UD-MULTIGENRE – a UD-Based Dataset Enriched with Instance-Level Genre Annotations

Vera Danilova and Sara Stymne
Department of Linguistics and Philology
Uppsala University, Sweden
`first_name.last_name@lingfil.uu.se`

Abstract

Prior research on the impact of genre on cross-lingual dependency parsing has suggested that genre is an important signal. However, these studies suffer from a scarcity of reliable data for multiple genres and languages. While Universal Dependencies (UD), the only available large-scale resource for cross-lingual dependency parsing, contains data from diverse genres, the documentation of genre labels is missing, and there are multiple inconsistencies. This makes studies of the impact of genres difficult to design. To address this, we present a new dataset, UD-MULTIGENRE, where 17 genres are defined and instance-level annotations of these are applied to a subset of UD data, covering 38 languages. It provides a rich ground for research related to text genre from a multilingual perspective. Utilizing this dataset, we can overcome the data shortage that hindered previous research and reproduce experiments from earlier studies with an improved setup. We revisit a previous study that used genre-based clusters and show that the clusters for most target genres provide a mix of genres. We compare training data selection based on clustering and gold genre labels and provide an analysis of the results. The dataset is publicly available.[1]

1 Introduction

In the context of cross-lingual transfer to low-resource target languages, a significant effort is put into identifying the most suitable source data for the transfer process. The source language, as a pivotal transfer factor, is subject to comprehensive research (e.g. Lin et al., 2019; Lauscher et al., 2020; Turc et al., 2021). Within cross-lingual dependency parsing, a direction of research explores the additional impact of the text genre dimension (Stymne, 2020; Müller-Eberstein et al., 2021a). These studies use data from Universal Dependencies (UD) (Nivre et al., 2020), which provides detailed cross-linguistically consistent morphosyntactic annotations for over 100 languages. Genre[2] information is represented by labels that are assigned at the treebank level. While UD has extensive guidelines for morphosyntactic annotations, Nivre et al. (2020) note that genre labels lack both exclusive boundaries and consistent criteria, and there is a lack of comprehensive descriptions of UD genres. This means that each contributor of a UD treebank may interpret the genre labels in a different way, leading to inconsistencies. Our investigation shows that it is indeed often the case that the actual texts included in a treebank do not match the assigned genre label(s). The inconsistencies in genre annotation in UD, limit the possibilities of exploring the effect of genre on parsing and other studies based on UD, and is a confounding factor in previous studies, such as Müller-Eberstein et al. (2021a).

We present UD-MULTIGENRE, a dataset of instance-level genre annotations for a highly multilingual subset of UD, based on a comprehensive manual analysis of documentation and metadata for individual UD treebanks. We analyze the existing UD genres, and propose modifications to achieve more coherent genres, resulting in a set of 17 target genres. We then go through a subset of UD treebanks and reorganize them into controlled single-genre subsets. The training and development part of the corpus covers all 17 genres and 38 languages from 63 UD treebanks. We also create a smaller test set covering 5 genres and 16 languages, based on 17 UD treebanks. In addition, we perform an experiment on genre-aware cross-lingual dependency parsing, where we revisit the most successful method in Müller-Eberstein et al. (2021a) and reanalyze it based on our gold genre

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

[2] We follow the terminology of UD, and use the term *genre* for the distinction between categories. We note though, that some of the categories used in UD are not strictly genres, such as *medical* (topic/theme) and *spoken* (medium).

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 253–267
December 7, 2023 ©2023 Association for Computational Linguistics

annotations.

Our work makes two contributions. First, it addresses the lack of morphosyntactically annotated multilingual multigenre datasets. Some of its potential applications include 1) exploring the impact of genre on cross-lingual transfer learning, 2) understanding the role of genre in the adaptation to languages with few or no resources, and 3) learning multilingual genre representations for genre prediction. Secondly, we build on Müller-Eberstein et al. (2021a) and investigate the performance of dependency parsing when sampling multilingual training instances by gold genre compared to clustering-based sampling.

2 Related Work

Besides UD, several cross-lingual datasets exist for multiple tasks, for instance, XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020), which, however, are not focused on genres, and typically mainly have a single genre per task. There are also many datasets available annotated for genre, including corpora of raw text collected from different genres for a single language, such as the BNC.[3] Multilingual genre corpora, also annotated for other aspects are less common; one example is MultiNERD, which covers 10 languages and 2 genres, annotated for NER (Tedeschi and Navigli, 2022).

Compared to other datasets, UD stands out as covering a high number of languages for a diverse set of genres, and annotation of morphosyntax. The UD treebanks are contributed by independent teams, who are expected to follow the UD guidelines, which, however, are missing for genres. There is a mix between single-genre treebanks, and multi-genre treebanks, containing a mix of different genres. Some treebanks contain additional sentence-level annotations. However, these are specific to each treebank and are not standardized. In addition, each treebank comes with some additional documentation, more or less detailed, and in some cases refers to papers that describe some aspect of the treebank. Previous work has explored the distribution and properties of genres in UD (Müller-Eberstein et al., 2021b), noticing the diverse and contradictory nature of UD genre labels. Besides the available single-genre treebanks in UD, they were able to identify readily available instance-level annotations from 6 treebanks with training data and 20 with test data only. They then explored methods to automatically classify genres in the remaining multi-genre treebanks.

There has also been work attempting to improve parsing by using genre information from UD. Stymne (2020) focused on two genres, spoken and Twitter, a sub-genre of social, showing that using in-genre data from other languages led to improvements compared to using only out-of-genre data from the same or related languages. Müller-Eberstein et al. (2021a) continue this line of research and present findings showing the significance of genre features in the training data. They propose a set of data-driven methods for collecting training data for a specific target genre, mainly based on clustering. They found that it was better to use genre-based clustering or bootstrapping, rather than to just match sentences using an LLM. Including all multi-genre treebanks containing a given target genre, led to worse results than even a random baseline, even though this gave the largest training data sets. One of their best methods is clustering based on Gaussian mixture models (GMMs), originally explored by Aharoni and Goldberg (2020) for monolingual domain clustering. The idea is to cluster each multi-genre treebank into the same number of clusters as their assigned genres, and then select the cluster that is closest to a target genre embedding, calculated based on 100 sentences from the target treebanks.

Another line of work has tried to improve UD parsing for a given language by combining all treebanks for the language. While not directly related to genre, it is one of the relevant aspects. Overall the findings are that concatenation of treebanks does not work well and that a more advanced method is needed to take advantage of the different treebanks, such as single-treebank fine-tuning (Che et al., 2017; Shi et al., 2017), treebank embeddings (Stymne et al., 2018), or adversarial networks (Sato et al., 2017).

