Language Mixture to Develop Accurate Galician Dependency Parsers: An Exploration of Its Effects

Xabier Irastortza-Urbieta[1], José M. García-Miguel[2], Marcos Garcia[3]
[1] HiTZ Center - Ixa, University of the Basque Country
[2] iLingua, Universidade de Vigo
[3] CiTIUS, Universidade de Santiago de Compostela
xabier.irastorza@ehu.eus, gallego@uvigo.gal, marcos.garcia.gonzalez@usc.gal

Abstract

The development of accurate syntactic parsers remains a challenge for low-resource languages. To overcome it, the literature has proposed leveraging syntactic annotations from typologically related languages. This work investigates the viability and adequacy of this approach for Galician, evaluating the use of annotations from major Romance languages as source data. Our methodology extends beyond standard automatic evaluation to incorporate a detailed error analysis, which precisely quantifies the effects of multilingual training and assesses the practical scalability of the method. The results establish the necessity of embedding models for effective cross-lingual transfer and demonstrate that even languages not particularly close can yield adequate parsers. This work confirms the benefits of cross-lingual data augmentation while delineating its scalability limits. Furthermore, the error analysis identifies specific, typologically conditioned grammatical dependencies that remain persistent challenges for accurate dependency parsing.

1 Introduction

Galician is a Romance language spoken in the southwestern part of the Iberian Peninsula, with approximately 2.4 million speakers. It is the official language of the region of Galicia, alongside Spanish. Like many other minoritized languages, it faces significant challenges in developing language technologies, primarily due to a scarcity of data and resources. This scarcity is a particularly critical challenge in the field of syntax parsing, the focus of this work, as it requires costly, expert human annotation to develop high-quality systems.

To address this data scarcity, we focus on Galician's membership in the Romance language family. Other major languages in this family, such as Portuguese, French, and Spanish, benefit from abundant human-annotated data for parser development (Nivre et al., 2020). Building on this, our work investigates methods for combining data from these languages with Galician data to leverage their syntactic similarities and, consequently, develop more accurate syntactic parsers.

1.1 Goal

This work investigates the effects of combining syntactic annotations from different languages, focusing on Galician. We aim to identify the most effective source languages for Galician and analyze the specific parsing errors that are introduced or resolved by incorporating data from other languages. To guide this investigation, we base our work on two research questions:

1. RQ1: Typological Proximity. What is the minimum degree of typological proximity required between Galician and a source language to improve syntactic parsing performance?

2. RQ2: Error Analysis. What are the most prominent syntactic dependency errors in the resulting parsers? To what extent can these errors be attributed to (a) inherent treebank structure, (b) specific language combinations, or (c) the use of cross-lingual embedding models? Furthermore, do these factors primarily exacerbate or mitigate specific error types?

Ultimately, this work aims to propose guidelines for constructing more accurate Galician syntactic parsers. Although focused on a single language, the core approach—leveraging typological proximity within a language family—provides a reproducible framework for improving parsing systems for other low-resource languages (Dione, 2021).

This paper proceeds as follows. We first survey related work in Section 2. Section 3 describes our methodology, the results of which are discussed in Section 4. We present the final conclusions in Section 5 and the limitations in Section 6.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 58–69
March 29, 2026.

Language	Nb. Tr.	Nb. Sent.	UAS/LAS
English	13	97,896	96.8 / 94.4
Portuguese	7	82,183	96.5 / 95.2
Italian	11	42,297	95.8 / 93.8
Basque	1	8,993	88.2 / 85
Galician	**3**	**5,993**	**86.8 / 82.3**
Welsh	1	2,717	88.4 / 82.6

Table 1: Syntax parsing resources of languages. Respectively, the number of treebanks in UD, the sum of the total number of sentences of them and the highest UAS and LAS obtained by each language with the default UDPipe 2 models (version of 2025).

Parser	UAS	LAS	Innovations
UDPipe2	78.66	74.25	Graph-based neural architecture
UDify	84.08	76.77	Multilingual parser
Stanza	77.27	71.86	Bi-LSTM parser
Tower	77.57	66.87	Hierarchical clustering
Sarym.	89.03	85.30	Mixture of languages, embeddings

Table 2: Performance of various models in Galician. The rows follow a chronological order. Scores were obtained by parsing the TreeGal treebank.

2 Related Work

Linguistically, the majority of modern syntactic parsers are based on dependency parsing. The Universal Dependencies (UD) project, a large-scale collection of treebanks, has been instrumental in the widespread adoption of this framework. UD provides treebanks for over 178 languages and has significantly contributed to the standardization of annotation guidelines, ensuring cross-linguistic consistency (de Marneffe et al., 2014).

Grounded in Universal Dependencies, several dependency parsers have been successfully developed using machine learning techniques, including UDPipe (Straka and Straková, 2017) and Stanza (Qi et al., 2020). Subsequently, the rise of multilingual parsers like UDify (Kondratyuk and Straka, 2019) and the creation of algorithms such as TOWER (Glavaš and Vulić, 2021) to combine annotations from different languages were important milestones in the field's development. Furthermore, some authors have proposed using embeddings as input to parsers instead of raw text, demonstrating the positive effects of this approach (Adelmann et al., 2021).

The Large Language Model revolution in recent years has also influenced the field; some authors have employed them to create synthetic training data (Zhang et al., 2024), while others have used them directly as parsers (Ezquerro et al., 2025). Some research has suggested that LLMs are capable of achieving state-of-the-art performance in dependency parsing, at least in some languages (Hromei et al., 2024). However, it is important to note that some authors have found significant performance gaps of LLMs between high- and low-resource languages in common NLP tasks such as translation (Court and Elsner, 2024) or annotation (Jadhav et al., 2025), which may also be the case for dependency parsing in low-resource languages.

