Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 29-41 recibido 22-02-2024 revisado 07-05-2024 aceptado 16-05-2024

Multi-label Discourse Function Classification of Lexical Bundles in Basque and Spanish via transformer-based models

Clasificación Multietiqueta de la Función Discursiva de Conjuntos Léxicos en Euskera y Español mediante Modelos Basados en Transformers

Josu Goikoetxea,[1] **Markel Etxabe,**[1]
Marcos García,[2] **Eleonora Guzzi,**[3] **Margarita Alonso**[3]
[1]Euskal Herriko Unibertsitatea
[2]Universidade de Santiago de Compostela
[3]Universidade da Coruña
josu.goikoetxea@ehu.eus, metxabelizarazu@gmail.com,
marcos.garcia.gonzalez@usc.gal,{eleonora.guzzi,margarita.alonso}@udc.es

Abstract: This paper explores the effectiveness of transformer-based models in the discourse function multi-label classification of lexical bundles task in two languages, Basque and Spanish. The study has a dual focus: firstly, to evaluate the impact of manually and automatically annotated datasets in the fine-tuning for this task; secondly, to demonstrate the efficiency of multilingual language models in a cross-lingual transfer learning context for this task. First and foremost, our findings reveal their ability to generalize discourse function classification of lexical bundles beyond specific sequence of words forms in the mentioned task in both monolingual and cross-lingual transfer learning contexts. In the former setting, this research highlights the superiority of manually annotated datasets over the automatically annotated ones as long as dataset size is sufficiently large. In the latter case, despite the transfer learning occurring between two typologically different languages, results also suggest the superiority of manually annotated datasets along with the capability to surpass the monolingual results when ratios of target and source language training and fine-tuning corpora are balanced.
Keywords: lexical bundles, discourse function classification, manual annotation, multilingual transfer learning.

Resumen: Este artículo explora la efectividad de los modelos basados en transformers en la clasificación multietiqueta de la función discursiva de tareas de conjuntos léxicos en dos idiomas, euskera y español. El estudio tiene un doble enfoque: en primer lugar, evaluar el impacto de los conjuntos de datos anotados manual y automáticamente en el fine-tuning para esta tarea; en segundo lugar, demostrar la eficiencia de los modelos de lenguaje multilingües en un contexto de aprendizaje de transferencia entre idiomas para esta tarea. En primer lugar, nuestros resultados revelan la capacidad de los transformers de generalizar la clasificación de funciones discursivas de conjuntos léxicos más allá de las formas específicas de secuencia de palabras, en contextos tanto de aprendizaje monolingüe como de transferencia de aprendizaje entre idiomas. En el primer contexto, esta investigación destaca la superioridad de los conjuntos de datos anotados manualmente sobre los anotados automáticamente, siempre que el tamaño del conjunto de datos sea lo suficientemente grande. En el último, a pesar de que el aprendizaje de transferencia ocurre entre dos idiomas tipológicamente diferentes, los resultados también sugieren la superioridad de los conjuntos de datos anotados manualmente, así como la capacidad de superar los resultados monolingües cuando se equilibran las proporciones de los corpus de entrenamiento y ajuste fino en el idioma objetivo y de origen.
Palabras clave: conjuntos léxicos, clasificación de funciones discursivas, anotación manual, aprendizaje de transferencia multilingüe.

ISSN 1135-5948 DOI 10.26342/2024-73-2

1 Introduction

Lexical bundles (LB) are defined as recurrent lexical sequences with high frequency and dispersion (Biber et al., 1999). They are n-grams which are automatically extracted from a corpus using statistical metrics. We refer to sequences such as *it is clear that*, *as can be seen*, *it should be noted*, etc. In the context of academic English, LBs have been studied extensively (Hyland, 2008a; Simpson-Vlach y Ellis, 2010; Biber, Conrad, y Cortes, 2004). In particular, they have been used to provide resources for novice writers (Granger y Paquot, 2015). The interest of LBs lies in the discourse function they are associated with. This is because although many of them are compositional, their relevant meaning does not come from their propositional meaning, but from their discourse meaning. Thus, a novice writer has to know that *it should be noted* is one of the ways to highlight their statement, but *as can be seen* serves to resend something mentioned previously.