3 UD-MULTIGENRE Dataset

The main purpose of this effort is to provide consistent and comprehensive instance-level genre annotation of UD treebanks covering many genres and multiple languages per genre. We achieve this by splitting existing UD treebanks into subsets with a single genre, which we reclassify by going through treebank documentation. The dataset enables new research as well as re-evaluation and a deeper un-

[3] https://www.english-corpora.org/bnc/

derstanding of prior research on genre-based data selection for cross-lingual dependency parsing. In addition, it is highly relevant for the research direction that investigates cross-lingual genre representation and classification (Petrenz, 2012).

Our dataset is based on treebanks from UD version 2.11 and focuses mainly on training and development sets. Additionally, for the experiments carried out in this paper, we collected a small test set including additional languages. Collecting test sets across all covered genres is left for future work.

3.1 General Overview

The main part of the dataset is made up of training and development data, collected from training and development sets of 25 single-genre and 38 multi-genre UD treebanks in 38 languages from 15 language families, as well as the English-Tweebank (Liu et al., 2018), which contains Twitter data in UD format.[4] Currently, the total size of the dataset (training and development) in tokens is 11096.9k, and in sentences - 657.4k. In addition, the test set currently includes data from 17 treebanks for five genres and 14 low-resource languages (119k tokens and 7.2k sentences).

In order to get a coherent and useful dataset, we decided on the following limitations, and excluded:

- data in ancient languages including data attributed, among others, to the *bible* and *poetry* genres. Genres in ancient languages are likely to have their distinctive properties and their annotation requires further analysis, which will be addressed in future research;
- data that requires paid subscription;
- subsets of training instances with less than 500 tokens per genre in a treebank;
- data corresponding to UD labels *grammar_examples* and *web*, which cannot be viewed as single genres.

3.2 Genres in UD

UD contains 18 treebank-level genre labels, see Müller-Eberstein et al. (2021b)) for an overview. As pointed out by Nivre et al. (2020) for UD v2.7, the distribution of the 18 genre labels is skewed towards a few genres. In UD v2.11, which we work on, *news* is the most frequent label included in 60% of all treebanks in training data. While it might seem to be the most consistent in terms of text sources, this is not always the case. We found it to be represented both by daily mainstream news and long reads from magazines and periodicals. *Nonfiction* is the second most frequent and diverse genre in UD that subsumes many subgenres including *academic*, *legal*, and others, as noted earlier in Müller-Eberstein et al. (2021b).

The descriptions in the underlying documentation often mismatch the assigned genre labels, even for single-genre treebanks. To provide some examples, the Dutch-Alpino treebank is labelled as *news*, however, its metadata description on GitHub lists several types of genre annotation patterns that cover both news and data from other sources, such as a QA project, the Dutch reference grammar, suites for grammar maintenance, periodicals and magazines. English-Atis and Turkish-Atis are labelled as *news* and *non-fiction*, however, they belong to the *spoken* genre in fact, since they include transcriptions of human speech interactions where people request flight information through automated inquiry systems. Tamil-MWTT is labeled as *news* and it comprises sentences primarily sourced from a grammar on Tamil. Development and test sets of the same treebank may have different genre distributions. We identified similar issues by analyzing the documentation and metadata in other treebanks including those that have multiple genre labels.

41% of UD treebanks contain only test sets and have no data for training. Only a small portion (35%) of treebanks that have training data are single-genre, and many of them are small. Single-genre treebanks cover 13 genre labels where 18% of labels belong to *bible*, *grammar_examples*, and *medical*. Moreover, some of the genres are not adequately represented in single-genre treebanks, which complicates the use of methodologies from prior research in Müller-Eberstein et al. (2021a,b). While *news* constitutes 32% of single-genre treebanks in training data, *nonfiction* is represented only by data in ancient languages (Latin-ITTB, Old East Slavic-Birchbark, Sanskrit-Vedic).

3.3 Genres in UD-MULTIGENRE

The 18 original treebank-level UD genre tags serve as an initial reference. Our label set uses 11 out of these tags, for which we provide new definitions. This led to reassigning some treebanks that do not fit the new definitions. In addition, we add 6 new tags, based on coherent subgenres, most of which are currently subsumed under *nonfiction*. Table 1

[4] Converted to avoid multiple roots following Stymne (2020).

genre	in UD	Criteria
academic	✓	scientific articles and reports from different fields (medicine, oil and gas, humanities, computer science), and popular science articles
blog	✓	texts proceeding from blogging platforms like WordPress
email	✓	email messages
fiction	✓	fiction novels, stories, fairy tales. Documentation and patterns tend to include author or story names
guide		Wikihow, travel guides, instructions
interviews		prepared interviews with celebrities, politicians and businessmen
learner_essays	✓	essays of language learners on different topics that tend to contain grammar errors
legal	✓	legal and administrative texts, including texts from governmental webs
news	✓	mainstream daily (online) news, Wikinews. We stick to short articles and exclude long-read newspaper articles since they often belong to popular science
nonfiction_prose		documentary prose, biographies, autobiographical narratives, memoirs, essays
parliament		transcriptions of parliamentary speeches and debates
QA		data from Question Answering competitions
reviews	✓	messages containing reviews and opinions
social	✓	informal social media posts and discussions (e.g., Twitter, Telegram, Reddit, forum messages and comments etc.)
spoken	✓	transcriptions of spontaneous spoken speech: monologues and conversations
textbook		educational literature, textbooks
wiki	✓	main Wikipedia articles. Wikihow, Wikinews, Wikitravel, and Wikianswers are not considered in this category

Table 1: Genre selection criteria

gives an overview of all our genres with definitions.

As stated earlier, we exclude both labels and data related to treebanks in ancient languages and the extremely diverse UD genres *web* and *grammar_examples*. *nonfiction* and "topical" labels (*medical*, *government*) are discarded as labels, but the underlying data is categorized based on the analysis of the documentation and metadata patterns. Within *nonfiction*, we find the following major types of data sources that correspond to a specific metadata pattern in each treebank: 1) academic reports and popular science articles, 2) guides (wikihow, Wiki travel) and instructions, 3) textbooks, 4) nonfictional prose that includes documentary prose, biographical narratives, and essays, 5) interviews. We group the sources in 1), 2) and 4) into the corresponding new metadata-based genres. The categorization is based on concepts shared by these sources that closely align with the idea of communicative purpose. Although communicative purpose is itself a complex and multilayer concept as discussed in Askehave and Swales (2001), it has often been considered a key characteristic feature for genre identification and categorization. Academic reports and popular science articles deliver scientific knowledge and are attributed to *academic*. Guides and instructions provide step-by-step guidance on how to perform a specific task or function and are assigned the *guide* label. Documentary prose, biographical narratives and essays are literary works based mainly on factual information[5] and are assigned the *nonfiction_prose* label. Textbooks and interviews are assigned the labels *textbook* and *interviews*, respectively.