With regard to low-resource languages, the lack of manual syntactic annotations has traditionally hindered the training of accurate parsers (Taghizadeh and Faili, 2022). To address this problem, previous work has proposed using annotations from other languages by means of different types of transfer learning (Ghiffari et al., 2023). Table 1 compares the available resources for each language in Universal Dependencies. Note that the major languages have a considerably larger number of annotated sentences; thus, their parsers achieve considerably higher scores on standard metrics, indicating better performance.

Universal Dependencies has released three Galician treebanks, two of which consist of manually revised annotations: TreeGal (Garcia, 2016) and PUD-gl (Sánchez-Rodríguez et al., 2024). It is worth mentioning the treebank PUD-gl, which is the Galician version of a parallel corpus available in several languages; it was first published for the CoNLL Shared Task of 2017 (Zeman et al., 2017).

Regarding dependency parsers for Galician, the earliest systems were rule-based (Gamallo and González, 2012). Subsequently, Galician was included among the languages supported by UDPipe 2 (Straka, 2018), and further research has been conducted to improve these parsers—most notably the study by Sarymsakova et al. (2024). In that study, the authors explored the use of annotations from Portuguese and Spanish for data augmentation, as well as the use of embeddings from several language models, to develop more accurate parsers for Galician. The present work takes that study as its point of departure.

3 Methods

This work uses UDPipe 2 (Straka, 2018) as the basis for creating syntactic parsers, following the approach of several previous studies (Lopes and Pardo, 2024; Sarymsakova et al., 2024). In all experiments, we provided the models with gold-standard tokenization. All treebanks used for training were sourced from Universal Dependencies. For the Galician treebanks PUD-gl and TreeGal, we split the sentences into three sets: from each treebank, 800 sentences were assigned to the training set, 50 to the development set, and 150 to the test set. All other treebanks used in this work were employed exclusively as training data. We excluded the Galician treebank *Corpus Técnico do Galego*, CTG (Guinovart, 2017), because it contains semi-automatically annotated data, unlike the manually annotated treebanks used in this work.

In parallel, across all experiments we investigate the impact of using embeddings as input to the parsers. To this end, we generated embeddings using three language models: Bertinho (Vilares Calvo et al., 2021), BertGL (Garcia, 2021), and XML-Roberta (Conneau et al., 2020). In all cases we have used the *Base* version of those models. The first two are monolingual models trained on Galician, while the latter is a multilingual model. All models were sourced from HuggingFace, and embeddings were calculated using the Wembedding Service tool provided by UDPipe 2.

Evaluation was conducted consistently across all experiments. First, we assessed the general performance of the parsers using two standard automatic metrics: Unlabeled Attachment Score or UAS (Eisner, 1996) and Labeled Attachment Score or LAS (Nivre et al., 2004). We placed greater emphasis on the LAS metric due to its closer alignment with human judgments, as established in prior work (Plank et al., 2015). Second, we performed a fine-grained error analysis, disaggregating results by grammatical dependency types to measure parser precision for each specific relation. This approach allowed us to describe in detail the impact of cross-lingual data mixtures on parser performance and to pave the way for future qualitative analyses. The following subsections describe each experiment.

3.1 Impact of Typological Proximity

In this experiment, we trained parsers using separate, monolingual datasets from Galician and other languages, ensuring an equal number of sentences per language without mixing data across languages within a single training run. Our goal was to confirm that using data from languages typologically close to Galician—even in the absence of any Galician training sentences—can produce effective parsers, which should perform significantly better than parsers trained on data from typologically distant languages. To this end, we selected five high-resource languages representing a gradient of typological proximity to Galician: Portuguese and Spanish (both Ibero-Romance languages, the subgroup to which Galician belongs)[1]; French and Italian (Romance languages outside the Ibero-Romance subgroup); and German (a Germanic, non-Romance language). Thus, the first pair is phylogenetically closest to Galician, while German is the most distant.

To conduct this experiment, we used the Parallel Universal Dependencies (PUD) treebanks for the five languages mentioned above. We selected 800 sentences from each language's PUD treebank and trained all models exclusively on this data. For each language, we trained four models: three models used the embeddings generated by each of the three language models (Bertinho, BertGL, and XML-Roberta) as input, and one baseline model used no embeddings. All models were trained on monolingual data (i.e., each model used data from only one source language).

In line with the above, we created six distinct training data partitions, each corresponding to a single source language: G (Galician), P (Portuguese), S (Spanish), F (French), I (Italian), and D (German). These codes will be used in the following sections to reference the corresponding partitions.

3.2 Comparison Between Treebanks

In this experiment, we compared parsers trained on the two manually annotated Galician treebanks, PUD-gl and TreeGal. The aim was to quantify the impact of differences in annotation guidelines and minor formatting conventions between treebanks, even when they represent the same language. To this end, we constructed two separate training sets of 800 sentences each, one from each treebank. For each set, we trained four parser variants: three using embeddings from the respective language models (Bertinho, BertGL, XML-Roberta) and one

[1] Galician and Portuguese are traditionally considered varieties of the same language with different orthographies, while Galician and Spanish, despite intense contact and a shared writing system, are distinct languages (Carvalho Calero, 1985).

baseline model without embeddings. Finally, we evaluated all trained parsers using the corresponding 150-sentence test sets defined for each treebank in the previous section.