With the aim of building up resources helping novice writers in academic Spanish and Basque, we compiled lists of academic LBs, referred here as the umbrella term of *formulae* (Alonso-Ramos y Zabala, 2022). These formulae comprise sequences of diverse nature, including discourse markers, connectors and modal operators, as well as other elements that are not easily classified through a specific tag. With these lists of formulae, we have annotated both a Spanish (Guzzi et al., 2023) and a Basque corpus, and, to the best of our knowledge, these corpora are the first to be annotated in this way. The goal of this annotation has been to obtain a gold-standard corpus to train and evaluate Large Language Models (LLM) on the identification and classification of academic formulae in new corpora. The automatic identification and classification is a challenge for these models, and an open question is whether discourse functions can be effectively modeled using distributional approaches. In our study, we aim to address this question by leveraging BERT-like models. Indeed, the fine-tuning framework of BERT-like (Devlin et al., 2018) encoder-only models have proven to be suitable for capturing intricate and nuanced discourse relations (Kishimoto, Murawaki, y Kurohashi, 2020; Hou, 2020), as they effectively grasp intricate linguistic relationships and contextual dependencies within the discourse, critical for identifying diverse discourse functions that are often context-sensitive. Furthermore, the capacity to fine-tune a single multilingual model on multiple languages allows cross-lingual transfer learning, making them a cost-effective and efficient choice in the discourse field (Kurfali y Östling, 2021; Liu, Shi, y Chen, 2020), especially in resource-scarce languages like Basque. For these reasons, this study employes BERT-like models in order to tackle the multi-label classification task in Basque and Spanish. Moreover, due to the importance of the manually annotated datasets in the discourse field in NLP (Nie, Bennett, y Goodman, 2019; Sileo et al., 2019), we also intend to evaluate the impact of those datasets in the multi-label classification of formulae, comparing its results with those derived from automatically constructed ones.

2 Related Work

The development of large pre-trained language models has proved to be very effective in most of the NLP fields. Discourse-related tasks like discourse parsing (Koto, Lau, y Baldwin, 2021; Liu, Cohen, y Lapata, 2019), RST parsing (Xiao, Huber, y Carenini, 2021; Kwon et al., 2021) or argumentation mining (Fergadis et al., 2021) have also been successfully tackled by BERT-like models.

The studies most related to our research are indeed those dealing with discourse analysis (Chiarcos, 2022; Ru et al., 2023), mostly focused on the identification of connectors that establish discourse relations between sentences (Zhou et al., 2010; Braud y Denis, 2016). This type of analysis is crucial for understanding how the information is structured within a discourse, and its implementation can be useful for different NLP tasks, such as natural language generation (Callaway, 2003; Leopold, Mendling, y Polyvyanyy, 2014), sentiment analysis (Mukherjee y Bhattacharyya, 2012; Bayoudhi, Ghorbel, y Belguith, 2015), textual entailment (Pan et al., 2018) or machine translation (Meyer y Webber, 2013; Hardmeier, 2014).

While the majority of research on discourse markers has been conducted in English, contributions in other languages like Basque and Spanish, the languages under examination in this study, have also surfaced. One of the few contributions in the discourse field in Basque is from (Iruskieta et al., 2013),

who created the Basque discourse TreeBank[1] annotated with Rhetorical Structure Theory (RST). Regarding the Spanish language, (da Cunha, Torres-Moreno, y Sierra, 2011) also released a RST treebank[2] and (Guzzi et al., 2023) published a corpus with academic lexical bundles annotated with discourse functions.

More recent work have explored the use of automated methods to identify discourse markers from corpora. For example, (Nazar, 2021) introduced a method for identifying and categorizing discourse markers in English, Spanish, German and French using statistical analysis and clustering methods. Although the majority of works related to discourse marker predictions with transformers are in English (Pandia, Cong, y Ettinger, 2021; Huber y Carenini, 2022), some authors have extended it to the cross-lingual scenario and to other languages. For example, (Kurfali y Östling, 2021) evaluated several transformer-based multilingual models in various discourse-related tasks, and Spanish was included. Similarly, (Liu, Shi, y Chen, 2020) trained a RST multilingual parser out of a language independent shared semantic space derived from multilingual models, including Basque and Spanish among others. The work of (Toro, Zamorano, y Moreno-Sandoval, 2022) also use transformer-based models for conducting a binary classification of discourse markers in Spanish in the financial field, achieving a F1-score of 0.933.

3 Dataset creation

In this section we summarize the creation of the two types of datasets employed in this research, that is, the manually annotated one and the automatically annotated one.

3.1 Basque and Spanish academic corpora

The point of departure of the research are two academic corpora, one in Basque (Aranzabe, Gurrutxaga, y Zabala, 2022) and the other one in Spanish (Villayandre y others, 2018; Salido et al., 2018). The Basque corpus encompasses 295 Bachelor's theses and 105 Master's theses, altogether totaling 3.2 million tokens. Conversely, the Spanish corpus comprises 413 research articles (Salido et al., 2018) and 176 Bachelor's and Master's theses (Villayandre y others, 2018), containing a total of 5 million tokens. In this study, we haven not distinguished between research articles, Master's and Bachelor's theses within the corpus analysis, thus this distinction we acknowledge is an avenue for potential future investigation.

3.2 Label definition

Before creating any of the mentioned datasets, the authors agreed the labels which referred to the different discourse functions with the annotators, who decided to group all labels in three different discourse groups that encompass their respective discourse functions. These discourse functions have been organized according to a three-fold division (Biber, Conrad, y Cortes, 2004; Hyland, 2008b): *EST*, which is related to text structure; *REF* for referring to research content; *READER* for positioning and addressing the reader.