The UD *medical* label is quite rare, and the underlying data is categorized as *academic*. It is mostly represented by Romanian-SiMoNERo where texts predominantly come from scientific books. Moreover, it includes European Medicines Agency reports where medicines are their properties are mainly discussed.

The UD *government* label contains texts from governmental websites or parliamentary debates. We categorize the data from governmental websites as *legal*, since it generally aims to provide legal and administrative guidance. Parliamentary debates are attributed to *parliament*. The UD label *spoken* also contains parliamentary debates, which we include in *parliament* since we limit the *spoken* genre to contain spontaneous speech, rather than speech that is planned or scripted.

Finally, we include *QA* as a new genre. This data originates mostly from Question Answering competitions and its purpose is roughly to provide clear answers to specific questions in various domains. In UD, *QA* is mostly included in *news* or *web*.

The final assignment of a subset of instances to a genre is based on the criteria for data sources listed

[5] encyclopedia Britannica: https://www.britannica.com/topic/nonfictional-prose

Genre	*L*	*T*	*S*
academic	13	960.0k	42.8k
blog	6	92.9k	5.4k
email	1	51.2k	4.3k
fiction	20	769.3k	57.9k
guide	2	48.5k	3.5k
interview	4	62.8k	3.7k
learner_essays	1	28.6k	952
legal	11	217.0k	9.6k
news	29	6534.0k	361.6k
nonfiction_prose	9	85.0k	5.8k
parliament	11	191.7k	8.5k
QA	4	154.2k	12.2k
reviews	5	475.8k	44.0k
social	11	455.0k	32.6k
spoken	12	410.6k	34.0k
textbook	1	9.1k	430
wiki	14	549.3k	29.8k
Total	155	11096.9k	657.4k

Table 2: Number of covered languages (L) and size of each genre in tokens (T) and sentences (S) in the training and development sets.

in Table 1. As a result, annotation patterns that cannot be associated with any of these criteria are not considered and the corresponding subsets of instances are not included in UD-MULTIGENRE.

3.4 Procedure

UD-MULTIGENRE contains subsets of treebanks with consistent genres. Each subset contains information about genre, source UD treebank, language, and language family, as well as all sentences matching the subset identifiers. These subsets may originate both from multi-genre and single-genre UD treebanks. Due to the diversity of descriptions in the repositories of different UD treebanks, proper assignment of annotated subsets of instances to the corresponding genres required significant manual effort and is done as follows. For a given UD treebank (multi-genre or single-genre) we use information from the UD github repository, as well as any documentation of source corpora and treebank-related papers, and available document- and/or sentence-leval metadata. To ensure higher confidence in the retrieved patterns, original data sources are identified and provided for less clear cases. We compare the original UD labels with the official description of sources in the corresponding GitHub repositories and reclassify (parts of) treebanks when necessary. We also identify metadata patterns for each of the genres in UD-MULTIGENRE, and attribute sentences matching this metadata to the corresponding genre.[6]

Treebanks are considered good candidates to be included in the dataset when their documentation provides references to text sources, bibliographies and metadata patterns of various granularity together with the lists of genres. The procedure is more complex when detailed genre descriptions can only be found in project papers and additional documents that are available on original corpora websites. For some treebanks, scarce information on metadata patterns and their correspondence to genres is available. In this case, we verify whether the number of patterns corresponds to the number of genres and examine each annotation pattern in detail. We specifically focus on sentence-level metadata patterns `sent_id`, `newdoc id`, `genre`. For sources like *wiki*, *blog*, *fiction* and others, they often contain the exact genre names or their parts. In the case of *fiction*, they tend to contain the names of authors or literary pieces. Aligning patterns with treebank genres becomes more challenging when annotations include genre names or other identifiers in the language of the treebank. For instance, the annotations of fiction in Estonian-EDT start with `sent_id = ilu` where ilu refers to *Ilukirjandus* (eng. "fiction"). In less clear cases, we determine the origin of the texts by tracing the sources they come from.

Table 2 summarizes the size of the resulting genres in tokens and sentences, as well as the number of languages available for each genre.

3.5 Limitations in Training and Development Data

Some of the UD treebanks included in our dataset either lack development data or do not have some of their genres available in the development data. In these cases, where possible, a 20% split of training data is left for development. At least 10k tokens are left for training since our experiments require this minimum. It was done for the following treebanks and their corresponding genres: Russian-Taiga (*reviews* and *QA*), Lithuanian-ALKSNIS (*academic*), Dutch-Alpino (*QA* and *news*), Indonesian-CSUI (*news*), Slovenian-SST (*spoken*), Slovak-SNK (*fiction*), Russian-Syntagrus (*news* and *nonfiction_prose*). Slovak-SNK treebank's genres in training and development do not match. The development set contains only Wikipedia data and, since its size is sufficient to share between training and development, we use 0.9 of it (10.8k instances) as training data. For

[6] A detailed description of metadata patterns is available in the UD-MULTIGENRE repository.

Italian-ParlaMint, which lacks development data and its training data size is lower than 10k, we add 4.8k test instances to the development set since it has a test set of over 9k tokens,

3.6 Test Data

The current test data is targeted at the experiment described in Section 4, and covers five genres. We plan to collect test sets across all genres and for more languages in future work. Test data is extracted from 17 UD treebanks including test-only treebanks that satisfy the requirement of our experimental setup: maximum genetic distance should be achieved between test and training data to minimize transfer from close languages. Hence, we select test sets of treebanks in languages that do not belong to the language families of the training set. In some cases, this was not possible, such as for *fiction*, where all the available test sets belong to the Uralic language family. Consequently, we exclude Uralic languages from the training set during the experiments. UD includes PUD corpora (*wiki* and *news*) that have only test sets available. We retrieve instances for these genres for our test data in Indonesian, Japanese, Chinese, and Thai, and split them into subsets for *news* and *wiki*.

4 Experiment

We present a pilot experiment, designed to shed some further light on genre-based data selection, explored in Müller-Eberstein et al. (2021a). We limit the experiments to five genres explored in their work: *news*, *wiki*, *spoken*, *social*, *fiction*, excluding *grammar_examples*, which is not in UD-MULTIGENRE. We analyze their GMM clustering strategy for data selection and compare it to using gold genre annotated data. In addition, we aim to control for dataset size as well as minimize the impact of related languages in the data selection.