3.3 Mixture of Languages

For this experiment, we created training partitions by combining manual annotations from different treebanks, thereby mixing languages. The purpose was to study the parsing errors that are either introduced or resolved as a direct result of this cross-lingual data combination. Specifically, we created four training partitions based on the PUD-gl Galician treebank. All partitions contained the core set of 800 training sentences from PUD-gl. To this core, we added sentences from other languages to form the following four mixed-data conditions:

1. G+P: PUD-gl + Portuguese (800 Galician + 800 Portuguese sentences).
2. G+S: PUD-gl + Spanish (800 Galician + 800 Spanish sentences).
3. G+P+S: PUD-gl + Portuguese + Spanish (800 Galician + 400 Portuguese + 400 Spanish sentences).
4. G+I+F: PUD-gl + Italian + French (800 Galician + 400 Italian + 400 French sentences).

All partitions consisted of 1,600 sentences each. All trained parsers were evaluated on the 150-sentence PUD-gl test set defined earlier. As in previous experiments, we tested the impact of different embedding types (Bertinho, BertGL, XML-Roberta, and no embeddings).

3.4 Scalability

As a final scalability experiment, we investigated whether large, high-resource treebanks from closely related languages could be used directly—without any pre- or post-processing—to develop more precise parsers for Galician, thereby addressing its data scarcity. For this purpose, we incorporated two additional large treebanks: CINTIL (Portuguese) and AnCora (Spanish).

We constructed three training partitions. Each partition contained the combined training sentences from both Galician treebanks (800 from PUD-gl + 800 from TreeGal = 1,600 sentences). To this Galician base, we added sentences from the large external treebanks to create the following conditions:

1. G+PP: Galician + CINTIL: 1,600 Galician + 2,400 Portuguese (CINTIL) sentences.
2. G+SS: Galician + AnCora: 1,600 Galician + 2,400 Spanish (AnCora) sentences.
3. G+PPSS: Galician + CINTIL + AnCora: 1,600 Galician + 1,200 Portuguese (CINTIL) + 1,200 Spanish (AnCora) sentences.

For each of the three data combinations, we trained four parser variants (using the three embedding types—Bertinho, BertGL, XML-Roberta—and one baseline without embeddings). All models were evaluated on the standard PUD-gl test set of 150 sentences.

4 Results

In this section we will present the results of the experiments described in Section 3, followed by a discussion of them.

4.1 Typological Proximity

UAS-LAS analysis. Figure 1 shows the UAS and LAS scores achieved by parsers trained on different language versions of the PUD treebank, all evaluated on the Galician PUD-gl test set. The parsers trained on Portuguese and Spanish data obtain UAS scores very close to the parser trained on Galician (within 2 and 4 points, respectively), though their LAS scores are somewhat lower (5 and 12 points difference). This proximity in performance is only observed when using embeddings; without embeddings, the performance of parsers trained on Portuguese (P) and Spanish (S) is significantly more limited. For these two closely related languages, the choice between monolingual and multilingual embedding models has a minimal effect.

In contrast, for more typologically distant languages—Italian (I), French (F), and German (D)—the type of embedding model matters substantially. When using the multilingual embedding model, parsers trained on Italian and French achieve UAS scores only 7 points below the Galician-trained parser, compared to a 30–40 point drop for parsers trained without embeddings. However, performance on the stricter LAS metric remains more limited for all parsers trained on languages other than Galician.

Dependency-level analysis. Figure 2 visualizes the error analysis for the trained parsers. Each row corresponds to one of the 14 most frequent dependency relations in the PUD treebank, and

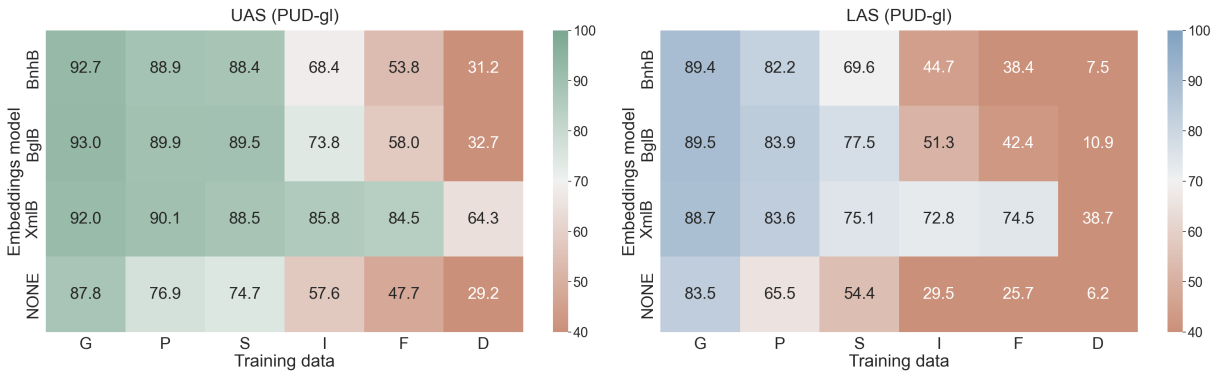

Figure 1: Performance of models trained in a sole language. Both images columns represent a different language on which the parsers was trained and each row a distinct embedding model used by the parser.

each cell shows the precision of correctly inferred dependencies for a given parser. In the row *Others* we calculated the average of that proportion for all other dependencies that were not represented in the Figure, since the total amount of dependencies is of 32.