Thus each label comprises two annotation levels, the discourse group and the discourse function itself, both of them separated by an underscore. In Spanish we defined 38 labels and in Basque we added 3 extra labels to the latter ones. In light of this, we cite the following sentence in example 1 to illustrate the use of the formulae "sailkatzen da" in Basque:

(1) Saiakera bost ataletan **sailkatzen da**[3]

This discourse function associated with this formulae is part of the REF group and the SETGROUPS subgroup which is meant to establish groups in the discourse. Example 2 shows the formulae "por eso" in Spanish, which belongs to the EST group and the EXPCAUS subgroup used for expressing cause:

(2) **Por eso** nos manejamos con las cifras antes citadas[4]

In this paper, our analysis has focused primarily on the classification of formulae into discourse functions. Given the complexity and potential significance of this aspect, we have chosen to defer a comprehensive analysis of the discourse groups and their corresponding subgroups to future research.

[1] `http://ixa2.si.ehu.eus/diskurtsoa/en/`

[2] `http://www.corpus.unam.mx/rst/`

[3] In English, "The essay **is categorized into** five parts".

[4] In English, "**Thus** we handle the previously mentioned ciphers.

This paper addresses the automatic identification and multi-label classification of academic *formulae* in both monolingual and cross-lingual settings in Basque and Spanish using automatically and manually annotated datasets in both settings.

3.3 Automatically annotated dataset

As mentioned before, in our research we have employed an automatic annotation process to annotate both corpora with their respective labels. In order to facilitate this annotation, our annotators have curated for each language a dictionary that links lexical bundles with their corresponding labels.

It is important to highlight that in the annotation process it was observed that certain lexical bundles exhibited the potential to be associated with multiple discourse functions. To maintain clarity and simplicity when automatically annotating the corpus, the annotators included only the most frequent discourse function label for the ambiguous formulae. In the Basque dictionary the annotators delineated 366 formulae within the *EST* group, 209 within the *READER* and 290 within the *REF* one. Conversely, the Spanish one encompassed 428 within the *EST* group, 122 within the *READER* and 464 within the *REF*.

Employing the aforementioned dictionaries as reference, we developed an in-house script for the automated labeling of both corpora and also for splitting the latter into training, development and test datasets.

The resulting output for the automated labelling part generated BIO annotated datasets in CONLL format that comprise the entirety of corpora for each of the languages. The Basque corpus is composed by 87,140 sentences and 54,794 labelled formulae while the Spanish one encompassed 146,362 sentences and 68,856 labelled formulae. The percentage between annotated to non-annotated sentences in both languages approximates 33 %.

In initial experimentation, we partitioned the annotated corpora randomly and conducted classification experiments. However, we noted that the F1 scores exceeded 98 % because the lexical forms present in the test subset also occurred in the training and development ones. Within this scenario, it is not possible to prove the models' capacity for generalizing the discourse function classification task beyond word forms. We adressed the latter issue following the next splitting strategy for both labelled corpora: first, only sentences with one labelled formulae were considered; second, only labels with more than five formulae were taken into account; third, 20 % of the formulae were randomly set aside to create the test set; last, 80 % of the remaining formulae were employed to create the training and development sets.[5] Note that with this strategy we make sure that the lexical bundles found exclusively in the test set do not overlap with the training and development data. This partitioning is motivated by our desire to assess the generalization capabilities of transformers, given that the formulae present in the test set remain unobserved within the training and development. Sticking to the same proportions as previously described for the partitioning, we allocated all non-annotated sentences to the three respective sets. We have intentionally divided the data this way to prevent any overlap of word forms between the test subset and the training and development subsets. However, it is worth emphasizing that we recognize the importance of the excluded data segments, and we are actively considering their inclusion in future research.

As a result, the filtered Basque dataset encompasses 72,130 sentences, comprising 15,027 annotated instances, and the Spanish dataset has a total of 125,601 sentences, with 27,261 of them being annotated. The ratio of annotated to non-annotated sentences in both languages in the filtered dataset approximates the 22 %. Furthermore, the division of these datasets into training, development, and test subsets adheres to the proportions mentioned in the previous paragraph.

Note that when applying the partitioning script, although it doesn't have a significant impact on the overall number of sentences within each dataset, both languages undergo a reduction in the amount of the annotated lexical bundles.

3.4 Manually annotated dataset

The specifics of the manual annotation process and annotation criteria for the Spanish dataset are comprehensively outlined in the work by (Guzzi et al., 2023). Following these established criteria, the Basque corpus un-

[5] Out of which 75 % are used for the training set and 25 % for the development one.

derwent manual annotation, and a detailed account of the annotation procedure for the Basque dataset will be provided in an upcoming publication.

As a result, we obtained a 43,909 sentence corpus with with 56,271 annotated lexical bundles in the Basque corpus, and 82,175 in the Spanish one with 60,715 annotated lexical bundles, both of them labelled via BIO annotation and in CONLL format. Note that the the proportion of annotated bundles per corpus size in comparison to the automatically annotated ones is higher.