4.1 Experiment Motivation

As we have pointed out, earlier research on genres in cross-lingual UD parsing is affected by the inconsistent genre annotations in UD. In this experiment, we validate whether by addressing the limitations of prior research and obtaining a cleaner genre signal, we can confirm the statement of the previous work. Specifically, we revisit the GMM clustering method of Müller-Eberstein et al. (2021a), taking advantage of the clean genre annotations in UD-MULTIGENRE, which allows us to explore the content of GMM clusters with respect to the gold genre. In addition, we modify the GMM strategy compared to Müller-Eberstein et al. (2021a) to avoid using the target data for mean genre embedding calculation, made possible by the fact that UD-MULTIGENRE provides target genre annotations for multiple languages that are necessary to calculate mean genre embeddings for genre representation. We also control for the size of the training data and exclude all languages that are closely related to the target language from the training data, in order to isolate the genre feature as far as possible.

Our main questions can be formulated as follows: 1) Does the GMM clustering approach, based on Müller-Eberstein et al. (2021a), extract genre-specific subsets? 2) Is selecting gold target genre instances better or worse than GMM clusters? 3) What is the mix of genres in GMM clusters, especially when GMM outperforms gold?

4.2 Training Data Selection

We compare the performance of a parser trained on two types of training sets for each genre. The first type uses gold multilingual training instances from UD-MULTIGENRE subsets. The second is based on instances selected from multigenre UD treebanks using GMM, inspired by Müller-Eberstein et al. (2021a), as well as sentences from single-genre treebanks. For GMM, we use multigenre UD treebanks, from which subsets are derived. It allows us to clearly see how gold instances are distributed within clusters.

We consider several enhancements to the workflow in Müller-Eberstein et al. (2021a). First, we avoid target-like data when calculating the mean genre embeddings to represent genres. The data come neither from the same treebank (training data) nor from the same language. It is based entirely on UD-MULTIGENRE subsets derived from single-genre UD treebanks with verified labels and single-subset treebanks in UD-MULTIGENRE. This allows us to exclude the bias towards topical and language features of target data. Secondly, we minimize the influence of genetically close languages by excluding the members of the target language family from the training data for each genre.

Gold: For the gold data, we collect all subsets from UD-MULTIGENRE that are labelled with the target genre. This includes both data that was originally from UD single-genre as well as multigenre treebanks.

GMM: For each genre, mean genre embeddings are calculated by mean pooling XLMRoberta-base embeddings of $n = 100$ instances that are randomly selected from all single-genre subsets. All subsets in UD-MULTIGENRE that originate from a multi-genre UD treebank that contains the target genre are then clustered. The number of clusters is set to the number of genres in each set. Next, we compute cluster centroids and measure the cosine distance from each cluster centroid to our mean genre embeddings. The closest cluster is selected, and all sentences in it are added to the GMM training data for that genre. In addition, we add data from all matching single-genre UD treebanks, controlled in UD-MULTIGENRE, since such data is readily available.

In order to balance the size of the training data for each target, we select the number of sentences in the smallest set of gold and GMM, and sample that amount of sentences from the larger set. This ensures that the two datasets for each genre have the same size. Table 8 (Appendix) shows the sizes of the training sets.

4.3 Training Setup

We use the MaChAmp v4.2 (van der Goot et al., 2021) for dependency parsing. An older version of the same framework was used in the previous work (Müller-Eberstein et al., 2021a). Instead of mBERT, as used there, XLMRoberta base is used as the base MLM. XLMRoberta was observed to be more suitable for multigenre data since it was trained not only on Wikipedia but on a large selection of multilingual CommonCrawl resources (Lepekhin and Sharoff, 2022) and it typically gives better results for cross-lingual parsing than mBERT (see e.g. de Lhoneux et al., 2022). The performance is assessed using labelled attachment scores (LAS). We use the test data described in Section 3.6, controlling for language family in the training sets. Therefore, we remove Uralic languages from the training data for *fiction* and *social* (Finnish-OOD). It allows us to add Uralic development sets from UD-MULTIGENRE to the evaluation (Estonian-EDT *fiction*, Finnish-TDT *fiction*, Estonian-EWT *social*).

4.4 Results and Discussion

Table 3 shows the proportion of genres in the GMM clusters. It is clear that all clusters contain a mix of genres, with *news* and *fiction* containing the largest part of matching genre data, and *spoken* very little from its own genre. Only for *news*, the majority of instances come from this genre (66.64%). This indicates that GMM clustering is not solely capturing genre, but also other aspects, as also noted by Aharoni and Goldberg (2020), who suggest that cluster assignments are sensible to the presence of topical terms. Note that when using GMM for training, we concatenate it with data from single-genre treebanks, which means that additional in-genre data is added for each target genre. The proportion of such data is 24% for *fiction*, and over 50% for all other genres, up to 88% for *spoken*

	news	wiki	fiction	spoken	social
news	***66.73***	*33.80*	26.77	20.52	*23.77*
wiki	1.86	**9.13**	1.89	0.75	0.90
fiction	8.42	23.12	***43.19***	*39.74*	10.29
spoken	0.47	0.02	2.63	**18.60**	0.75
social	3.30	15.99	12.07	3.41	**21.20**
academic	8.60	4.20	1.92	0.36	0.00
blog	0.86	2.53	0.45	1.56	1.77
email	0.00	0.00	0.00	0.00	4.55
guide	0.80	0.00	0.95	3.28	0.00
interview	0.62	1.27	2.68	3.83	0.00
legal	1.92	2.85	0.69	0.00	0.00
nf_prose	1.94	2.70	3.25	5.13	1.85
parliament	3.04	1.38	3.14	0.65	0.00
QA	1.34	2.79	0.00	0.06	15.88
reviews	0.07	0.21	0.19	1.69	19.04
textbook	0.03	0.00	0.16	0.42	0.00

Table 3: Distribution in percent of gold genres in GMM-based training data for each genre. The matching genre is marked in bold, and the largest genre in each cluster is marked in italics. *nf_prose* is short for *nonfiction_prose*

The results of zero-shot dependency parsing are shown in Table 4. The performance with data selected with the GMM-based approach is generally on par with data based on gold instances. The average score is slightly higher for GMM, whereas gold is better for 12 out of 21 targets. For *fiction*, gold is the best option in all cases, with an average improvement over GMM of 2 LAS points. For all other genres, however, there is a variation between target treebanks of which option performs the best. In two cases, both in spoken, the LAS scores are equal, but very low, showing that neither of the training sets are a good fit in that case.