The analysis reveals that certain dependencies are consistently more difficult to parse across all models, regardless of the source language or embedding type. The most problematic relations are the clausal modifier of noun (*acl*) and numeric modifier (*nummod*), while oblique nominals (*obl*), nominal modifiers (*nmod*), and nominal subjects (*nsubj*) also exhibit high error rates despite their high frequency. As expected, parsers trained on other source languages struggle most with infrequent syntactic dependencies. For instance, even when using embeddings, the precision gap between the Galician-trained parser (G) and the Portuguese-trained parser (P) is of 0.25 points. This gap widens dramatically for more typologically distant languages.

Although embeddings generally improve performance for all languages (noting that distant languages require multilingual embeddings), they enable competitive accuracy on some otherwise problematic dependencies for specific languages (e.g., determinants in Italian and Spanish, or oblique nominals in Portuguese). In that sense, the Figure 2 shows that each source language tends to accentuate specific error patterns. For instance, parsers trained on Portuguese and French more frequently misidentify oblique nominal (*obl*) and nominal modifier (*nmod*) relations, whereas those trained on Spanish and Italian show higher error rates on direct objects (*obj*).

In summary, typological proximity has a stronger positive effect on UAS than on the stricter LAS. Portuguese and Spanish are sufficiently close to Galician to train parsers of decent quality, with results not far from those trained on Galician data itself. Moreover, our experiments reveal a clear performance gradient when using different source languages to train parsers for a target language. For Galician, Ibero-Romance languages (Portuguese, Spanish) yield results closest to the Galician-trained baseline. Other Romance languages (French, Italian) can achieve relevant performance, but only when multilingual embeddings are employed. Parsers trained on German data, however, fail to achieve decent results even with multilingual embeddings.

Notably, the performance gap between using other Romance languages and using German is substantially larger than the gap between Ibero-Romance and other Romance languages. This holds despite the fact that, for example, both French and German are largely intelligible to Galician speakers, suggesting that mutual intelligibility is a poor proxy for parser transferability compared to formal syntactic similarity.

4.2 Mixture of Languages

In this section we present the results of combining diverse languages with Galician in the training of parsers. The Table 3 ilustrates that results.

UAS and LAS improvements. As can be seen in the Table 3, the improvements gained from mixing languages are limited. The results obtained on the PUD-gl test partition by models trained solely on Galician (G) are similar to those obtained by models trained with partition composed by multiple languages (e.g., G+P or G+S). It is true that, when no embedding models are used, language mixing

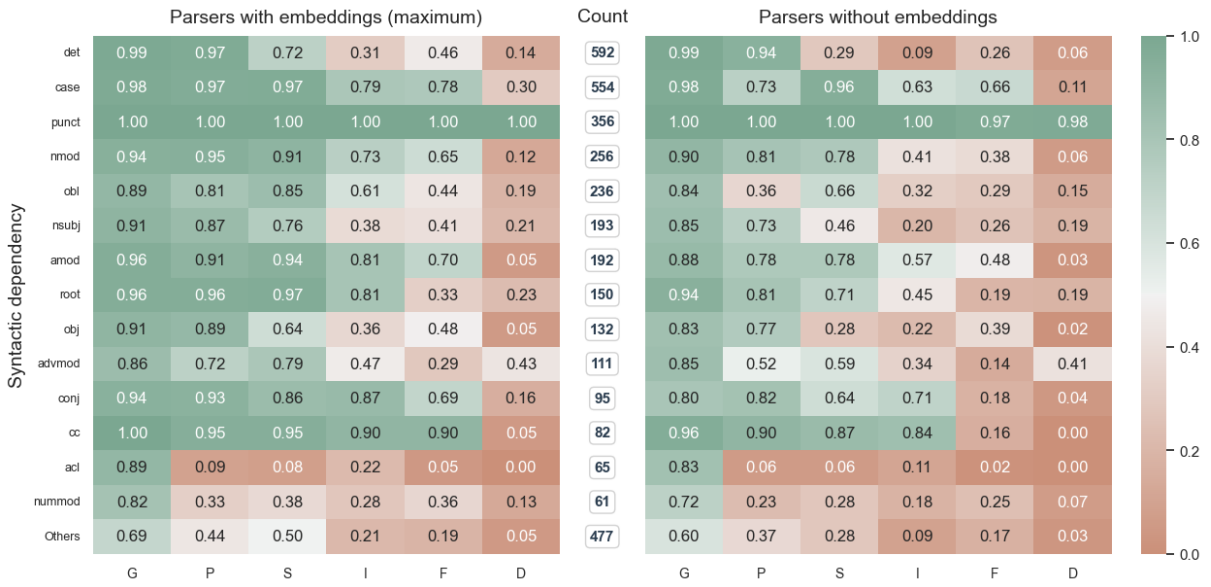

Figure 2: Error analysis disaggregated by dependencies. The left image shows the precision of the parsers that use embeddings (the maximum score obtained between the three models previously explained) and the right image the scores obtained by the parsers without embeddings. *Count* indicates the number of occurrences of each dependency.

Parsers	BglB	XmlB	NONE
G	93.0 | 89.5	92.0 | 88.7	87.8 | 83.5
G+P	93.1 | 89.5	**93.0 | 89.8**	88.0 | **84.9**
G+S	**93.3 | 90.1**	92.6 | 89.3	**89.3** | 84.8
G+P+S	93.1 | 89.9	92.4 | 89.0	87.8 | 84.7
G+I+F	92.4 | 89.2	92.4 | 89.2	87.9 | 83.6

Table 3: Results of language mixture in training. Each row represent a different training partition -thus, a specific language mixture- and each column a different embedding model. Each cell contains: UAS | LAS.

brings greater improvements: the gain between G and G+S models is 0.3 in UAS when using Bert-GL embeddings, and 1.5 when using no embeddings.