After manually annotating the corpora, we employed the same filtering procedure as used in the automatically annotated dataset. This resulted in 1,842 annotated sentences for the Basque dataset and 5,522 for the Spanish dataset, which constitute 12 % and 20 % of the size of their respective automatically annotated counterparts. In order to ensure that the datasets are large enough for training and evaluation, we introduced more non-annotated sentences than in the automatically created ones, thus decreasing the annotated and non-annotated ratio to approximately 10 %, but we kept the same partitioning proportions for splitting the dataset in training, development and test subsets.

4 Experimental setting

The present section is dedicated to detail the experimental settings of the monolingual and cross-lingual experiments of the research.

4.1 Monolingual experiments

This section is divided in the following parts: the description of the experimental setting with the automatically annotated datasets, the setting with the manually annotated datasets and the baseline experimental setting for both types of datasets.

4.1.1 Transformer-based experiments

In a preliminary phase we conducted experiments employing several transformer-based models from the Hugging Face library.[6] These experiments were carried out in the discourse function classification task for both languages with the automatically created datasets, including both monolingual and multilingual models.

From this preliminary exploration, we retained the top-performing models (comprising multilingual and monolingual) for each language. We executed a grid search with various learning rates, including 1e-6, 5e-6, 7e-6, 1e-5, 2e-5, 3e-5, 4e-5 and 5e-5, while keeping the rest of the parameters in the default setting. For the large models we doubled the epoch numbers to 6, because we observed that these modifications were necessary to optimize the convergence and performance of these models.

In our initial experiments involving the automatically annotated Basque datasets, we observed that `xlm_r_large` (Conneau et al., 2020) and `ixambert` (Otegi et al., 2020) proved to be the top-performing multilingual models, while the best-performing monolingual models included `rob_eu_large` (Mikel Artetxe, 2022) and `bert_eu` (Agerri et al., 2020).

In the context of Spanish, the initial experiments revealed that `ixambert`, `xlm_r_large` and `mdebertav3_base` (He et al., 2021) were the top-performing multilingual models. Among the monolingual models, `beto_uncased` (Cañete et al., 2020) and `rob_es_base` (Fandiño et al., 2022) showed the best performance.

We carried out the manually annotated dataset experiments employing the top-performing models over the automatically annotated ones, to enable a comparative analysis of the models' performance when confronted with two distinct types of datasets.We applied the same grid-search criteria as in the automatically annotated ones, but due to the smaller size of the corpora we incremented the epoch number to 10. Both automatically and manually annotated dataset results are shown side-by-side in table 1. At this point it is important to emphasize that even though we acknowledge that the results of these two types of datasets are not directly comparable, we also consider that the side-by-side comparison of their results shall provide insights into the impact of manual review for training models.

4.1.2 Baseline experiments

As our baseline approach, we employed a Bi-LSTM-CNN-CRF neural architecture, as proposed by (Ma y Hovy, 2016), and we implemented it using the PyTorch framework, following the method described by (Chernodub et al., 2019).[7] For both the Basque and Spanish languages, we initialized the network

[6] `https://huggingface.co/`

[7] `https://github.com/achernodub/targer`

using official *fastText* embeddings.[8]

During training, we conducted 50 epochs and selected the model that exhibited the best performance on the development set. We used the Adam optimizer, set a learning rate of 0.001, and employed 100-dimensional vectors. Additionally, we maintained the rest of the hyper-parameters at their default settings.

4.2 Cross-lingual experiments

The present section explains the experimental setting of cross-lingual experiments in this research.

4.2.1 Transformer-based experiments

In the cross-lingual transfer learning setting with automatically annotated datasets we employed the two top-performing multilingual transformer-based models from the monolingual experimentation setting, `ixambert` and `xlm_r_large`, which encompassed the Basque language. The multilingual model `mdebertv3` yielded very competitive results in the monolingual setting for Spanish, consistent with the findings in (Agerri Gascón y Agirre Bengoa, 2023). However, the latter model has a suboptimal performance with Basque language in both monolingual and cross-lingual settings, so that it has been excluded from the cross-lingual transfer learning experiment. We also performed experiments with `mbert` in both settings and it yielded significantly lower results in both scenarios when compared to `ixambert` and `xlm_r_large` for both languages, leading to its exclusion as well. Due to space limitations in the paper, we have not included the results for these two models, so our focus was on highlighting the best-performing models for a more concise and informative presentation of results.

Given the variety of dataset sizes and combinations in the fine-tuning of the cross-lingual setting with automatically annotated datasets, we didn't conduct a grid-search. In the case of `xlm_r_large`, we decreased the learning rate to 1e-5 and increased the epochs to 6, in `ixambert` we kept the default parameters.

