Relation between Cross-Genre and Cross-Topic Transfer in Dependency Parsing

Vera Danilova[1] **and Sara Stymne**[2]
[1]Department of History of Science and Ideas, Uppsala University, Sweden
[2]Department of Linguistics and Philology, Uppsala University, Sweden
vera.danilova@idehist.uu.se, sara.stymne@lingfil.uu.se

Abstract

Matching genre in training and test data has been shown to improve dependency parsing. However, it is not clear whether the used methods capture only the genre feature. We hypothesize that successful transfer may also depend on topic similarity. Using topic modelling, we assess whether cross-genre transfer in dependency parsing is stable with respect to topic distribution. We show that LAS scores in cross-genre transfer within and across treebanks typically align with topic distances. This indicates that topic is an important explanatory factor for genre transfer.

Keywords: dependency parsing, cross-genre transfer, topic modelling

1. Introduction

It has been shown that genre is a valuable signal for cross-lingual dependency parsing (Stymne, 2020; Müller-Eberstein et al., 2021), particularly for low-resourced languages when no good transfer languages are available.[1] However, these studies suffer from noisy genres from Universal Dependencies (UD) or use automatically generated genre clusters. Another factor that is likely relevant to transfer is topic similarity between data sets. It has not yet been explored how topic interacts with genre for dependency parsing.

Petrenz (2012) investigated the influence of covariances between genre and topic on genre classification. He found that shifts in topic distribution within training and test sets of the same genre affect the performance, and thus recommended removing topical features, for the task of genre classification. For the dependency parsing task, clustering has been proposed for selecting data matching a specific target genre (Müller-Eberstein et al., 2021). The best-performing option is GMM clustering, which has previously been used to identify domain clusters (Aharoni and Goldberg, 2020), but is highly likely also to capture topic information.

In this paper, we present a first investigation of the relationship between the influence of genre and topic in the context of cross-genre dependency parsing. Unlike earlier cross-lingual work, we perform this study on a set of monolingual multi-genre treebanks (Danilova and Stymne, 2023), to avoid the confound of the influence of language in earlier cross-lingual parsing studies. Our hypothesis is that the proximity of topic distributions in cross-genre transfer can positively influence the parsing performance. We believe that topic transfer plays an important, but so far unexplored, role in cross-genre transfer. In this initial exploration of the theme, we ask the following questions: 1) Can we see any relationship between the topical distance and LAS scores across datasets from different domains for a set of different languages? 2) Within the same language, will we always observe the best transfer within the same genre? If not, does topical distance explain why this is not the case?

We measure topic similarity between data sets based on a state-of-the-art topic model, BERTopic (Grootendorst, 2020). We find that in 10 out of 14 treebanks LAS scores moderately correlate with topic distances. For 5 out of 6 genre pairs with sufficient observations across treebanks, we find moderate and strong correlations with topic distances. In summary, our results suggest that topic is an important explanatory factor in the success of cross-genre transfer, which has previously been overlooked.[2]

2. Related Work

Stymne (2020) showed that when no in-genre data is available for a target language+genre, it is beneficial to add data from that genre from other languages. Müller-Eberstein et al. (2021) focused on parsing into low-resource languages and showed that selecting data matching genre based on GMM or LDA clusters or bootstrapping genre annotations led to better results than selecting sentences based on mBERT embeddings or using all multi-genre treebanks that contained the target data. However,

[1]While genre is typically used to describe works sharing a communicative purpose (e.g. Kessler et al., 1997), work on genre in dependency parsing is typically based on the genre categories in UD, which includes labels like *spoken* and *medical* (de Marneffe et al., 2021).

[2]Code at: https://github.com/UppsalaNLP/genre_topic_transfer

LREC-COLING 2024, pages 13879–13884
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Treebank	Lang.	Train	Dev	|Topics|	|Genres|
HSE	bel	240k	24.7k	276	3/5
CAC	cze	270k	10.9k	198	3/3
EWT	eng	330k	24.8k	403	5/5
GUM	eng	165k	39.0k	232	6/9
EDT	est	270k	43.9k	274	3/3
TDT	fin	220k	13.7k	365	6/6
Sequoia	fre	80k	17.9k	97	3/4
ISDT	ita	160k	10.8k	153	4/5
Nynorsk	nor	250k	41.2k	312	4/4
RRT	rum	300k	20.5k	360	4/5
SynTagRus	rus	290k	84.6k	266	3/6
Taiga	rus	150k	32.4k	270	3/3
BOUN	tur	140k	9.9k	197	3/3
ArmTDP	hyw	170k	31.9k	263	3/9

the number of topics generated using BERTopic, and the number of genres used for training/testing. Language codes are given according to ISO-639-2.

the number of topics generated using BERTopic, and the number of genres used for training/testing. Language codes are given according to ISO-639-2.

