Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 247-257 recibido 04-04-2024 revisado 13-05-2024 aceptado 17-05-2024

Towards accurate dependency parsing for Galician with limited resources

Hacia un análisis de dependencias preciso para gallego usando recursos limitados

Albina Sarymsakova,[*1] Xulia Sánchez-Rodríguez,[*2] Marcos Garcia[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]Language Variation and Textual Categorisation (LVTC) Universidade de Vigo
{albina.sarymsakova, marcos.garcia.gonzalez}@usc.gal, xulia.sanchez@uvigo.gal
[*]Equal contribution

Abstract: Automatic syntactic parsing is a fundamental aspect within NLP. However, effective parsing tools necessitate extensive and high-quality annotated treebanks for satisfactory performance. Consequently, the parsing quality for low-resource languages such as Galician remains inadequate. In this context, the present study explores several approaches to improve the automatic syntactic analysis of Galician using the UD framework. Through experimental endeavors, we analyze the quality of the model incrementing the size of the initial training corpus by adding data from Galician PUD treebank. Additionally, we explore the benefits of incorporating contextualized vector representations by comparing the use of various BERT models. Lastly, we assess the impact of integrating cross-lingual training data from similar varieties, analyzing the models' performance across used treebanks. Our findings underscore (1) the positive correlation between augmented training data and enhanced model performance across used treebanks; (2) superior performance of monolingual BERT models compared to their multilingual analogues; (3) improvement of overall model performance across utilized treebanks by incorporation of cross-lingual data.
Keywords: Galician, automatic parser, Universal Dependencies, BERT.

Resumen: El análisis sintáctico automático es fundamental dentro del PLN. Sin embargo, las herramientas eficaces requieren bancos de árboles extensos y de alta calidad para el entrenamiento satisfactorio. En consecuencia, la calidad del análisis sintáctico sigue siendo inadecuada para lenguas de escasos recursos como el gallego. En este contexto, el presente estudio explora varios enfoques para mejorar el análisis sintáctico del gallego utilizando el marco de UD. Nuestros experimentos analizan la calidad del modelo incrementando el tamaño del corpus de entrenamiento inicial añadiendo datos del PUD gallego. Además, exploramos los beneficios de incorporación de las representaciones vectoriales contextualizadas y el uso de varios modelos BERT. Por último, evaluamos el impacto de la integración de datos interlingüísticos para el entrenamiento de variedades similares, analizando el rendimiento del modelo en los bancos de árboles usados. Nuestros hallazgos subrayan (1) la correlación positiva entre los datos de entrenamiento aumentados y el rendimiento mejorado del modelo; (2) el rendimiento superior de los modelos BERT monolingües en comparación con sus análogos multilingües; (3) el rendimiento mejorado general del modelo en los bancos de árboles tras la incorporación de datos interlingüísticos.
Palabras clave: gallego, análisis sintáctico, Dependencias Universales, BERT.

ISSN 1135-5948 DOI 10.26342/2024-73-18

1 Introduction

The Universal Dependencies (UD) initiative serves as a multilingual framework for natural language processing (NLP), offering a standardized system for morphological and syntactic annotation across languages. It facilitates collaborative efforts to create annotated corpora in multiple languages, thereby establishing a growing repository of essential resources for linguistic research. Currently, the UD project spans more than 217 treebanks, covering 122 languages across 24 different language families[1]. However, a significant discrepancy exists in the amount of available syntactic analysis models for each language. Specifically, the limited availability of automatic parsing models trained with treebank data for low-resource languages like Galician presents a challenge for researchers interested in conducting both cross-lingual and NLP studies (Vania et al., 2019).

As posited by Kondratyuk and Straka (2019), the efficacy of automatic parsing tools requires meticulously annotated treebanks for methodical training. Consequently, the parsing quality for languages characterized by limited linguistic resources, exemplified by Galician, is currently insufficient. Moreover, models trained exclusively on a singular treebank may encounter challenges pertaining to domain adaptation, whereby they assimilate idiosyncratic features inherent to the training data. Furthermore, the usage of contextualized vector representations, spanning both monolingual and multilingual models, underscores the necessity for rigorous systematic evaluation to identify the most effective parsing solution. Hence, in navigating the landscape of parsing methodologies, a thorough understanding of these factors is essential to discern and implement the most suitable approach for achieving parsing precision and efficacy.