In order to create a framework for the cross-lingual transfer learning with automatically created datasets, we integrated both the target and source languages within the bilingual training and development datasets, introducing varying proportions of the source language (10 %, 30 %, 50 % and 100 %) into these subsets, while preserving the original data in the target language subset. Furthermore, when inserting each subset proportion, we randomly chose three subsets for each proportion and averaged the results. Using the subsets outlined in Section 3.3, it is worth noting that we prevent the overlap of lexical bundles between the test subset and the training and development ones for both target and source languages. This approach was employed during the fine-tuning phase, thus exploring the strategy with Basque as target language and Spanish as source one, and vice-versa (see tables 2 and 3 respectively).

In the evaluation of these strategies, we carried out the testing for both bilingual configurations in both languages, enabling a comparative analysis of the contributions made by the two configurations for each target language. In order to further explore the contributions of the source language in the transfer learning, we also compared side-by-side the results of each bilingual configuration (tested on the target language) with those obtained from proportions of the target language subset (see table 4).

Regarding the cross-lingual setting with manually annotated datasets, we followed a simpler approach. Due to the smaller size of the datasets, we shuffled both languages' labelled datasets in the training and development, and tested them in Basque and Spanish labelled test subsets (see table 5). We also proceeded with the same grid-search as in its monolingual manually annotated counterpart.

4.2.2 Baseline experiments

In the context of our cross-lingual experiments, we made use of the same setting as described in its monolingual counterpart, but initializing the network with bilingual embeddings (Basque-Spanish) that were previously mapped using the unsupervised method of *vecmap* (Artetxe, Labaka, y Agirre, 2018).

5 Results

In this section, we analyze the outcomes of two distinct categories of experiments.

5.1 Monolingual experiments

In this setting we compare side-by-side the results of the top-performing transformer-

[8] FastText embeddings by (Grave et al., 2018).

based models both monolingual and multilingual for Basque and Spanish.[9]

Table 1 presents the precision, recall and F1-score results of the aforementioned experiments. The table is structured as follows: on the one hand, the first five rows are dedicated to the Basque language and the last five rows to Spanish; on the other hand, first three columns show the automatically constructed datasets' results, while the last three show the manually constructed ones. For the Basque language, as mentioned in Section 4.1.1, we report the results for the baseline model, followed by the top-performing models, specifically, `ixambert` (imbrt), `xlm_r_large` (xml), `rob_eu_large` (robeu), and `bert_eu` (bereu), across both dataset modalities. Similarly, the Spanish language section provides results for the baseline model, along with `xlm_r_large`, `mdebertv3` (mdb3), `beto_uncased_large` (beto), and `rob_es` (robes).

		Aut. created			Man. created		
		P	R	F1	P	R	F1
eu	*base*	0.54	0.22	0.31	**0.63**	**0.47**	**0.54**
	imbrt	**0.69**	0.65	**0.64**	0.67	**0.71**	**0.64**
	xlm	**0.80**	**0.66**	**0.72**	0.54	**0.66**	0.6
	robeu	**0.78**	**0.68**	**0.73**	0.66	0.67	0.66
	bereu	**0.73**	0.65	**0.69**	0.60	**0.66**	0.63
es	*base*	0.50	0.39	0.44	0.80	0.52	0.63
	imbrt	0.78	0.62	0.70	**0.82**	**0.78**	**0.79**
	xlm	**0.83**	0.73	0.77	0.82	**0.80**	**0.81**
	mdb3	0.79	0.67	0.72	**0.84**	**0.82**	**0.83**
	robes	0.82	0.58	0.64	**0.86**	**0.78**	**0.82**
	beto	**0.81**	0.68	0.74	0.79	**0.76**	**0.78**

Tabla 1: Results in the monolingual experiments, **test in corresponding language**. Best results within a model and across both dataset configurations in bold, best results within a language across all models and both dataset configurations underlined.

5.2 Cross-lingual experiments

This section resumes the results of the experiments described in in section 4.2.

5.2.1 Automatically annotated dataset

Table 2 presents the results of the cross-lingual experiments with the following table structure: within each five-row group, we focus on a multilingual transformer-based model, with each row representing varying proportions of the source language to be inserted into the bilingual datasets; on the other hand, first three columns display the results of the bilingual dataset using Basque as target language and Spanish as source one, with test in Basque. The remaining three columns show the results of the opposite dataset configuration, also with test in Basque. Conversely, table 3 has the same structure, but with the test in Spanish. We also indicate the zero-shot (*zs*) scenario results in all cases. Due to the space constraints and the comparatively low performance of the baseline in this experimental setting, we have omitted its results from this study.

		Autom. created (**test eu**)					
		es+ %eu			eu+ %es		
		P	R	F1	P	R	F1
imbrt	zs	0.57	0.07	0.12	-	-	-
	10 %	0.63	**0.59**	0.61	0.71	0.58	0.64
	30 %	0.65	0.57	0.6	0.74	**0.59**	0.65
	50 %	0.68	**0.59**	0.63	0.74	**0.59**	**0.66**
	100 %	**0.75**	0.58	0.65	**0.76**	0.57	0.65
xlm	zs	0.49	0.07	0.13	-	-	-
	10 %	0.53	0.59	0.56	0.62	0.64	0.62
	30 %	0.6	0.63	0.61	0.70	0.65	0.67
	50 %	0.64	0.65	0.64	**0.71**	**0.66**	**0.68**
	100 %	**0.74**	**0.68**	**0.71**	**0.71**	**0.66**	**0.68**

Tabla 2: Results in the cross-lingual experiments with Basque dataset plus percentage of Spanish dataset, **test in Basque**. Best results within the same model and dataset configuration in bold, and best results across all models and dataset configurations underlined.