To further investigate the impact of genres, we additionally performed cross-genre experiments, applying the GMM and gold models for each genre, to all target genres. The full results of this experiment are shown in Table 7 (Appendix). Here we did not fully control for language relatedness, and

		GMM	gold
fiction	Erzya_JR	17.33	**18.28**
	Estonian_EDT	72.35	**74.22**
	Finnish_TDT	74.45	**75.31**
	Komi-Zyrian_Lattice	14.56	**17.34**
	Moksha_JR	18.51	**19.80**
news	Chinese_PUD	**45.88**	45.61
	Japanese_PUD	**41.50**	40.71
	Tamil_TTB	46.61	**47.76**
	Thai_PUD	57.59	**58.30**
social	Estonian_EWT	59.71	**60.75**
	Finnish_OOD	66.78	**67.85**
	Irish_TwittIrish	**47.01**	45.62
spoken	Abaza_ATB	3.07	3.07
	Beja_NSC	0.82	0.82
	Cantonese_HK	**33.40**	32.32
	Chukchi_HSE	10.60	**10.92**
	Gheg_GPS	32.61	**33.78**
	Komi-Zyrian_IKDP	**21.96**	20.57
wiki	Albanian_TSA	**82.97**	79.83
	Indonesian_PUD	**73.54**	73.37
	Japanese_PUD	33.14	**31.65**
	Average	**40.8**	**40.7**

Table 4: Zero-shot dependency parsing results (LAS)

it is clear that the best results for both gold and GMM when the training data include the same language, as for Indonesian_PUD, when trained on *news* containing Indonesian_CSUI, or when trained on related languages, such as for Irish_Twittrish trained on *news*, containing Scottish Gaelic data. This shows the importance of controlling for languages. However, there are still cases when it is preferable to train on other genres than the target genre. This is the case for *spoken*, where training on *social* is the best option for Chukchi and Cantonese and training on *fiction* is best for Komi Zyrian. This to some extent matches the content of these treebanks, which include folk stories and fairy tales in Komi Zyrian and political discussions in Cantonese.

When GMM outperforms gold we mainly observe 2 scenarios. In the first case, GMM-based training data contains a significant portion of a non-target genre g, and, at the same time, the gold parser for g scores the highest across all parsers. In the second case, gold underperforms GMMs in all or most genres, which suggests that another genre beyond the five target genres of this experiment contributes to the performance. The latter is the case for Japanese *news* and Albanian *wiki*. Examples of the former are *spoken* Cantonese and Komi Zyrian, discussed above, where we note that the *spoken* GMM cluster contains a relatively high proportion of both *fiction* and *social*. For Cantonese-HK, the parser trained on gold *social* achieves the best score of 37.4 LAS. This test set includes sentences from a council meeting discussion and an interview. Social media discussions on political issues involving several participants are quite typical of the *social* genre. Hence, although this test set contains unprepared speech with many disfluencies, characteristic of the *spoken* genre, we assume that the input from *social* in the GMM-based training data (18.47%) contributes useful instances and increases the performance. For Indonesian *wiki*, the scores when training on *news* are high, and the GMM cluster contains a high proportion of it.

This experiment provides valuable knowledge on the influence of genre distribution on dependency parsing performance. Gold training data is more advantageous in *fiction*. GMMs work better for several treebanks in *social*, *spoken*, *news*, and *wiki*, where we assume a larger diversity in terms of topics and author styles. Therefore, input from other genres can be useful. Nevertheless, on the majority of test treebanks, GMM-based genre distributions do not improve performance. On the one hand, it may be explained by a higher genre consistency. On the other hand, genre distributions may not match the target due to the use of single-genre sets instead of the target test samples for mean genre embedding calculation as in the original paper (Müller-Eberstein et al., 2021a). As stated earlier, we attempt to isolate genre from topic and language features, which would be impossible if we calculated mean genre embeddings based on target test data. In summary, the results of our experiment indicate that the distribution of genre in the training data influences the results of zero-shot dependency parsing, and minimizing the differences between distributions in training and target sets can improve the results.

5 Conclusions

This paper presents UD-MULTIGENRE, a UD-based dataset with instance-level genre annotations for 17 genres in 38 languages. It provides fine-grained verified labels for 63 treebanks with training data and 17 test-only treebanks. It constitutes a robust basis for further exploration of text genre from a multilingual perspective.

A pilot experiment illustrates the application of UD-MULTIGENRE to genre-related research. We revisit previous work that builds on treebank-level UD labels to perform training data selection for zero-shot dependency parsing. Our dataset has facilitated in-depth analysis of training sets produced by a top-performing clustering approach. We show that GMM clusters are not limited to the target genre, but contain a mix of different genres. Instead, this approach can sometimes produce training data containing genre mixtures that are advantageous for certain test treebanks. However, gold training data from UD-MULTIGENRE produces better results on the majority of test treebanks.

6 Limitations

Genre data in UD-MULTIGENRE is grouped based on UD data sources and documentation. This information is more or less detailed, however, we cannot be completely confident about it. Also, it should not be the sole basis for defining terms. More comprehensive and UD-independent genre definitions can help to further reorganize and improve the dataset.

Furthermore, genre descriptions and instance-level patterns are not available for all UD treebanks. Therefore, UD-MULTIGENRE currently cannot provide full coverage of UD. The documentation and referenced project reports contain detailed descriptions of genres for a few treebanks, such as Pomak-PHILOTIS and Welsch-CCG, however, instance-level annotations cannot be associated with them and no source documents corresponding to the annotation patterns are available on the project websites. Also, UD-MULTIGENRE currently does not cover genres encountered in ancient texts, which is a limitation for the investigation of genre-aware dependency parsing of ancient languages. Finally, additional collaboration with contributors of less documented treebanks is needed to increase confidence in annotation patterns and further enhance the clarity of genres.

Acknowledgements

This work is supported by the Swedish strategic research programme eSSENCE. Computations were enabled by resources in project UPPMAX 2020/2-2 at the Uppsala Multidisciplinary Center for Advanced Computational Science. We want to thank all contributors to UD treebanks for their work; without your effort, this work would not have been possible.

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.

Inger Askehave and John M. Swales. 2001. Genre identification and communicative purpose: a problem and a possible solution. *Applied Linguistics*, 22(2):195–212.