Table 1: Treebanks used, with total data sizes, the number of topics generated using BERTopic, and the number of genres used for training/testing. Language codes are given according to ISO-639-2.

even though using clustered data improved parsing, it is not clear that the clusters only capture genre, and not language and topic features, among others.

Petrenz and Webber (2011) explored the influence of topic distribution shifts on genre classification and concluded that topic distribution plays a key role in the stability of genre classification. Furthermore, it has been shown that topics are well captured when clustering LLM embeddings and topic coherence is high for the main topic modelling benchmarks (Sia et al., 2020; Zhang et al., 2022). Moreover, Aharoni and Goldberg (2020) who introduced domain clustering of LLM embeddings suggested that cluster assignments are sensible to the presence of topical terms. All this suggests that the influence of topic similarity on genre transfer for dependency parsing needs to be explored. Our work presents the first experiments in this direction.

3. Data

For our experiments we use the UD-MULTIGENRE dataset (Danilova and Stymne, 2023),[3], which is a reorganization of a highly multilingual subset of the Universal Dependencies (de Marneffe et al., 2021) treebanks, v2.11 (Zeman et al., 2022). our goal with UD-MULTIGENRE was to enrich the UD genre annotation, which was only available at the treebank level, by adding instance-level genre annotations. We manually studied the documentation, associated publications, and metadata of individual treebanks, to arrive at 17 consistent genre labels, avoiding clearly topical genres like *medical*, into which we reorganized these treebanks. UD-MULTIGENRE contains training and dev data for 38 languages from 63 UD treebanks.

In this work, we select 14 UD-MULTIGENRE treebanks in 12 languages with training data for at least 3 genres in each, see Table 2. This allows comparing the relationship between LAS and topical distance for several source genres within each language-specific treebank. For each genre, we set the minimum size in tokens per training sample to 10k, which is generally enough data to reach in-genre LAS scores of at least 80, which we find sufficient for this study.

Where possible, we collect up to 3 random non-overlapping samples with 3 different seeds which results in a maximum of 9 samples varying in sentences. This variance in sentence composition across samples is likely to be associated with differences in topic distribution and provides additional data for the analysis. For testing, we use the dev set for each genre/treebank. For many treebanks, the number of genres available for testing is higher than for training.

4. Methods and Tools

4.1. Dependency Parsing

For dependency parsing, we use the MaChAmp toolkit (van der Goot et al., 2021), which is based on fine-tuning an LLM, in our case XLM-Roberta. The parser is graph-based, using biaffine attention (Dozat and Manning, 2017) and the CLU algorithm (Chu and Liu, 1965; Edmonds, 1967).

We train a separate parsing model for each of the training sets. Each parsing model is thus trained on a single genre, source from one treebank, covering one language, as specified in Table 2. Each of these genre-specific models is thus applied to all test sets for the given treebank, which means that we have parsing results both for matching genres and for all possible mismatching genres in each treebank.

4.2. Topic Modelling

For topic modelling, we use a particular version of neural topic models, BERTopic introduced in (Grootendorst, 2020). This method leverages a well-known hierarchical density-based clustering algorithm HDBSCAN (McInnes et al., 2017) to cluster sentence-transformer embeddings after dimensionality reduction with Uniform Manifold Approximation Projection (UMAP) (McInnes et al., 2018), and class-based TF-IDF to improve cluster representation. In previous research, BERTopic has been reported to be advantageous for short-text topic modelling, e.g. Twitter data (Egger and Yu, 2022; Kellert and Mahmud Uz Zaman, 2022).