Based on the outcomes and with the objective of extending contributions to the field of automatic parsing models for Galician, we formulate our hypothesis in the present investigation as follows: Consequential augmentation of training resources by incorporation of supplementary data, such as additional training sentences, and pre-trained word embeddings, brings a significant improvement in the performance of the automatic syntactic analysis models of Galician using the Universal Dependencies methodological framework and elucidates the comparative dynamic using different Galician treebanks. To test this hypothesis, this study aims to achieve the following objectives:

a To assess the performance of the model by incrementally augmenting the training data, and subsequently evaluating the parsing efficacy through the examination of LAS and UAS metrics.

b To employ three different BERT-based embedding models in order to ascertain the best configuration for achieving superior parsing performance.

c To evaluate the model's performance subsequent to the integration of cross-lingual data from related linguistic varieties into the training dataset.

d To analyze possible inconsistencies of the newly trained model in terms of syntactic dependencies.

Our experimental findings validate the favorable impact of augmenting training data and indicate the superiority of monolingual BERT models over multilingual ones. Moreover, to our knowledge, the model proposed in our work achieves superior results in dependency parsing for Galician.[2]

The subsequent sections of this paper are organized as follows. Section 2 provides a brief overview of existing Galician parsing models. Section 3 describes the Treebanks and explains its relevance to our work. Section 4 presents the design of our experiments and its outcomes. Section 5 refers to the discussion that emerges from the main findings collected within Section 4. Finally, we conclude in Section 6.

2 Related work

Several antecedent works have addressed the problem of development of automated processing resources for the Galician language. Regarding the parsing models for Galician, previous tools relied on rule-based approaches (Gamallo and González, 2012). This strategy was commonly used until the release of manually annotated treebanks.

[1] https://universaldependencies.org/

[2] The proposed model will soon be released and available for community. The newly introduced treebank utilized for training and testing datasets is available within UD last release.

The integration of the Galician-TreeGal treebank[3] within the UD framework facilitated the training of several statistical models for Galician language employing different parsing tools, among them UDPipe 1 and 2. This advancement was particularly evident in the context of CoNLL 2017 and 2018 UD Shared Tasks (Zeman et al., 2017; Zeman et al., 2018). Notably, the best performance of the Galician TreeGal model achieved the Labeled Attachment Score (LAS) equivalent to 74.34% and Unlabeled Attachment Score (UAS) of 79.17% of accuracy[4], using validation mode from raw text. More recently, a sequence labeling approach for dependency parsing using the BERTinho-base monolingual model achieved a higher accuracy with LAS 75.26% and UAS 80.27% using a golden tokenization validation mode (Vilares, Garcia, and Gómez-Rodríguez, 2021).

Nonetheless, the aforementioned results have been obtained using relatively small-sized treebanks. As a result of the utilization of multiple cross-lingual models for the training of the Galician TreeGal parsing model, there was a noticeable improvement in both LAS and UAS metrics. Specifically, the highest achieved accuracy for LAS when employing predicted POS-tags was 70.16%, and 76.54% was attained when utilizing gold POS-tags with the cross-lingual models. In terms of UAS scores, there was an augmentation up to 78.63% when employing predicted POS-tags, and a further increase to 82.43% when employing gold POS-tags (Garcia, Gómez-Rodríguez, and Alonso, 2018).

Current dependency parser models that fuse multiple training corpora along with contextual embeddings have shown significantly improved performance. For instance, using the UDPipe 2 parser, LAS of 77.69% and UAS of 82.72% were achieved with golden tokenization validation mode. Meanwhile, the UDify parser attained an accuracy of LAS 76.77% and UAS 84.08% (Glavaš and Vulić, 2021). Furthermore, a model incorporating UD-specified genre as an alternative signal for data selection achieved LAS of 80.94% and UAS of 85.51% (Müller-Eberstein, van der Goot, and Plank, 2021).

In light of these findings, our hypothesis, as formulated in Section 1, aims to confirm empirically that utilizing cross-lingual data leads to improved model performance.