Wanxiang Che, Jiang Guo, Yuxuan Wang, Bo Zheng, Huaipeng Zhao, Yang Liu, Dechuan Teng, and Ting Liu. 2017. The HIT-SCIR system for end-to-end parsing of Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 52–62, Vancouver, Canada. Association for Computational Linguistics.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–587, Dublin, Ireland. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Mikhail Lepekhin and Serge Sharoff. 2022. Estimating confidence of predictions of individual classifiers and TheirEnsembles for the genre classification task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5974–5982, Marseille, France. European Language Resources Association.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6008–6018, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021a. 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.

Max Müller-Eberstein, Rob van der Goot, and Barbara Plank. 2021b. How universal is genre in Universal Dependencies? In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 69–85, Sofia, Bulgaria. 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 Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

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*, pages 71–79, Vancouver, Canada. Association for Computational Linguistics.

Tianze Shi, Felix G. Wu, Xilun Chen, and Yao Cheng. 2017. Combining global models for parsing Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 31–39, Vancouver, Canada. 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.

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.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *ArXiv*, abs/2106.16171.

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.

A Target data selection

Table 5 provides a detailed description of test sets including their size in tokens and the information on text sources. Also, for some treebanks, metadata patterns are used to extract data corresponding to target genres. The use of patterns is explained in the description column. PUD *news* test treebanks in Thai, Japanese, and Chinese represent translations of sentences randomly extracted from multiple daily news media in English: The Washington Post, The Independent, BBC and others. From PUD treebanks, *news* instances are selected using metadata where `sent_id` starting with `n` and `w` correspond to *news* and *wiki*, respectively.

B Genre in UD and UD-MULTIGENRE

Table 6 lists all UD treebanks included in UD-MULTIGENRE, and, for each of them, the official UD labels (treebank-level) and the ones that we assign to subsets of instances derived from these treebanks based on the performed analysis.

C Additional Results

Cross-genre evaluation results are shown in Table 7. As stated earlier, we control for language family distribution in the training data for each genre. Therefore, within a specific genre, we manage to avoid the influence of genetically close language. However, when we perform cross-genre evaluation, this influence takes place for some targets. Japonic, Dravidian, Caucasian, Chukotko-Kamchatkan, and IE.Albanian language families are not present in either of the training sets. Therefore, for the corresponding targets, we assume no transfer from genetically close languages across genres.

Instances from a Sino-Tibetan language (Chinese-GSD) are included only in the *wiki* training sets. Therefore, the higher scores of the *wiki* parsers on Chinese-PUD *news* and Cantonese-HK *spoken* are not taken into consideration.

Austronesian language family instances (Indonesian-CSUI) are included only in the *news* training data. Hence, the higher scores of the *news* parsers on Indonesian-PUD *wiki* are not considered.

A Celtic language, Scottish Gaelic, is present in *spoken*, *news*, and *fiction*. Therefore, the higher performance of parsers on the Irish-Twittirish test set in these genres can be due to the transfer of language features from a genetically close language.

Instances from Afroasiatic languages (Hebrew, Maltese) are part of *news*, *wiki*, and *fiction* training data. For the Beja test set, we exclude from consideration the scores of the corresponding genre-specific parsers.

Uralic languages are present in *news* and *wiki* training data. Hence, the scores of the corresponding parsers for Estonian, Finnish, Komi Zyrian, Erzya, and Moksha test sets are not taken into account.

The results where the transfer from genetically close languages takes place are marked in bold italics in Table 7.

D Additional Statistics

Table 8 shows, for each genre, the number of instances in the single-genre set (shared), in gold and GMM samples derived from the multigenre set together with the total number of instances in the balanced training data. To save computational resources, we reduce the size of the training data based on the multigenre set for news to the mean multigenre set size (38436 instances). We randomly select this number of instances from the corresponding gold and GMM training data.

Table 9 shows the distribution of language families in the clustering-based training data. Language families are the same in the gold data since it is based on the same multigenre sets. Table 10 displays the distribution of language families in the single-genre sets.