[3] https://github.com/UppsalaNLP/UD-MULTIGENRE

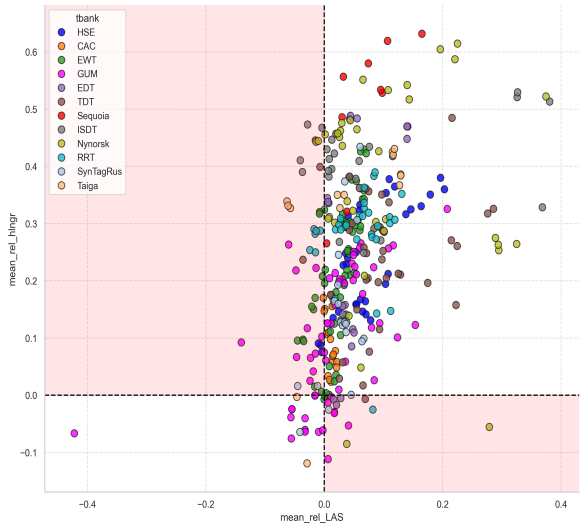

Figure 1: Relationship between mean (across seeds) relative values of LAS and Hellinger distance across all genre pairs in all treebanks.

Each sentence is considered a document. Preprocessing includes the removal of web addresses, emojis, and special symbols, which are rather genre than topic-specific. Sentences are encoded using a multilingual sentence-transformers model, paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019), trained on 50+ languages. The number of topics is inferred by the clustering algorithm and it varies from 97 to 403 across treebanks in our case.

We generate a topic model for each of the 14 treebanks shown in Table 2. Each sample includes all training and development sets in all genres available for each treebank. The produced document-topic distributions are aligned with genre-specific training and development sets within each treebank and transformed into distributions of topic proportions. For the resulting 14 topic models, we perform a reduction of outlier sentences, by allocating sentences that were not assigned any topic by HDBSCAN to topic clusters using cosine distance between topic and sentence embeddings.

4.3. Topic Distance Measurement

To calculate the topic distance between genre pairs, we perform the following steps. For each document (sentence) in each sample within a given treebank, we select the topic with the highest probability as estimated by BERTopic. Next, we calculate topic distributions for each sample, based on the sentences in the sample. The distance between the resulting document-topic probability distributions for each source-target pair is measured using Hellinger distance (HD), which together with KL divergence has long been used in topic modelling (Huang, 2008; Zhu et al., 2012) to calculate distances between document-topic and topic-word probability distributions.[4]

4.4. Correlation between LAS and Topic Distances

We assess the relationship between parsing performance (LAS) and topic distances across various genre pairs and treebanks, using the Spearman (ρ_s) correlation.

This investigation serves a dual purpose: firstly, it offers insights into the impact of topic distribution on cross-genre transfer across different treebanks and languages. Secondly, it examines the similarity between LAS score distribution and patterns of topical distance within treebanks. Our underlying hypothesis posits that, in most cases, whether in cross-genre transfer or within individual treebanks, we will observe a correlation between topical distances and parsing performance.

To allow exploration of the relationship between topic distribution and parsing performance *across* genres and treebanks, we use relative LAS scores and distances by normalizing the cross-genre scores by the in-genre scores: $RelLAS_{g_a \to g_b} = LAS_{g_a \to g_a} - LAS_{g_a \to g_b}$. For HD, we normalize the values as follows: $RelHD_{g_a \to g_b} = HD_{g_a \to g_b} - HD_{g_a \to g_a}$.

This approach allows us to investigate whether superior cross-genre transfer in terms of LAS corresponds to greater topical similarity between training and development data and vice versa. Smaller relative values of LAS and HD suggest closer alignment between source and target genres in both LAS and topic spaces.

5. Results

Overall and in-treebank correlation. The overall correlation between relative LAS and relative HD for in-treebank cross-genre transfer is shown in Figure 1. The overall correlation is moderate, but significant ($\rho_s = 0.32$, $p < 0.01$). Observations in the upper right quadrant, where most points are concentrated, correspond to cases where LAS scores for non-matching genre pairs (source $\neq$ target genre) are lower than for in-genre (source = target genre) for the corresponding treebanks and the topic distance is larger, which is the expected behavior. We note that the majority of values are in this quadrant. LAS and HD corresponding to the in-genre transfer are at the zero coordinates in this plot, since we normalize by these values.