We also show a comparison of the results to quantify the contribution of an external language, so that we compare side-by-side the results of the proportions of Basque and Spanish datasets and their corresponding bilingual counterparts. Table 4 indicates, on the one hand, the F1-score results of the Basque and Spanish proportion datasets, and, on the other hand, the contribution of an external language to the former[10] represented as absolute gains with respect to the former results.

5.2.2 Manually annotated dataset

Lastly, we summarize all results of the tree types of previous experiments[11] along with

[9] In this research we have used the ISO-639-1 language codes for Basque and Spanish, *eu* and *es* respectively.

[10] Extracted from F1 column of tables 2 and 3.

[11] That is, monolingual scenario and automatically and manually created dataset, and cross-lingual transfer scenario and automatically created dataset,

		Automat. created (**test es**)					
		es+ %eu			eu+ %es		
		P	R	F1	P	R	F1
imbrt	zs	-	-	-	0.07	0.02	0.03
	10 %	0.77	0.57	0.65	0.55	0.54	0.54
	30 %	0.75	0.54	0.63	0.73	0.59	0.65
	50 %	0.77	0.55	0.64	0.71	0.53	0.61
	100 %	0.77	0.54	0.63	0.78	0.56	0.65
xlm	zs	-	-	-	0.03	0.02	0.02
	10 %	0.75	**0.73**	0.72	0.48	0.6	0.54
	30 %	0.79	0.69	0.74	0.69	**0.68**	0.69
	50 %	**0.82**	0.7	**0.76**	0.69	0.67	0.68
	100 %	0.81	0.68	0.74	**0.8**	0.67	**0.73**

Tabla 3: Results in the cross-lingual experiments with Spanish dataset plus percentages of Basque dataset, **test in Spanish**. Best results within the same model and dataset configuration in bold, and best results across all models and within a dataset configuration underlined.

		Autom. created			
		test eu		**test es**	
		%eu	es+ %eu	%es	eu+ %es
imbrt	10 %	0.34	**+0.27**	0.49	**+0.05**
	30 %	0.59	**+0.01**	0.62	**+0.03**
	50 %	**0.64**	-0.01	**0.65**	-0.04
	100 %	0.64	**+0.01**	**0.67**	-0.02
xlm	10 %	0.26	**+0.30**	0.39	**+0.15**
	30 %	0.44	**+0.17**	0.58	**+0.11**
	50 %	0.56	**+0.08**	**0.7**	-0.02
	100 %	0.67	**+0.04**	0.72	**+0.01**

Tabla 4: Results that compare side-by-side the F1-scores of portions of Basque and Spanish datasets with their bilingual counterparts, with **test in corresponding language**. The results of bilingual datasets show the absolute gain with respect to its monolingual counterpart. Best results between a proportion and its bilingual counterpart in bold, and best results across all models and within a dataset configuration underlined.

the scores of the cross-lingual transfer learning experiment with manually annotated datasets in table 5. As mentioned before, by comparing the performance of models trained on both types of datasets, researchers can assess the effectiveness of manual annotation and thus have a deeper understanding of different annotation approaches (see section 6) for informing future decisions regarding dataset creation and model training strategies for discourse function classification task.

their results have been extracted from 1, 2 and 3, respectively.

			P	R	F1
eu	*base*	cross man.	0.81	0.56	0.66
	imbrt	mono auto.	**0.69**	0.65	0.64
		mono man.	0.67	0.71	0.64
		cross auto.	0.74	0.59	0.66
		cross man.	0.71	**0.69**	**0.70**
	xlm	mono auto.	**0.80**	0.66	**0.72**
		mono man.	0.54	0.66	0.60
		cross auto.	0.74	0.68	0.71
		cross man.	0.71	**0.72**	**0.72**
es	*base*	cross man.	0.72	0.51	0.6
	imbrt	mono auto.	0.78	0.62	0.70
		mono man.	0.82	**0.78**	0.79
		cross auto.	0.78	0.56	0.65
		cross man.	**0.84**	0.77	**0.80**
	xlm	mono auto.	**0.83**	0.73	0.77
		mono man.	0.82	0.80	0.81
		cross auto.	0.82	0.7	0.76
		cross man.	0.82	**0.83**	**0.82**

Tabla 5: Comparison of best results of all types of experiments conducted in the study. Best result of a model within a language in bold, and best result among all models within a language underlined. **Test in corresponding language**.

6 Discussion

This section is a summary of the main findings of this research.