3 Treebanks

The datasets used for training our new Galician parsing model primarily consist of two treebanks. The first one, Galician-TreeGal v0.42, comprises 1000 sentences, totaling approximately 25,000 tokens. Due to its relatively small size, this dataset contains 600/400 sentences splits for training and testing purposes. While all information except syntactic details was semi-automatically converted to UD format from the original resource[5], dependency labels were assigned using cross-lingual parsing techniques (Garcia, Gómez-Rodríguez, and Alonso, 2018). Subsequently, these labels were manually corrected by an expert linguist.

As studies referenced in Section 2 showed, implementation of extensive treebank datasets for training parsing models enhances its performance. In order to expand the Galician treebanks manually annotated alongside the Galician TreeGal corpus, a new treebank of 1000 sentences (Galician PUD) was created (Sánchez-Rodríguez et al., 2024). These sentences were meticulously reviewed by professional translators and aligned with 23 other available PUD treebanks.

Initial linguistic annotation was conducted using state-of-the-art NLP tools for Galician, followed by thorough review by two experts, resulting in high inter-annotator agreement. As posited by aforementioned authors, during the training phase, scores ranged from 0.83 (κ for the syntactic head) to 0.91 (90.78 UAS) for both annotators, and these scores improved significantly with the final 50 sentences, reaching 93.79 LAS, 96.48 UAS, and $\kappa = 0.96$ for both the Head and Deprel columns. The Galician-PUD has been published in Universal Dependencies v2.14.[6]

4 Experiments

Looking for higher performance of our novel Galician parser model, we conducted a series of experiments on a corpus comprising both TreeGal and Galician PUD treebanks.

[3]We excluded Galician-CTG due to its syntactic analysis not been revised manually. We reference TreeGal because, to the best of our knowledge, it stands as the unique manually annotated Galician treebank.

[4]http://universaldependencies.org/conll17/results.html

[5]http://corpus.cirp.es/xiada/

[6]https://universaldependencies.org/treebanks/gl_pud/index.html

In order to design a definitive dataset for each experiment, we divided the 1000 sentences in training (train), development (dev), and test (test) sets, respectively with 80% (800 sentences), 5% (50 sentences), and 15% (150 sentences) of both TreeGal and Galician PUD corpora. Seeking for fulfillment of our objectives described in Section 1, in concrete evaluating the performance of the newly trained model through the integration of additional data from the Galician PUD into the TreeGal corpus, we created five training splits for our experiments. These splits include the same development (50 TreeGal + 50 Galician PUD), test (150 TreeGal + 150 Galician PUD), and train datasets as follows: 800 TreeGal, 1000 (800 TreeGal + 200 Galician PUD), 1200 (800 TreeGal + 400 Galician PUD), 1400 (800 TreeGal + 600 Galician PUD), and 1600 (800 TreeGal + 800 Galician PUD). The motivation behind these splits arises from our intention to enhance the parsing model performance. We aim to achieve this by systematically incorporating sentences from the Galician PUD training dataset.

4.1 Architecture selection

Treebank	Parser	UAS	LAS
TreeGal	UDPipe 1	81.70	77.50
	UDPipe 2	86.99	82.78
	UDify	84.08	76.77
	TOWER	77.57	66.87

Table 1: Accuracy of the official models trained with standard splits (600/400) of TreeGal, gold tokenization.

Split	Parser	UAS	LAS
800	UDPipe 1	81.60	76.71
	UDPipe 2	88.15	83.08
	UDify	86.62	78.60

Table 2: Accuracy of a novel model trained with the Galician PUD treebank first split in 800/100/300, gold tokenization.

As Table 1 shows, among the results of the four parsers currently utilized for automatic syntactic analysis of Galician texts, UDPipe 2 demonstrates superior metrics in terms of LAS and UAS. Therefore, we performed our initial experiments using this parsing tool and applied our first split of train (800 TreeGal), dev (50 TreeGal + 50 Galician PUD), and test (150 TreeGal + 150 Galician PUD) sets with gold tokenization and sentences to the default settings of the three parsers. Comparing the results from Table 1 and Table 2, we notice a better performance of UDPipe 2 (88.15% UAS and 83.08% LAS), which is superior to both UDPipe 1[7] and UDify. Consequently, we will employ the UDPipe 2 parser for our further experiments since its higher accuracy results. Regarding Table 2 data, its outcomes not only revealed enhanced performance of our parser model, but also showed improvement within the TreeGal parser model itself as expected due to incorporation of extended training and development data.