The lower left quadrant represents the relatively few cases where the parsers achieved the highest

[4]We also experimented with KL divergence, which gave similar results, and thus is not reported.

treebank	ρ_s	$g_m \uparrow$
HSE	-0.66***	3/3
CAC	-0.19	3/3
EWT	-0.32***	3/5
GUM	-0.40***	3/6
EDT	-0.50***	3/3
TDT	-0.67***	5/6
Sequoia	-0.82***	3/3
ISDT	-0.38***	3/4
Nynorsk	-0.47***	3/4
RRT	-0.48***	3/4
SynTagRus	-0.20	2/3
Taiga	-0.56***	1/3
BOUN	0.29	2/3
ArmTDP	0.16	1/3

***$p < 0.01$

Table 2: Within-treebank correlation between absolute LAS and HD. Significance levels are estimated using permutation tests. $g_m \uparrow$ denotes the number of source genres out of all source genres for each treebank where the highest LAS corresponds to the matching-genre transfer $g_a \rightarrow g_a$.

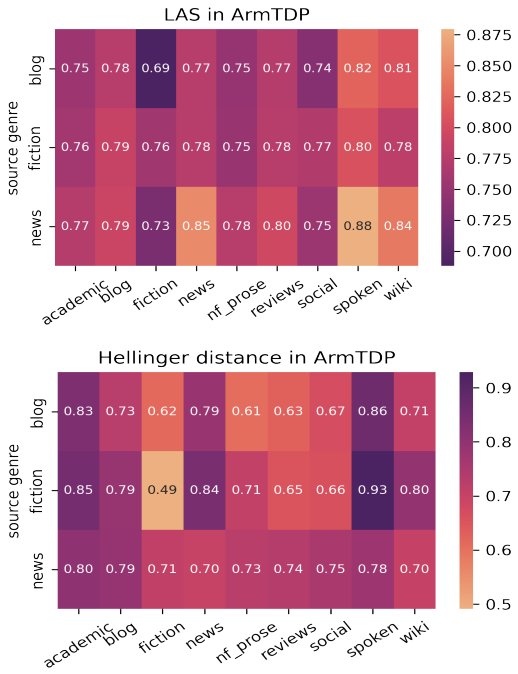

Figure 2: Mean absolute LAS scores (across samples and seeds) and mean HD in ArmTDP. The lightest colour indicates the highest LAS and the shortest topic distance. nf_prose denotes nonfiction_prose

LAS for non-matching genre pairs and the corresponding topic distances are also shorter than for the in-genre.

The upper left quadrant represents cases where LAS is relatively high even though the topic distance is large. Many of these instances are from the ArmTDP and BOUN treebanks. The lower right quadrant corresponds to the cases where LAS is relatively low, although the topics are close. It mostly includes instances from ArmTDP and GUM treebanks.

In summary, the transfer cases observed in the white quadrants are associated with our hypothesis that the higher the topical proximity between the source and target - the higher the LAS score. The cases in the pink quadrants deviate from the expected behaviour and we investigate them closer further in this paper.

Table 2 shows the correlation between absolute LAS and HD within each treebank. Negative correlations mean that higher LAS scores correspond to lower distances, which is expected, and occur for 10 out of 14 treebanks.[5]

We see exceptions for ArmTDP and BOUN, with positive correlations. When we exclude these treebanks and measure the overall correlation, we obtain $\rho_s = 0.42$ under $p < 0.01$. To better understand the situation within these treebanks, we explore heatmaps of LAS and HD. In Figure 2, we observe that for ArmTDP, the spoken dev set has the highest LAS scores for all source genres although the topic distance to spoken is the largest. This may indicate that this dev set is the easiest for the parser for some reason, and topic distance plays no role here. Fiction, in contrast, has the overall lowest LAS scores.

The pattern is similar for BOUN, where nonfiction_prose has low LAS scores in all cases, which is not explainable by topic distances.

Based on this, we can conclude that most parsing scores can be associated or partially explained with topic distances, but that the difficulty of particular test sets is also influential.

Correlation in genre pairs. Table 2 ($g_m \uparrow$ column) shows that for 4 out of 14 treebanks matching genre pairs for all source genres receive the highest LAS scores, and for 7 treebanks this is the case in the majority of cases. We analyze the LAS and HD heatmaps (as in Figure 2) for each treebank to interpret this. In SynTagRus, TDT, GUM and RRT, topic distances contribute to the best LAS scores of non-matching genre pairs. In EWT and Taiga, it is due to dev sets easy for the parsers that receive the highest LAS for all source genres. For Nynorsk and ISDT, the best LAS on non-matching genre pairs (news→blog and news→QA and wiki) is not clear.