Mixing closely related languages does not introduce noise into the training; results always improve, even if only slightly. This is the case for Spanish (G+S) and Portuguese (G+P) with Galician: mixing either with Galician yields small but consistent gains. Increasing the number of mixed languages also does not seem to add more noise to the training, provided the added languages are themselves typologically close (Spanish and Portuguese, G+P+S partition), although the improvements are not greater either. However, as the typological proximity of the mixed languages decreases, they become noisier: combining Italian or French with Galician yields poorer results than training with Galician alone.

Dependency-level error analysis. The analysis follows the trend observed with the automatic metrics. When using embedding models, parsers augmented with data from other languages rarely surpass the performance of the parser trained exclusively on Galician (G), and the observed gains are limited to specific dependencies. It can be affirmed that language mixing is generally unable to yield substantial improvements for the problematic dependencies identified in Section 4.1, as they remain persistent challenges.

However, when embedding models are not used, small but consistent improvements become apparent in several grammatical dependencies: oblique nominals, nominal subjects, adjectival modifiers, etc. In all cases, an improvement of at least 0.02 is observed. When evaluation is performed on the TreeGal treebank, these improvements persist, whether weaker or stronger (see Figure 4). It is noteworthy, furthermore, that for low-frequency dependencies (grouped as *Others*), language mixing has a positive effect, particularly on the TreeGal treebank. Concerning the mixing of specific languages, the most widespread improvements across dependencies are observed when combining Galician and Portuguese (the G+P parsers). There are very few dependencies for which the parsers trained with mixed languages make more errors than those trained only on Galician, even when the added languages are Italian or French.

Depen.	PUD-E	PUD-N	Tre-E	Tre-N
obl	0.02	0.03	**0.06**	0.01
nsubj	0.01	**0.04**	0.02	0.02
amod	0	**0.04**	0.01	0.03
obj	-0.01	**0.03**	0.01	0.01
conj	-0.01	**0.04**	-0.01	0.01
acl	0.03	0.02	**0.07**	0.01
Others	0.02	0.02	0.03	**0.06**

Table 4: Improvements in problematic dependencies. Values indicate the gain in precision when using language mixture training (G+P, G+S, G+P+S, G+I+F) over Galician-only training (G). Columns are organized by test corpus (PUD-gl, TreeGal), with each pair showing results without embeddings (first column) and with embeddings (second column).

4.3 Comparison Between Treebanks

Table 5 compares the performance scores of parsers trained and evaluated on the PUD-gl and TreeGal treebanks, using the four embedding models described previously.

Different performances between treebanks. Despite both treebanks being in Galician and having the same size, models evaluated on the treebank they were not trained on performed significantly worse than those trained and tested on the same treebank. For example, parsers trained on PUD-gl suffered a performance drop of 8–13 UAS points and 11–16 LAS points when evaluated on TreeGal, compared to their evaluation on PUD-gl. Furthermore, parsers trained and evaluated exclusively on TreeGal outperformed those trained on PUD-gl and evaluated on TreeGal, achieving roughly 5 more UAS points and 8 more LAS points. Similarly, parsers trained on TreeGal performed worse when evaluated on PUD-gl compared to parsers trained and evaluated on PUD-gl.

TreeGal more challenging than PUD-gl. The results indicate that the TreeGal treebank appears to present a greater parsing challenge. This is observable in Table 5, where, in some configurations, parsers trained on TreeGal achieve higher scores when evaluated on PUD-gl than when evaluated on their own training treebank, TreeGal (compare columns Tr-Tr and Tr-Pu for UAS).

Embeddings as flexibility providers. This counterintuitive effect occurs only for parsers that use embeddings. The use of embeddings not only improves overall performance but also reduces the performance gap between evaluation on the in-domain treebank (the one used for training) and the out-of-domain treebank (the other treebank). For instance, for parsers trained on PUD-gl with embeddings, the performance drop when evaluating on TreeGal (versus PUD-gl) is approximately 8 UAS points and 12 LAS points. In contrast, for parsers trained on PUD-gl without embeddings, this cross-treebank performance drop increases to 13 UAS and 16.3 LAS points. Therefore, the inclusion of embeddings enhances the parser's generalization capacity, yielding models that are more robust and flexible across different annotation schemes.

Greater performance gap on LAS. Finally, the results demonstrate that the performance gap between in-domain evaluation (same treebank used for training and testing) and cross-treebank evaluation is wider for the LAS than for the UAS—by an average of three points. This is consistent with expectations, as LAS is a more restrictive metric that requires correct grammatical labels in addition to syntactic structure.

Based on the findings presented in this subsection, we conclude that automatic evaluation scores vary considerably across treebanks, even when they represent the same language and are of comparable size. These variations likely stem from differences in annotation quality, guideline compatibility, and the generalization capacity of the parsers themselves.

4.4 Scalability

The last objective of this work was to measure the scalability of language mixing as a step toward building more precise syntactic parsers for Galician. Table 6 presents the results of these experiments. Following what was shown in the previous section, this experiment also demonstrates that language mixing can only yield minor improvements when robust embedding models are used, even if the size of the added partition is five times larger. The UAS difference between the G parser and the scaled parsers (G+PP, G+SS, and G+PPSS) is, at best, 1.1 points in favor of the latter; the LAS difference, however, is 1.7 points. In contrast, when embedding models are not used, the scalability of language mixing is clearly observed: the GCP parser achieves a 3.2-point higher UAS and a 3.5-point higher LAS.