4.2 Contextualized embeddings

In our pursuit of enhancing the performance of the model proposed in this work, we experimented with three different contextualized vectors extracted from Transformers models, mBERT (Devlin et al., 2019) used by default in UDPipe 2, and the 'base' variant of both Bertinho (Vilares, Garcia, and Gómez-Rodríguez, 2021) and BERT-gl (Garcia, 2021), as Table 3 exposes. Before training a UDPipe2 model, we use the provided scripts to perform the embedding computation stage. Following standard procedures, we used the last four layers of each pretrained model.

We employed the following splits of test datasets in our study: Galician PUD test, comprising 150 sentences from the recently developed Galician treebank; TreeGal test, containing 150 sentences sourced from the Galician TreeGal treebank; and Mixed, a merged dataset combining samples from both sources. This segmentation was implemented in order to observe the evolution of metrics across the treebanks, which include different textual domain. Table 3 shows a consistent enhancement in LAS and UAS metrics using the Mixed test dataset from the initial split to the final one. However, our analysis indicates that the BERT-gl embedding monolingual model, developed specifically for the Galician language, significantly improves the accuracy of the parser proposed

[7] We additionally utilized a pre-trained FastText model, which yielded inferior results in terms of UAS and LAS performance, 81.41% and 75.67% respectively. Therefore, it has been excluded from Table 1.

Splits	Test dataset	UAS mBERT	LAS mBERT	UAS BERT-gl	LAS BERT-gl	UAS Bert-inho	LAS Bert-inho
800	Galician PUD test	88.67	82.84	88.73	82.72	89.05	82.93
	TreeGal test	87.70	83.28	88.55	84.82	87.85	83.71
	Mixed test	88.15	83.08	88.64	83.84	88.41	83.34
1000	Galician PUD test	90.40	85.40	90.63	86.20	90.60	85.54
	TreeGal test	87.72	83.21	88.45	84.59	87.85	83.71
	Mixed test	88.97	84.23	89.47	85.35	89.13	84.57
1200	Galician PUD test	91.35	87.12	91.78	87.67	91.52	87.64
	TreeGal test	88.45	83.84	88.53	84.29	87.82	83.51
	Mixed test	89.81	85.37	90.05	85.87	89.55	85.44
1400	Galician PUD test	91.35	87.90	91.92	88.56	91.81	88.30
	TreeGal test	87.95	83.71	88.55	84.64	88.25	83.94
	Mixed test	89.54	85.67	90.13	86.47	89.91	85.98
1600	Galician PUD test	91.64	88.42	**92.35**	**89.11**	92.27	89.08
	TreeGal test	88.55	83.94	**88.63**	**84.67**	88.10	83.79
	Mixed test	89.99	86.03	**90.37**	**86.74**	90.05	86.26

Table 3: Performance of the parser proposed in our work using three different BERT embedding models.

in the present work. Specifically, the UAS score for the first 800 sentences split using the BERT-gl model achieved 88.64%, escalating to 90.37% in the final 1600 sentences split. Similarly, the LAS score achieved an accuracy of 83.84% during the evaluation of the first split, progressively improving to 86.74% in the final split.

Given the emergence of such complex data, statistical testing became imperative. We opted for the Friedman Aligned Ranks test due to the fact that the experimental data are not normally distributed and the property of homocedasticity is not satisfied. The resulting p-value, determined to be 0.00052 (with a significance level of 0.05), was observed for BERT-gl UAS and LAS scores. This implies that these metrics bear statistical significance within our study

BERT-gl shows higher performance than Bertinho across various treebanks and amounts of training data, whereas mBERT exhibits lower performance. These findings confirm the monolingual models are more efficient for training parsers of the same language, highlighting that increased training data, the primary distinction between BERT-gl and Bertinho, leads to improved parsing results. In light of the previously mentioned results, we employed the BERT-gl model for experiments involving cross-lingual data with UDPipe 2, as it has demonstrated superior performance.