Transformer-based models prove to be able to generalize discourse functions. On the one hand, results in tables 1 and 5 show that transformer-based models outperform the baseline in every possible configuration. On the other hand, given the configuration of the datasets, which avoids the overlap of formulae between training/development subsets and the test one, results in practically all tables indicate that transformers have the ability to classify discourse functions without relying on specific word forms. The capacity of transformers to handle tasks that require intricate comprehension of text beyond surface-level features in several NLP tasks features is well-known. This ability has been previously acknowledged in the context of binary discourse function classification for Spanish, as demonstrated by Toro, Zamorano, y Moreno-Sandoval (2022). Thus we have extended these insights to the considerably more complex multi-label classification scenario as well as to the cross-lingual transfer learning one in our current study.

Multilingual and monolingual transformer-based models exhibit si-

milar performances in the monolingual setting. The results in table 1 fail to provide substantial evidence regarding the comparative performance between multilingual and monolingual models in the discourse function classification task. This observation does not entirely align with the findings in Agerri Gascón y Agirre Bengoa (2023), which suggested that large multilingual models outperformed their monolingual counterparts across various tasks. It is important to emphasize that our work is focused on discourse function classification while the mentioned authors research covers a wide variety of tasks. Discourse function classification of lexical bundles places a strong emphasis on contextual understanding and subtleties in language use, making it a unique testing ground for model performance.

A manually annotated dataset of sufficient size yields superior results compared to automatically annotated ones in the monolingual setting. Results in table 1 show that in the monolingual setting transformer-based models and the baseline exhibit enhanced performance in Spanish when utilizing manually annotated datasets as opposed to automatically created ones, while in the case of the Basque language, the results indicate the opposite trend. Considering the disparity in dataset sizes, this observation suggests that beyond a certain dataset size threshold, manual curation can significantly enhance model performance in token classification tasks. In order to back this statement, we conducted an experiment reducing the manually annotated dataset in the best model in Spanish from table 1, `mdbertv3`, making it equivalent in size to the Basque dataset. Its results were considerably reduced, under-performing its automatically annotated counterpart.[12] The sensitivity of the fine-tuning dataset size will also appear in the last finding which is related to the cross-lingual scenario.

Transformer-based models demonstrate suboptimal performance in zero-shot learning scenarios within the multilingual setting. Results in tables 2 and 3 demonstrate that transformer-based models are unable to perform zero-shot discourse function classification tasks effectively in a multilingual context. This phenomena has already been discussed by Lauscher et al. (2020), who observed that large multilingual models exhibit limited performance in zero-shot transfer to distant target languages, particularly for languages with limited monolingual data for pre-training. The languages in our study, Basque and Spanish, fulfill both aspects.

Automatically annotated datasets in the cross-lingual scenario do not outperform their monolingual counterparts. A comparison of the results for both languages in the automatic setting of the datasets in `xlm_r_large` and `ixambert` models in table 1 with their cross-lingual counterparts in tables 2 and 3 show that employing automatically annotated datasets is not effective in the cross-lingual setting. The only expcetion is `ixambert` with the *eu+ %es* dataset setting, where the language model has enough target language corpus in the training and the source language contribution in the bilingual dataset is big enough. We consider this phenomena interesting aspect to be studied in the future, utilizing a language model with enough amount of data in both target language corpus training and source language dataset.

Automatically annotated datasets show their effectiveness in small datasets sizes in the cross-lingual setting. Results in table 4 show that the influence of an external language is evident within the range of 10 % to 30 %, with the only exception of `xlm_r_large` in the Basque setting that extends to the 100 % of the configuration. The contribution of the Spanish to Basque portions of corpus (es+ %eu column) seems to be more pronounced than the opposite, and this could be because the size of the Spanish training corpus in the models is bigger than the Basque one, thus being more difficult to improve the former's results. However, note that despite the difference of Basque and Spanish languages, Basque contributes significantly in the 10 % to 30 % range with `xlm_r_large`. Despite those results, none of them in any of the models of table 4 outperform the their monolingual counterparts in table 1.[13]

Manually annotated dataset in the

[12] Precision, recall and f1 scores dropped from 0.84, 0.82 and 0.83 to 0.708, 0.685 and 0.696, respectively.

[13] The only exception is the eu+ %es column result in `ixambert` model with test in Basque, which slightly surpasses its respective monolingual counterpart (0.65 vs 0.64, respectively).