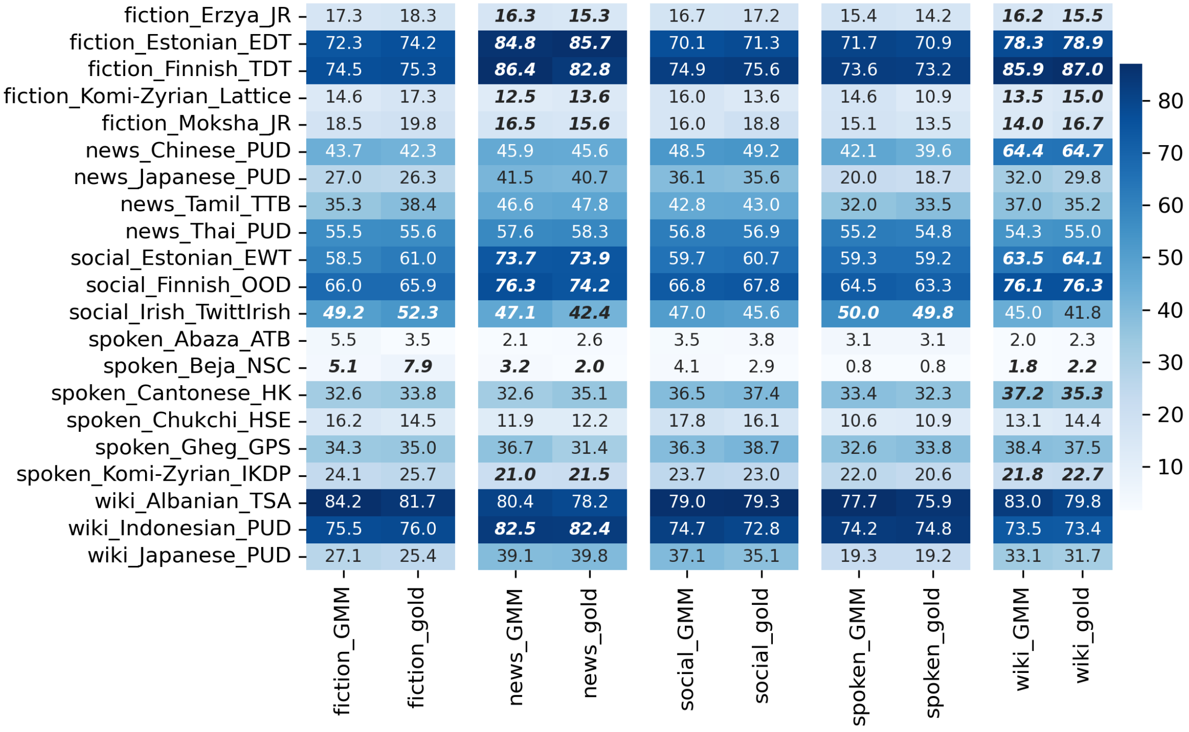
genre	treebank	language family	description	tokens
spoken	Abaza-ATB	Caucasian	spontaneous stories about the speakers' lives, village traditions, tales and legends (source: corpus website)	652
spoken	Beja-NSC	Afroasiatic	a collection of fairy tales and stories narrated by Beja speakers(source: corpus website)	856
spoken	Cantonese-HK	Sino-Tibetan	2 parts of this test set correspond to spontaneous spoken speech: send_id = 411 to 547, interview with unprepared dialogues, and send_id = 651 to 1004, meeting of the legislation council with unprepared dialogues	10231
spoken	Chukchi-HSE	Chukotko-Kamchatkan	anecdotes, songs, parables, autobiographical stories, fairy tales, everyday dialogues, retellings of silent movie fragments (source: corpus website)	5389
spoken	Gheg-GPS	IE.Albanian	narrations of Wallace Chafe's Pear Stories video (pearstories.org) by heritage speakers of Gheg Albanian. To extract the instances, metadata starting with `[sent_id = P]` is used (speakers from Prishtina), we exclude the instances of speakers from Switzerland since they contain a lot of code-switching (mostly Swiss-German).	2312
spoken	Komi Zyrian-IKDP	Uralic	Iźva dialect transcriptions of spoken Komi Zyrian (source: corpus website)	2304
wiki	Albanian-TSA	IE.Albanian	Wikipedia	922
wiki	Indonesian-PUD	Austronesian	Wikipedia	9823
wiki	Japanese-PUD	Japonic	Wikipedia	15124
news	Japanese-PUD	Japonic	traslated: Washington Post, BBC, etc.	13664
news	Tamil-TTB	Dravidian	daily news media (source: corpus website)	1772
news	Chinese-PUD	Sino-Tibetan	traslated: Washington Post, BBC, etc.	10531
news	Thai-PUD	Sino-Tibetan	traslated: Washington Post, BBC, etc.	10831
fiction	Erzya-JR	Uralic	texts from various authors of fiction who created original works in the Erzya language	10357
fiction	Komi Zyrian-Lattice	Uralic	all `sent_id` variants belong to *fiction* except for those that start with kpv (news) and `OKK` (grammar examples)	2662
fiction	Moksha-JR	Uralic	all instances belong to *fiction*, except for those `sent_id` starting with `MKS` (grammar examples)	1004
social	Irish-Twittirish	Celtic	Twitter data	15433
social	Finnish-OOD	Uralic	instances with `sent_id` starting with `thread` belong to forum discussions and `tweet` - to Twitter posts	5134

Table 5: Description of the test data grouped by genre. The selection of metadata patterns for the extraction of genre-specific subsets of instances is explained in the description column. IE stands for Indo-European language family

Treebank name	**UD labels**	**UD-MULTIGENRE labels**
Afrikaans-AfriBooms	legal, nonfiction	legal
Armenian-ArmTDP	blog, fiction, grammar-examples, legal, news, nonfiction	nonfiction_prose, blog, fiction, news, legal
Armenian-BSUT	blog, fiction, government, legal, news, nonfiction, web, wiki	nonfiction_prose, blog, fiction, news, legal, wiki
Belarusian-HSE	fiction, legal, news, nonfiction, poetry, social, wiki	social, news, nonfiction_prose, fiction, wiki
Bulgarian-BTB	fiction, legal, news	fiction, legal, academic, nonfiction_prose, news, interview
Catalan-AnCora	news	news
Chinese-GSD	wiki	wiki
Croatian-SET	news, web, wiki	news
Czech-CAC	legal, medical, news, nonfiction, reviews	legal, news, academic
Czech-FicTree	fiction	fiction
Czech-PDT	news, nonfiction, reviews	news, academic
Dutch-Alpino	news	news, QA
Dutch-LassySmall	wiki	wiki
English-Atis	news, nonfiction	spoken
English-EWT	blog, email, reviews, social, web	social, QA, reviews, blog, email
English-GUM	academic, blog, fiction, government, news, nonfiction, social, spoken, web, wiki	news, fiction, academic, nonfiction_prose, parliament, spoken, guide, interview, textbook
English-GUMReddit	blog, social	social
English-LinES	fiction, nonfiction, spoken	fiction, parliament
English-Tweebank		social
Erzya-JR	fiction	nonfiction_prose, fiction
Estonian-EDT	academic, fiction, news, nonfiction	news, academic, fiction
Estonian-EWT	blog, social, web	social
Finnish-TDT	blog, fiction, grammar-examples, legal, news, wiki	wiki, news, legal, blog, fiction, parliament
French-ParisStories	spoken	spoken
French-Rhapsodie	spoken	spoken
French-Sequoia	medical, news, nonfiction, wiki	wiki, academic, parliament, news
German-GSD	news, reviews, wiki	reviews
German-HDT	news, nonfiction, web	news
Greek-GDT	news, spoken, wiki	news, parliament
Hebrew-HTB	news	news
Hebrew-IAHLTwiki	wiki	wiki
Hindi English-HIENCS	social	social
Hindi-HDTB	news	news
Icelandic-Modern	news, nonfiction	parliament, news
Indonesian-CSUI	news, nonfiction	news
Italian-ISDT	legal, news, wiki	news, parliament, QA, wiki, legal
Italian-MarkIT	grammar-examples	learner_essays
Italian-ParlaMint	government, legal	parliament
Italian-PoSTWITA	social	social
Italian-TWITTIRO	social	social
Lithuanian-ALKSNIS	fiction, legal, news, nonfiction	academic, legal, news, fiction
Maltese-MUDT	fiction, legal, news, nonfiction, wiki	fiction, parliament
Naija-NSC	spoken	spoken
Norwegian-Nynorsk	blog, news, nonfiction	blog, parliament, legal, news
Norwegian-NynorskLIA	spoken	spoken
Polish-LFG	fiction, news, nonfiction, social, spoken	social, news, fiction, academic, spoken
Portuguese-PetroGold	academic	academic
Romanian-RRT	academic, fiction, legal, medical, news, nonfiction, wiki	legal, news, fiction, academic, wiki
Romanian-SiMoNERo	medical	academic
Russian-GSD	wiki	wiki
Russian-SynTagRus	fiction, news, nonfiction	news, fiction, academic, nonfiction_prose, interview, wiki
Russian-Taiga	blog, fiction, news, poetry, social, wiki	social, QA, reviews
Scottish Gaelic-ARCOSG	fiction, news, nonfiction, spoken	fiction, news, spoken, interview
Slovak-SNK	fiction, news, nonfiction	fiction, legal, nonfiction_uc, news, wiki
Slovenian-SSJ	fiction, news, nonfiction	wiki
Slovenian-SST	spoken	spoken
Swedish-LinES	fiction, nonfiction, spoken	fiction, parliament