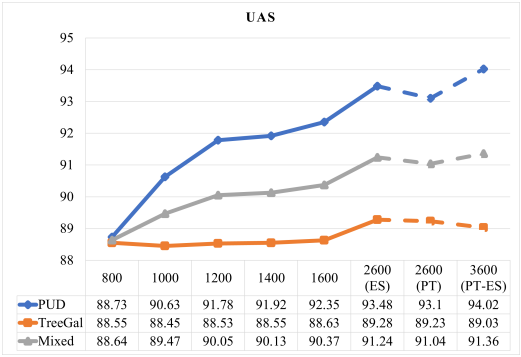
4.3 Introduction of training data from linguistically related languages

In order to test our hypothesis as formulated in Section 1, we aim to incorporate supplementary data from closely related linguistic variations. To accomplish this, we have chosen to include the Parallel Universal Dependency Spanish and Portuguese treebanks in our training dataset. These treebanks are comparable to our Galician treebanks in terms of size (each comprising 1000 sentences) and manual annotation. The results of these cross-lingual data integration are compiled in Table 4.

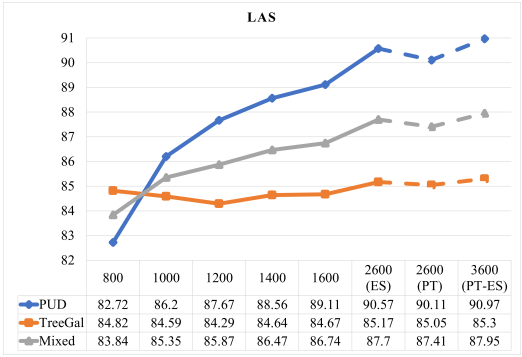
Within the initial split comprising 2600 sentences from the Spanish PUD, TreeGal, and the Galician new treebank for training, we achieved higher UAS (91.24%) and LAS (87.70%) scores compared to the second split involving the Portuguese PUD (UAS of 91.04% and LAS of 87.41%). Nonetheless, the most notable scores were obtained by fusing both Spanish and Portuguese PUD datasets alongside the Galician new and TreeGal treebanks. This 3600 sentences split yielded the highest results, with a UAS of 91.36% and LAS of 87.95%.

Splits	Test dataset	UAS BERT-gl	LAS BERT-gl
2600(+PUD-ES)	Galician PUD test	93.48	90.57
	TreeGal test	89.28	85.17
	Mixed test	91.24	87.70
2600(+PUD-PT)	Galician PUD test	93.10	90.11
	TreeGal test	89.23	85.05
	Mixed test	91.04	87.41
3600(+PUD-ES-PT)	Galician PUD test	94.02	90.97
	TreeGal test	89.03	85.30
	Mixed test	**91.36**	**87.95**

Table 4: LAS and UAS metrics of cross-lingual parsing models using BERT-gl embedding model, compiling Spanish and Portuguese PUD.



(a) Unlabeled Attachment Score.



(b) Labeled Attachment Score.

Figure 1: UAS (left) and LAS (right) BERT-gl monolingual vs cross-lingual learning curve in the different evaluation datasets.

4.4 Learning curves and epochs adjustment

Taking these outcomes into account, we obtained UAS and LAS learning curves, Figure 1 (a and b) for BERT-gl model for monolingual treebank training datasets and compared them with incorporated cross-lingual data dataset. The dashed line graphically illustrates the progressive increase in UAS and LAS scores with the addition of new cross-lingual data. In light of this, the data from Figure 1a shows that the consequential augmentation of the training corpus through the integration of cross-lingual data from related languages has archived 91.36% UAS score of mixed dataset within 3600 sentences split compared to 88.64% UAS of initial 800 split, and 87.89% of LAS within the last 3600 split compared to initial 83.84% 800 split, as Figure 1b evidences. These outcomes indicate improved performance of our automatic parser.