Considering these results, we conclude that scaling syntactic parsers for Galician by directly using large treebanks from closely related languages—without any pre- or post-processing—is not an especially promising approach, given the

Model	UAS					LAS				
	Pu-Pu	Pu-Tr	Tr-Tr	Tr-Pu	Dif.	Pu-Pu	Pu-Tr	Tr-Tr	Tr-Pu	Dif.
BnhB	91.8	82.9	87	88.4	8.9	88.4	76.2	83	83	12.2
BglB	91.4	83.1	87.4	87.9	8.3	88.3	76.8	83.8	82.5	11.5
XmlB	91	82.2	86.7	87.1	8.8	87.7	76	82.6	82	11.7
NONE	88	75	80.5	76.6	13	83.8	67.5	75.1	70.1	16.3

Table 5: Performance comparison across treebanks. The subcolumns represent the following configurations: parser trained and evaluated on PUD-gl (Pu–Pu), trained on PUD-gl and evaluated on TreeGal (Pu–Tr), trained and evaluated on TreeGal (Tr–Tr), and trained on TreeGal and evaluated on PUD-gl (Tr–Pu).

Parsers	BglB	XmlB	NONE
G	93.0 | 89.5	92.0 | 88.7	87.8 | 83.5
G+PP	93.8 | 91.0	93.2 | 90.0	90.5 | 86.7
G+SS	**94.1** | 90.0	93.1 | 90.0	**91.2** | **87.3**
G+PPSS	93.6 | **90.7**	**93.3** | **90.1**	91.0 | 87.0

Table 6: Scalability of language mixture in Galician. See Section 3 for information about parsers and experiment configuration. Each cell contains: UAS | LAS.

limited improvements. Two main reasons may explain this:

- The baseline performance of Galician parsers is already high, achieving UAS and LAS scores between 80 and 95 points. Therefore, the margin for improvement is not substantial, and, as shown in previous sections, only certain specific dependencies remain problematic.
- Galician already benefits from robust embedding models, which are capable of capturing on their own the types of gains that might otherwise be sought through language mixing. Thus, these models present a strong alternative to cross-lingual data augmentation.

Accordingly, we conclude that scaling parsers for Galician requires high-quality data. To effectively leverage language mixing, it would likely be necessary to develop specific pre- and post-processing techniques aimed at improving performance on particular problematic dependencies. However, the data also show that for languages and linguistic varieties with fewer resources than Galician—perhaps with greater interest in the latter—language mixing can still yield significant improvements as an alternative to embedding models.

To conclude the results section, we reproduced the experiments described above, this time evaluating them on the test partition of the TreeGal treebank, and we arrived at similar results. This suggests that our findings are treebank-independent, both for the scalability experiment and for the previous ones, as shown in the data provided in Appendix C.

5 Conclusion

A promising approach for training syntactic parsers in a low-resource language is to leverage syntactic annotations from major languages with significant typological proximity. In the case of Galician, parsers trained exclusively on Portuguese or Spanish can achieve accuracy close to those trained on Galician itself. Furthermore, the use of embeddings significantly enhances parser performance, to the point where models trained on more distantly related languages become viable. However, in the pursuit of more precise Galician parsers, we have identified three key limitations: (i) the gains from mixing Galician annotations with those of larger languages are modest; (ii) parsers exhibit a weak generalization capacity across different treebanks; and (iii) certain syntactic dependencies—such as oblique nominals and adverbial modifiers for Galician—, which could be language specific, remain problematic for parsers trained on any data source. Overall, this work demonstrates that strategically exploiting data from other languages can open pathways for many languages to obtain better parsers.

6 Limitations

An important limitation of this study is one commonly noted in the dependency parsing literature: the lack of diversity in text genres used for training and evaluation. Most available treebanks consist of journalistic and encyclopedic texts, as these are the easiest to crawl from the web. However, natural language is used in a much wider variety of genres—including literary works, transcribed speech, legal documents, and technical writing—which are

not well represented in current resources. Future work should address this gap by creating treebanks with broader genre coverage and subsequently developing parsers adapted to these diverse text types.

A more technical limitation of our study is that we did not perform hyperparameter tuning during parser training, and we used only the UDPipe 2 architecture. Future research could apply our experimental framework to other systems—such as Stanza or LLM-based parsers—to determine whether our findings generalize across different architectures. In that sense, exploring the instruction of the Galician-trained LLM Carballo (Gamallo et al., 2024) for dependency parsing is a promising path. Additionally, future work should undertake a more realistic evaluation by parsing raw text directly, rather than relying on gold tokenization.

Acknowledgments

This paper was funded by MCIU/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62), and by the Galician Government (ED431G 2023/04 and ED431B 2025/16). Xabier Irastortza-Urbieta is supported by a doctoral grant from the Basque Government (PRE_2025_1_0026).

References

Benedikt Adelmann, Wolfgang Menzel, and Heike Zinsmeister. 2021. The impact of word embeddings on neural dependency parsing. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 1–13, Düsseldorf, Germany. KONVENS 2021 Organizers.

Ricardo Carvalho Calero. 1985. O problema ortográfico. *Agália: Publicaçom internacional da Associaçom Galega da Lingua*, 2:127–134.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).