cross-lingual transfer learning scenario benefits low-resourced languages. Surprisingly, table 5 shows that the only combination where manually annotated datasets outperform the rest of the setting, and particularly its monolingual counterpart, is in the Basque as target language using `ixambert` model, and not `xlm_r_large` as expected. The ratio between Basque and Spanish in the source fine-tuning datasets is aproximately 33 %, and the ratios of the training corpora in the mentioned languages in `ixambert` and `xlm_r_large` are 32 % and 2.8 % respectively.[14] This finding suggest that cross-lingual transfer learning with small sized manually annotated datasets surpasses its monolingual counterpart when the distribution of training corpora and fine-tuning datasets of target and source languages is quite comparable, with the target one being lower-resourced. In the rest of the cases when target language has more resources like Spanish or the dataset distributions are not comparable, this setting does not seem perform effectively the transfer learning.[15] Apparently, this phenomena doesn't seem to be aligned with the previous finding, where `xlm_r_large` effectively conducted the transfer learning in the 10-30 % range of dataset portions. Considering the substantial difference in size between automatically and manually annotated fine-tuning datasets, with the 10 % range one in the former doubles the latter, this results suggests that the cross-lingual transfer learning capabilities of these models are highly sensitive to the fine-tuning dataset size. Therefore, to enhance those models' performance, it is crucial for the sizes of these datasets to be more comparable as they reduce their size. Recognizing the promise of cross-lingual transfer learning for low-resourced languages in discourse function classification, we aim to conduct an in-depth research of this particular field in future works.

7 Conclusion

The study reveals several insights into the use of transformer-based models for the discourse function classification task. First, our findings demonstrate the generalization capabilities of transformer-based models for classifying nuanced discourse functions without being reliant on specific word forms. Additionally, the study emphasizes the critical role of dataset size and curation in model performance. Manually annotated fine-tuning datasets significantly outperform automatically annotated ones in the monolingual setting as long as the dataset size is sufficient. This finding suggests, on the one hand, that rich and contextually nuanced information is necessary for addressing discourse function classification, and, on the other hand, the high sensitivity of the models to the dataset sizes. Lastly, the cross-lingual transfer learning results show a more complex scenario with manually annotated datasets, in which achieving the right balance between the size of the source language training corpus and size of fine-tuning dataset is needed in order to successfully harness the transfer learning capabilities of multilingual models. Thus automatically annotated datasets would be a viable alternative in two cases: first, in monolingual scenarios where manually annotated data is unavailable or doesn't yield a sufficient size; second, in cross-lingual transfer learning scenarios with small-sized automatically annotated dataset in the target language and a large one in the source side.

This preliminary study holds particular significance for low-resourced languages like Basque in the discourse function classification field, as the insights derived from our findings can serve as valuable guidance.

[14] Corpus sizes extracted from Otegi et al. (2020) and Conneau et al. (2020)

[15] The only exception is `xlm_r_large` with target language in Spanish, but the difference of the cross-lingual setting with respect of the monolingual one is not statistically significant (P-value=0.7657).

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation through the projects PID2019-109683GB-C21 and PID2019-109683GB-C22. We also acknowledge the support of the Basque Government through the grant IT-1570-22; the Xunta de Galicia, through the aid ED431C 2020/11; the Centro de Investigación do Sistema Universitario de Galicia, funded by Xunta de Galicia and the European Union (FEDER GALICIA 2014-2020), through the aid ED431G 2019/01; and the Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020.

References

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, y

E. Agirre. 2020. Give your text representation models some love: the case for basque. En *Proceedings of the 12th International Conference on Language Resources and Evaluation*.

Agerri Gascón, R. y E. Agirre Bengoa. 2023. Lessons learned from the evaluation of spanish language models.

Alonso-Ramos, M. y I. Zabala. 2022. Hartaes-vas: Combinaciones léxicas para una herramienta de ayuda a la redacción de textos académicos en español y en vasco. En *Pre-conference Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022). Co-located with the Conference of the Spanish Society for Natural Language Processings*, páginas 25–28.

Aranzabe, M. J., A. Gurrutxaga, y I. Zabala. 2022. Compilación del corpus académico de noveles en euskera hartaeus y su explotación para el estudio de la fraseología académica. *Procesamiento del Lenguaje Natural*, 69:95–103.

Artetxe, M., G. Labaka, y E. Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 789–798, Melbourne, Australia, Julio. Association for Computational Linguistics.

Bayoudhi, A., H. Ghorbel, y L. H. Belguith. 2015. Sentiment classification of arabic documents: Experiments with multi-type features and ensemble algorithms. En *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, páginas 196–205.

Biber, D., S. Conrad, y V. Cortes. 2004. If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405.

Biber, D., S. Johansson, G. Leech, S. Conrad, y E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Longman.

Braud, C. y P. Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. En *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, páginas 203–213, Austin, Texas, Noviembre. Association for Computational Linguistics.

Callaway, C. B. 2003. Integrating discourse markers into a pipelined natural language generation architecture. En *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, páginas 264–271, Sapporo, Japan, Julio. Association for Computational Linguistics.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, y J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. En *PML4DC at ICLR 2020*.

Chernodub, A., O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann, y A. Panchenko. 2019. Targer: Neural argument mining at your fingertips. En *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics (ACL'2019)*, Florence, Italy.

Chiarcos, C. 2022. Inducing discourse marker inventories from lexical knowledge graphs. En *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, páginas 2401–2412, Marseille, France, Junio. European Language Resources Association.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, y V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 8440–8451.

da Cunha, I., J.-M. Torres-Moreno, y G. