for contrastive analysis at the micro and macro levels of discourse: An investigation of Japanese EFL learners' and native-English speakers' writing*. PhD Dissertation, Leiden University.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. *RST Discourse Treebank*. Linguistic Data Consortium.

CEFR. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16.

Debopam Das and Manfred Stede. 2018. Developing the Bangla RST Discourse Treebank. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC)*.

Elnaz Davoodi and Leila Kosseim. 2016. On the Contribution of Discourse Structure on Text Complexity Assessment. In *Proc. 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Evelina Fedorenko, Rebecca Woodburym, and Edward Gibson. 2013. Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, 37(2):378–394.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proc. 25th International Conference on Computational Linguistics (COLING)*.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Shin'ichiro Ishikawa. 2013. The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, 1:91–118.

Jingyang Jiang, Peng Bi, and Haitao Liu. 2019. Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. *Journal of Second Language Writing*, 46:100666.

Peter J. Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. In *Research Branch Report 8–75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2019. Improved Document Modelling with a Neural Discourse Parser. In *Proc. 17th Annual Workshop of the Australasian Language Technology Association (ALTA)*.

Kerstin Kunz, Ekaterina Lapshinova-Koltunski, and Jose Manuel Martínez. 2016. Beyond Identity Coreference: Contrasting Indicators of Textual Coherence in English and German. In *Proc. Workshop on Coreference Resolution Beyond OntoNotes (CORBON)*.

Alan Lee, Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, and Bonnie Webber. 2006. Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex than in Syntax? In *Proc. 5th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level Discourse Dependency Parsing. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Olga Lyashevskaya, Irina Panteleeva, and Olga Vinogradova. 2021. Automated assessment of learner text complexity. *Assessing Writing*, 49:100529.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Elena Musi, Tariq Alhindi, Manfred Stede, Leonard Kriese, Smaranda Muresan, and Andrea Rocci. 2018. A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations. In *Proc. 11th International Conference on Language Resources and Evaluation (LREC)*.

Michael O'Donnell. 2000. RSTTOOL 2.4-A Markup Tool for Rhetorical Structure Theory. In *Proceedings of the First International Conference on Natural Language Generation (INLG)*.

Emily Pitler and Ani Nenkova. 2008. Revisiting Readability: a Unified Framework for Predicting Text Quality. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC)*.

Kateřina Rysová, Magdaléna Rysová, and Jiří Mírovský. 2016. Automatic Evaluation of Surface Coherence in L2 texts in Czech. In *Proc. Conference on Computational Linguistics and Speech Processing (ROCLING)*.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sophia Skoufaki. 2009. An exploratory application of rhetorical structure theory to detect coherence errors in L2 English writing: Possible implications for automated writing evaluation software. *International Journal of Computational Linguistics and Chinese Language Processing: Special Issue in Computer Assisted Language Learning*, 14(2):181–203.

Manfred Stede, Maite Taboada, and Debopam Das. 2017. *Annotation Guidelines for Rhetorical Structure*. Manuscript. University of Potsdam and Simon Fraser University.

Kun Sun and Wenxin Xiong. 2019. A computational model for measuring discourse complexity. *Discourse Studies*, 21(6):690–712.

Xinhao Wang, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2017. Discourse Annotation of Non-native Spontaneous Spoken Responses Using the Rhetorical Structure Theory Framework. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xinhao Wang, Binod Gyawali, James V. Bruno, Hillary R. Molloy, Keelan Evanini, and Klaus Zechner. 2019. Using Rhetorical Structure Theory to Assess Discourse Coherence for Non-native Spontaneous Speech. In *Proceedings of Discourse Relation Parsing and Treebanking (DISRPT2019)*.

Zarah Weiss and Detmar Meurers. 2019. Analyzing Linguistic Complexity and Accuracy in Academic Language Development of German across Elementary and Secondary School. In *Proc. 14th Workshop on Innovative Use of NLP for Building Educational Applications*.

Hengbin Yan and Yinghui Li. 2019. Beyond Length: Investigating Dependency Distance Across L2 Modalities and Proficiency Levels. *Open Linguistics*, 5:601–614.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural rst parsing with implicit syntax features. In *Proc. COLING*.

Amir Zeldes. 2016. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In *Proc. NAACL-HLT 2016 System Demonstrations*.