[5]The significance levels should be interpreted cautiously due to relatively small sizes of treebank samples (92 observations on average). Although Pearson correlation scores closely follow Spearman, we do not report them because our data does not satisfy the necessary assumptions.

source genre	target genre	ρ_s
academic	fiction	0.57***
academic	news	0.35**
fiction	academic	0.71***
legal	news	0.78***
news	fiction	-0.46***
news	nonfiction_prose	0.76***

$^{***}p < 0.01$, $^{**}p < 0.05$

Table 3: Correlation between relative values of LAS and Hellinger distance in genre pairs. Significance levels are estimated using permutation tests.

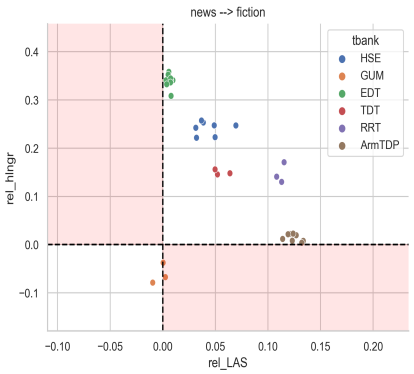

Figure 3: Correlation between relative values of LAS and HD in news → fiction genre pair.

We further explore the correlation between relLAS and relHD for specific genre pairs across all treebanks where they occur, considering only the genre pairs where we have data from at least 4 treebanks, see Table 3. Again we see that in most cases, there is an expected correlation between better LAS scores and lower topic distances. The only exception is news→fiction, for which the relationship is plotted in Figure 3. In all cases, the observations are placed in or very close to the upper right and lower left quadrants, meaning that higher LAS scores correspond to lower topic distances. In this case, the negative correlation is due to differences between treebanks. This suggests that topic is indeed an important explanatory factor for genre transfer.

6. Conclusion

We have presented an initial exploration of the impact of topic on cross-genre transfer for dependency parsing. For most treebanks and genre pairs, we find that lower topic distances indeed correspond to higher LAS, confirming that topic is an important explanatory factor of genre transfer. However, not all cases can be explained by topic distance. In some cases, the difficulty of the test set plays a role, but in other cases, further exploration is needed in future work. Another promising research direction would be to explore the role of topic versus genre in GMM-based clustering used in earlier work on selection of in-genre training data for dependency parsing (Müller-Eberstein et al., 2021; Danilova and Stymne, 2023). We would also like to expand this work to the cross-lingual setting explored in earlier work, in order also to try to disentangle the role of language in relation to genre and topic. An essential aspect to address in addition to this involves exploring potential confounding factors that might account for the observed correlation between LAS and topic distances. It is still an open question how topic similarity helps to improve parsing. An interesting line of research would be to explore the syntactic similarity of datasets with similar topic distributions, as well as across genres.

7. Bibliographical 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.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400.

Vera Danilova and Sara Stymne. 2023. UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the national Bureau of Standards B*, 71(4):233–240.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7.

Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.

Anna Huang. 2008. Similarity measures for text document clustering. In *New Zealand Computer Science Research Student Conference*, Christchurch, New Zealand.

Olga Kellert and Md Mahmud Uz Zaman. 2022. Using neural topic models to track context shifts of words: a case study of COVID-related terms before and after the lockdown in April 2020. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 131–139, Dublin, Ireland. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Leland McInnes, John Healy, and S. Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2:205.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arxiv:1802.03426.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Philipp Petrenz and Bonnie Webber. 2011. Stable classification of text genres. *Comput. Linguist.*, 37(2):385–393.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

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. 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.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3886–3893, Seattle, United States. Association for Computational Linguistics.

Shunzhi Zhu, Lizhao Liu, and Yan Wang. 2012. Information retrieval using Hellinger distance and sqrt-cos similarity. In *2012 7th International Conference on Computer Science & Education (ICCSE)*, pages 925–929.

8. Language Resource References

Zeman, Dan and Nivre, Joakim and Abrams, Mitchell and others. 2022. *Universal Dependencies 2.11*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. PID http://hdl.handle.net/11234/1-4923.