Cheikh M. Bamba Dione. 2021. Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92, Online. Association for Computational Linguistics.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Ana Ezquerro, Carlos Gómez-Rodríguez, and David Vilares. 2025. Better benchmarking LLMs for zero-shot dependency parsing. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 121–135, Tallinn, Estonia. University of Tartu Library.

Pablo Gamallo and Isaac González. 2012. Deppattern: A multilingual dependency parser. In *Demo Session of the International Conference on Computational Processing of the Portuguese Language (PROPOR 2012)*, Coimbra, Portugal.

Pablo Gamallo, Pablo Rodríguez, Iria de Dios-Flores, Susana Sotelo, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open Generative Large Language Models for Galician. *Procesamiento del Lenguaje Natural*, 73(0):259–270.

Marcos Garcia. 2016. Universal dependencies guidelines for the galician-treegal treebank. In *Technical Report*. LyS Group, Universidade da Coruña.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *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. Association for Computational Linguistics.

Fadli Aulawi Al Ghiffari, Ika Alfina, and Kurniawati Azizah. 2023. Cross-lingual transfer learning for Javanese dependency parsing. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational*

Linguistics: Student Research Workshop, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2021. Climbing the tower of treebanks: Improving low-resource dependency parsing via hierarchical source selection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4878–4888, Online. Association for Computational Linguistics.

Xavier Gómez Guinovart. 2017. Recursos integrados da lingua galega para a investigación lingüística. In *Gallæcia: Estudos de lingüística portuguesa e galega*, pages 1037–1048. Universidade de Santiago de Compostela.

Claudiu Daniel Hromei, Danilo Croce, and Roberto Basili. 2024. U-deppllama: Universal dependency parsing via auto-regressive large language models. *Italian Journal of Computational Linguistics*, 10(1):21–38.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. On limitations of LLM as annotator for low resource languages. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Lucelene Lopes and Thiago Pardo. 2024. Towards portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

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.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 49–56, Boston, Massachusetts, USA. Association for Computational Linguistics.

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320, Beijing, China. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Xulia Sánchez-Rodríguez, Albina Sarymsakova, Laura Castro, and Marcos Garcia. 2024. Increasing manually annotated resources for Galician: the parallel Universal Dependencies treebank. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 587–592, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Albina Sarymsakova, Xulia Sánchez Rodríguez, and Marcos García González. 2024. Towards accurate dependency parsing for galician with limited resources. *Procesamiento del Lenguaje Natural*, 73:247–257.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

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

Nasrin Taghizadeh and Heshaam Faili. 2022. Cross-lingual transfer learning for relation extraction using universal dependencies. *Computer Speech & Language*, 71:101265.

David Vilares Calvo, Marcos García González, and Carlos Gómez Rodríguez. 2021. Bertinho: Galician bert representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 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. Association for Computational Linguistics.

Meishan Zhang, Gongyao Jiang, Shuang Liu, Jing Chen, and Min Zhang. 2024. LLM-assisted data augmentation for Chinese dialogue-level dependency parsing. *Computational Linguistics*, 50(3):867–891.

A Available Resources

The code used to produce the results presented in this paper is publicly available at: https://zenodo.org/records/18007073.

B Glossary of Syntactic Dependencies

Through this paper we used abbreviations to reference syntactic dependencies. In this appendix we provide a glossary to explain each abbreviation. In general, the abbreviations we have used are the same employed by Universal Dependencies in all their treebanks.

- *Det*: Determinant.
- *Case*: Case marking.
- *Punct*: Punctuation.
- *Nmod*: Nominal modifier.
- *Obl*: Oblique nominal.
- *Nsubj*: Nominal subject.
- *Amod*: Adjectival modifier.
- *Root*: Root.
- *Obj*: Direct object.
- *Advmod*: Adverbial modifier.
- *Conj*: Conjunct.
- *Cc*: Coordinating conjunction.
- *Acl*: Clausal modifier of noun.
- *Nummod*: Numeric modifier.

C Supplementary Results

In this appendix, we provide a complementary set of results for the experiments described in Section 3, this time evaluating them on the TreeGal treebank. This contrasts with the primary results presented in the main paper, which were based on evaluation using the PUD-gl treebank. The strong similarity between the two sets of results reinforces the conclusions presented in Section 5.

With regard to **typological proximity**, Table 7 reiterates the results presented in Section 4.1. The influence of typological proximity when training parsers on different source languages remains a key factor for achieving high performance. Embedding models continue to provide a considerable improvement. It should be noted that the absolute results on TreeGal are lower than those on PUD-gl, consistent with the analysis in Section 4.3.

Parsers	BglB	XmlB	NONE
G	83.1 | **77.5**	81.7 | **76.3**	**77.1** | **70.3**
P	**83.5** | 76.2	**83.3** | 75.9	69.2 | 56.8
S	82.1 | 70.8	80.7 | 67.8	67.4 | 48.2
I	68.1 | 48.8	80.5 | 69.3	53.0 | 29.2
F	57.0 | 42.0	78.7 | 69.5	46.6 | 26.2
A	30.4 | 10.0	57.9 | 33.3	28.4 | 5.8

Table 7: Results from the typological proximity experiment (Section 4.1), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

Parsers	BglB	XmlB	NONE
G	83.1 | 77.5	81.7 | 76.3	77.1 | 70.3
G+P	84.6 | 79.4	83.2 | 77.5	78.1 | 71.4
G+S	84.5 | 79.1	**83.3** | 77.6	77.7 | 71.0
G+P+S	**85.0** | **79.7**	83.1 | 77.1	**78.2** | **71.6**
G+I+F	83.9 | 78.2	83.1 | **77.8**	78.1 | 71.2

Table 8: Results from the language mixture experiment (Section 4.2), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

Language mixture. Table 8 exposes the results for the parsers from Section 4.2—trained via language mixture—when evaluated on the TreeGal treebank. Note that all parsers were trained on the PUD-gl treebank. We reach the same conclusion as in the main analysis: while language mixture can improve results, its benefits are subject to significant limitations.