The subsequent set of experiments involved usage of the default parameters and

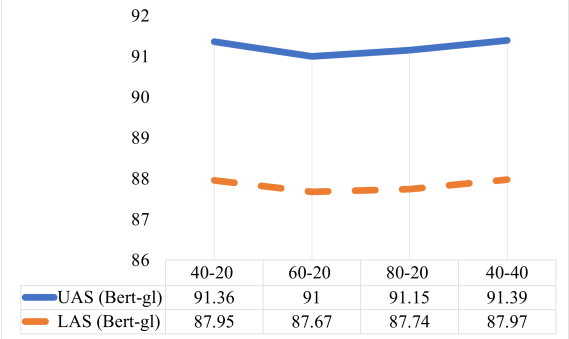


Figure 2: Accuracy variation for the novel model trained using BERT-gl model and 3600 sentences split according to epochs setting, Mixed test.

adjustment the number of epochs, specifically testing variations of 40-20, 60-20, 80-20, and 40-40 (all with initial epochs set at a learning rate of 10^{-3} and final epochs at 10^{-4}), as Lopes and Pardo (2024) propose. The rest of the hyperparameters remained constant, including a batch size of 32, character-level embedding dimension of 256, maximum sentence length of 120, LSTM with a dimension

	Galician PUD		TreeGal	
Number of tokens	**3480**		**3967**	
Type of error	Token #	Total %	Token #	Total %
All correct	3100	89.98	3352	85.50
HEAD	145	4.17	290	7.31
DEPREL	114	3.28	164	4.13
Both incorrect	121	3.48	161	4.06

Table 5: Comparison of annotation accuracy in Galician PUD and TreeGal.

Galician PUD				TreeGal			
DEPREL	**HEAD#**	**Error#**	**Error%**	**DEPREL**	**HEAD#**	**Error#**	**Error%**
conj	95	13	13.68	**punct**	**421**	**145**	**34.44**
appos	50	6	12.00	conj	144	30	20.83
punct	**372**	**40**	**10.75**	parataxis	12	2	16.67
advmod	108	10	9.26	advcl	62	10	16.13
acl	62	5	8.06	acl	103	15	14.56
cc	87	6	6.90	cop	63	7	11.11
cop	45	3	6.67	advmod	175	19	10.86
nmod	292	19	6.51	cc	117	11	9.40
xcomp	31	2	6.45	nmod	326	17	5.21
nummod	40	2	5.00	expl	51	2	3.92

Table 6: Comparison of HEAD annotation error between Galician PUD and TreeGal.

of 512 as the RNN cell type, word embedding dimension of 512, and the BERT-gl for initialization. Following Kann, Cho, and Bowman (2019), we selected the best performing model on the development set to obtain the results on the test datasets.

Figure 2 presents the outcomes of a proposed model training within various epoch parameters we used in UDPipe 2. The UAS and LAS show minimal variation compared to the results from Figure 1. Nonetheless, notable progress was achieved, particularly in the experiment conducted with 40-40 epochs, where we attained a UAS of 91.39% and LAS of 87.97%, as evidenced by Figure 2. This represents the highest accuracy of our parser achieved thus far.

Resuming, Figures 1 and 2 demonstrate a consistent increase in results on Galician PUD as we augment our training data. As indicated by the dashed blue line in Figure 1 1a, the UAS of Galician PUD has achieved a score of 94.02% on the mixed dataset within a split of 3600 sentences, compared to an initial UAS of 88.73% at the 800 sentence split. Similarly, Figure 1a demonstrates that the LAS reached 90.97% within the final 3600 sentence split, compared to an initial 82.72% at the 800 sentence split. However, the enhancements on TreeGal are relatively low. Observing the orange dashed line of Figure 1, the initial split's UAS score of 88.5% has merely increased to last split 89.03% accuracy and 84.82% LAS of 800 split has achieved 85.3% within 3600 split at most. Given these initial findings, it is crucial to conduct a qualitative analysis of parsing performance within both treebanks.

	BERT-gl		mBERT	
	gl-PUD	**TG**	**gl-PUD**	**TG**
Ok	88.98	65.56	87.60	67.93
HD	10.75	34.44	12.40	32.07
DEP	0.0	0.0	0.0	0.0
Both	0.27	0.0	0.0	0.0

Table 7: Comparison of annotation accuracy in Galician PUD and TreeGal across BERT models regarding punctuation. (Ok: all correct, HD: HEAD, DEP: DEPREL, Both: both incorrect).