Treebank name	**UD labels**	**UD-MULTIGENRE labels**
Turkish German-SAGT	spoken	spoken
Turkish-Atis	news, nonfiction	spoken
Turkish-BOUN	news, nonfiction	news, guide, nonfiction_prose
Turkish-Tourism	reviews	reviews
Uyghur-UDT	fiction	fiction
Western Armenian-ArmTDP	blog, fiction, news, nonfiction, reviews, social, spoken, web, wiki	nonfiction_prose, academic, news, fiction, blog, reviews, social, wiki, spoken

Table 6: Initial UD treebank-level genre labels compared to labels that correspond to each treebank in UD-MULTIGENRE



	fiction_GMM	fiction_gold	news_GMM	news_gold	social_GMM	social_gold	spoken_GMM	spoken_gold	wiki_GMM	wiki_gold
fiction_Erzya_JR	17.3	18.3	***16.3***	***15.3***	16.7	17.2	15.4	14.2	***16.2***	***15.5***
fiction_Estonian_EDT	72.3	74.2	***84.8***	***85.7***	70.1	71.3	71.7	70.9	***78.3***	***78.9***
fiction_Finnish_TDT	74.5	75.3	***86.4***	***82.8***	74.9	75.6	73.6	73.2	***85.9***	***87.0***
fiction_Komi-Zyrian_Lattice	14.6	17.3	***12.5***	***13.6***	16.0	13.6	14.6	10.9	***13.5***	***15.0***
fiction_Moksha_JR	18.5	19.8	***16.5***	***15.6***	16.0	18.8	15.1	13.5	***14.0***	***16.7***
news_Chinese_PUD	43.7	42.3	45.9	45.6	48.5	49.2	42.1	39.6	***64.4***	***64.7***
news_Japanese_PUD	27.0	26.3	41.5	40.7	36.1	35.6	20.0	18.7	32.0	29.8
news_Tamil_TTB	35.3	38.4	46.6	47.8	42.8	43.0	32.0	33.5	37.0	35.2
news_Thai_PUD	55.5	55.6	57.6	58.3	56.8	56.9	55.2	54.8	54.3	55.0
social_Estonian_EWT	58.5	61.0	***73.7***	***73.9***	59.7	60.7	59.3	59.2	***63.5***	***64.1***
social_Finnish_OOD	66.0	65.9	***76.3***	***74.2***	66.8	67.8	64.5	63.3	***76.1***	***76.3***
social_Irish_TwittIrish	***49.2***	***52.3***	***47.1***	***42.4***	47.0	45.6	***50.0***	***49.8***	45.0	41.8
spoken_Abaza_ATB	5.5	3.5	2.1	2.6	3.5	3.8	3.1	3.1	2.0	2.3
spoken_Beja_NSC	***5.1***	***7.9***	***3.2***	***2.0***	4.1	2.9	0.8	0.8	***1.8***	***2.2***
spoken_Cantonese_HK	32.6	33.8	32.6	35.1	36.5	37.4	33.4	32.3	***37.2***	***35.3***
spoken_Chukchi_HSE	16.2	14.5	11.9	12.2	17.8	16.1	10.6	10.9	13.1	14.4
spoken_Gheg_GPS	34.3	35.0	36.7	31.4	36.3	38.7	32.6	33.8	38.4	37.5
spoken_Komi-Zyrian_IKDP	24.1	25.7	***21.0***	***21.5***	23.7	23.0	22.0	20.6	***21.8***	***22.7***
wiki_Albanian_TSA	84.2	81.7	80.4	78.2	79.0	79.3	77.7	75.9	83.0	79.8
wiki_Indonesian_PUD	75.5	76.0	***82.5***	***82.4***	74.7	72.8	74.2	74.8	73.5	73.4
wiki_Japanese_PUD	27.1	25.4	39.1	39.8	37.1	35.1	19.3	19.2	33.1	31.7

Table 7: Complete results of zero-shot dependency parsing evaluation (LAS). In bold italics, we mark the results of cross-genre evaluation where the same and/or genetically close languages are present in the training data

	news	wiki	fiction	spoken	social
shared	190272	19552	11816	24156	14424
gold	131894	6548	29714	3062	14373
GMM	113643	22408	27222	4907	23311
Total balanced	228708*	26100	39038	27218	28797

Table 8: For each genre, the number of instances in shared (single-genre set), gold and GMM samples derived from the multigenre set, as well as the total number of instances in the balanced data (details on balancing are given in Section 4.2). *To save computational resources, the mean multigenre set size is used for *news* (38436 instances)

Language families	news	wiki	fiction	spoken	social
IE.Slavic	68.03	66.32	68.62	58.84	80.30
IE.Germanic	15.49	0.00	19.47	14.92	16.48
Uralic	7.80	9.45	0.00	0.00	0.00
IE.Romance	3.75	19.37	2.80	0.00	0.00
Altaic	1.88	0.00	0.00	0.00	0.00
IE.Armenian	1.34	4.86	3.94	14.56	3.22
IE.Greek	1.12	0.00	0.00	0.00	0.00
IE.Celtic	0.60	0.00	2.02	11.68	0.00
Afro-Asiatic	0.00	0.00	1.69	0.00	0.00
IE.Baltic	0.00	0.00	1.47	0.00	0.00

Table 9: Distribution of language families in clustering-based training data (from multigenre sets) for each genre (in percent)

Language families	news	wiki	fiction	spoken	social
IE.Germanic	80.43	29.61	0.00	31.82	22.62
IE.Indic	6.99	0.00	0.00	0.00	0.00
IE.Romance	6.90	0.00	0.00	11.09	63.29
Afro-Asiatic	2.75	21.98	0.00	0.00	0.00
IE.Slavic	2.65	27.97	85.99	6.88	0.00
Austronesian	0.28	0.00	0.00	0.00	0.00
Sino-Tibetan	0.00	20.44	0.00	0.00	0.00
Altaic	0.00	0.00	14.01	17.69	0.00
Creole	0.00	0.00	0.00	30.13	0.00
Code-switch	0.00	0.00	0.00	2.39	14.09

Table 10: Distribution of language families in single-genre sets for each genre (in percent)