Also, the results of the **scalability** experiment (Section 4.4) can be reproduced when evaluating on the TreeGal treebank, as shown in Table 9. Note that in this table, the parser in the row labeled *TRE* was trained exclusively on the 800-sentence training partition of TreeGal (in order to do a fair comparison). This same Galician data was also included in the training sets of the scaled parsers (G+PP, G+SS, G+PPSS), as described in Section 3.

On the other hand, the conclusions drawn from **error analyses** are also reproducible when evaluating on the TreeGal dataset. Figure 3 shows that the set of problematic dependencies remains similar to that identified using the PUD-gl treebank. For

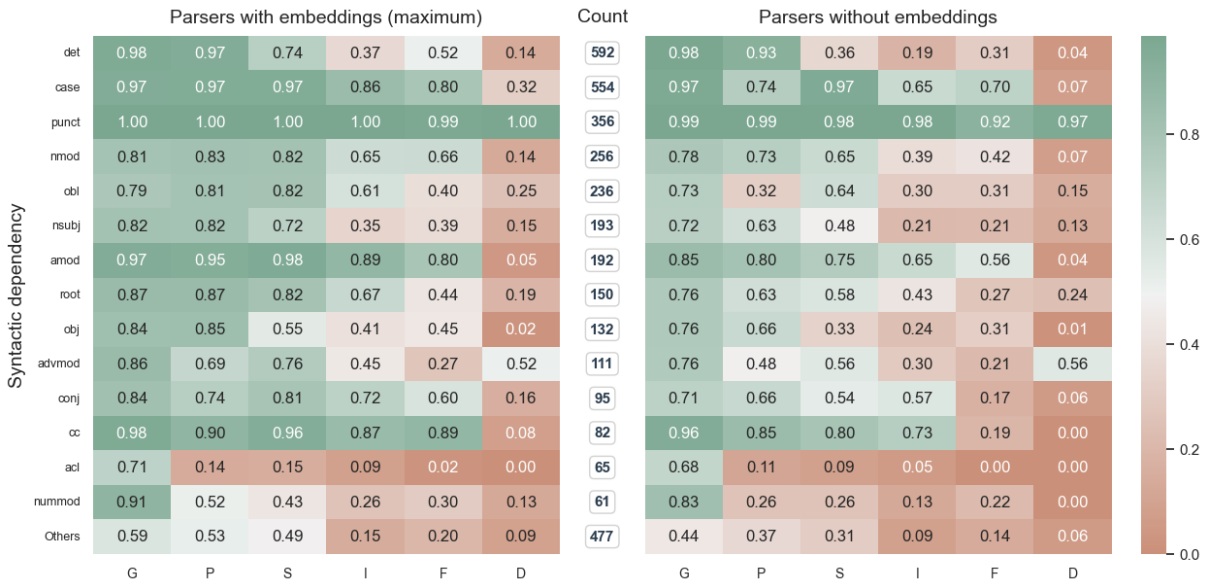

Figure 3: Error analysis disaggregated by dependencies, in the same format as Figure 2. The evaluation was conducted on the test partition of the TreeGal treebank.

Parsers	BglB	XmlB	NONE
TRE	88.8 | 85.7	86.8 | 83.7	81.5 | 76.5
G+PP	89.4 | 86.3	88.1 | 84.5	83.5 | 79.0
G+SS	**90.0 | 87.0**	**89.5 | 86.6**	**86.0 | 82.0**
G+PPSS	89.5 | 86.4	88.6 | 85.4	84.6 | 80.1

Table 9: Results from the scalability experiment (Section 4.4), evaluated on the TreeGal test partition. Each cell contains: UAS | LAS.

further analysis, the complete code used to generate the results in this paper is available via the link provided in Appendix A.

C.1 Ambiguity in Parsing

Finally, it should be noted that certain errors identified during the evaluation reflect **annotation decisions that are inherently debatable or ambiguous** even for human annotators. To illustrate this point, we provide the following examples:

- In the sentence "*Curio, [...], informou persoalmente a César das accións de Pompeio*" ("Curio, [...], personally informed Caesar of Pompey's actions"), the parser employing BglB embeddings (parser G) predicts *César* as a direct object, whereas the gold standard classifies it as an oblique nominal. This case is ambiguous due to the fuzzy boundaries in some cases between direct objects (obj), indirect objects (iobj), and obliques (obl) in Galician.
- The same parser frequently exhibits confusion between discourse elements and adverbial modifiers (advmod), as in the examples: *Porén, cando o Senado lle respondeu categoricamente, prohibindolle...* ("However, when the Senate answered him categorically, prohibiting him...") and *Agora, este departamento enfróntase a novos desafíos* ("Now, this department faces new challenges"). In such instances, the distinction between the adverbial function as a circumstantial modifier (advmod) and its role as a discourse marker is often subtle and not clearly defined.

Consequently, future qualitative analyses should also consider the annotation conventions of Universal Dependencies, along with the potential confusions or ambiguities they may introduce.