5 Discussion

To explain the disparity in learning curves between Galician PUD and TreeGal, the examination focused on the 1600 corpus (800

	Punctuation		No punctuation	
	Galician PUD	**TreeGal**	**Galician PUD**	**TreeGal**
LAS	89.11	84.67	90.67	88.71
UAS	92.35	88.63	93.81	92.65

Table 8: Examination of LAS and UAS for BERT-gl considering and excluding punctuation in Galician PUD and TreeGal.

Galician PUD				TreeGal			
DEPREL	**Total #**	**Error #**	**Error %**	**DEPREL**	**Total #**	**Error #**	**Error %**
csubj	5	5	100.00	**flat**	4	4	100.00
compound	2	2	100.00	**flat:foreign**	1	1	100.00
flat	2	2	100.00	**csubj**	9	7	77.78
discourse	8	5	62.50	**parataxis**	12	6	50.00
flat:foreign	3	1	33.33	**aux:pass**	8	4	50.00
ccomp	24	7	29.17	**nsubj:pass**	6	3	50.00
parataxis	14	4	28.57	**advcl**	62	24	38.71
fixed	62	17	27.42	**appos**	13	5	38.46
advcl	25	5	20.00	**fixed**	53	20	37.74
xcomp	31	6	19.35	**iobj**	30	10	33.33
obl	221	37	16.74	**ccomp**	34	8	23.53
iobj	6	1	16.67	**aux**	29	6	20.69
acl	62	10	16.13	**nummod**	25	4	16.00
nmod	292	37	12.67	**obl**	204	32	15.69
nummod	40	5	12.50	**acl**	103	15	14.56

Table 9: Most frequent DEPREL errors in Galician PUD and TreeGal annotations (BERT-gl).

TreeGal + 800 Galician PUD) within the BERT-gl framework, specifically analyzing the annotation of HEAD tags, which identify the governing word in a dependency relation, and DEPREL tags, which denote the syntactic relationship between a dependent word and its head.

The results of this analysis are presented in Table 5. Regarding Galician PUD, comprising 800 sentences, 3480 tokens were identified. Among these, 3100 tokens (89.98%) were accurately annotated, while 145 (4.17%) exhibited errors in HEAD annotation, 114 (3.28%) in DEPREL annotation and 121 (3.48%) in both annotations simultaneously. In the case of TreeGal, also consisting of 800 sentences, a total of 3967 tokens were analyzed. Among these, 3352 tokens (85.50%) were accurately annotated for both parameters, while 290 (7.31%) displayed errors in HEAD annotation, 164 (4.13%) in DEPREL annotation and 161 (4.06%) in both instances concurrently. As evident, there is a difference of over 3% in accuracy between HEAD annotation in Galician PUD and TreeGal.

For this reason, it was decided to analyze in detail the error percentage of the HEAD tags annotation based on their DEPREL (Table 6). Although there is a general increase in the error percentage in TreeGal, the augmentation of error in punctuation is particularly striking, surging by approximately a quarter. Upon further analysis, it was revealed that in Galician PUD, the accuracy rate regarding punctuation stood at 88.98% with 10.75% exhibiting HEAD errors, and a mere 0.27% (one instance) encountering both HEAD and DEPREL errors. Conversely, in TreeGal, sentence accuracy reached 65.56%, with head errors occurring in 34.44% of cases — a discrepancy of nearly 25%. This assessment was corroborated in the context of mBERT to ascertain its independence from model-specific errors, yielding closely aligned

results. In Galician PUD, an accuracy rate of 87.60% was observed, with 12.40% entailing head errors. However, in TreeGal, accuracy was achieved in 67.93% of cases, accompanied by a 32.07% head error rate — nearly 20% higher (Table 7).

Following this, an in-depth analysis was undertaken on LAS and UAS for BERT-gl, with and without considering punctuation, to find potential improvements, as shown in Table 8. When punctuation was considered, the Galician PUD dataset exhibited LAS of 89.11% and UAS of 92.35%. In contrast, the TreeGal dataset showed slightly lower LAS and UAS with scores of 84.67% and 88.63%, respectively. Nevertheless, upon removing punctuation from the embeddings, substantial enhancements were observed. LAS and UAS for the Galician PUD dataset elevated to 90.67% and 93.81% respectively, almost a two-point increase. Similarly, the TreeGal dataset experienced improvements, achieving LAS of 88.71% and UAS of 92.65%, increasing approximately four points. These findings suggest that the punctuation in both TreeGal and Galician PUD was annotated using different criteria, and its removal has a positive impact on the accuracy consistency across both datasets. Therefore, further work should focus on reviewing and standardizing the punctuation labeling criteria in both treebanks.

In relation to the annotation of DEPREL tags and their error rates, an analysis was conducted in both Galician PUD and TreeGal to examine potential disparities concerning the relations with annotation errors. Nonetheless, the results obtained, both in terms of percentage and the most common dependency relations concerning errors, were relatively similar in both cases, as it can be seen in Table 9.

6 Conclusions

The study we have carried out explored several factors in the UD dependency parsing for Galician. In summary, we highlight the following key aspects which meet the initial objectives aroused within our hypothesis formulated in Section 1:

a According to our data, the performance of our parser shows consistent improvement within treebanks with different domains in terms of UAS and LAS metrics while we augment the TreeGal corpus with additional data from the Galician PUD within our five main splits of training dataset.

b By employing three different BERT-based embedding models, we observed that monolingual models provided the superior parsing performance for Galician treebanks.

c Given the data obtained through evaluation of our parsing model's performance subsequent to the integration of cross-lingual data from Spanish and Portuguese data, we obtain higher UAS and LAS scores. These outcomes substantiate the validity of our hypothesis, as the extension of training corpus by incorporation of the aforementioned treebanks from related varieties brings improvement in the performance of our parsing model.

d After conducting a comparative analysis between the syntactic dependencies annotation of Galician PUD and TreeGal, it was shown that LAS and UAS improved in both cases when excluding punctuation, increasing approximately two points in the case of TreeGal and four points in the case of Galician PUD. This emphasises the importance of a thorough analysis and potential revision of punctuation labeling in both treebanks in the future.

Regarding future lines of research, we expect to augment training corpora by adding other treebanks (not only Galician, but from the other domains) and improve the manual annotation to train more robust parsers. In this regard, more additional data from related languages could also lead to improved performance, as shown in previous studies.

Finally, it is worth to mention that our paper contributes with a new treebank and model which achieves what we believe to be the best results in UD parsing for Galician.

Acknowledgments

This work was funded by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by MCIN/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00 and TED2021-130295B-C33, the latter also

funded by "European Union Next Generation EU/PRTR"), and by a Ramón y Cajal grant (RYC2019-028473-I). This publication was produced within the framework of the Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336.

References

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *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.

Gamallo, P. and I. González. 2012. DepPattern: a multilingual dependency parser. In *Demo Session of the International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, pages 17–20. Citeseer.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online, August. Association for Computational Linguistics.

Garcia, M., C. Gómez-Rodríguez, and M. A. Alonso. 2018. New treebank or repurposed? on the feasibility of cross-lingual parsing of romance languages with universal dependencies. *Natural Language Engineering*, 24(1):91–122.

Glavaš, G. and I. Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online, August. Association for Computational Linguistics.

Kann, K., K. Cho, and S. R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *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 3342–3349, Hong Kong, China, November. Association for Computational Linguistics.

Kondratyuk, D. and M. Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Lopes, L. and T. Pardo. 2024. Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 401–410, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Müller-Eberstein, M., R. van der Goot, and B. Plank. 2021. Genre as weak supervision for cross-lingual dependency parsing. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *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.

Sánchez-Rodríguez, X., A. Sarymsakova, L. Castro, and M. Garcia. 2024. Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G.

Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 587–592, Santiago de Compostela, Galicia/Spain, March. Association for Computational Linguistics.

Vania, C., Y. Kementchedjhieva, A. Søgaard, and A. Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *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.

Vilares, D., M. Garcia, and C. Gómez-Rodríguez. 2021. Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*.

Zeman, D., J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In D. Zeman and J. Hajič, editors, *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.

Zeman, D., M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinková, J. Hajič jr., J. Hlaváčová, V. Kettnerová, Z. Urešová, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, and J. Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.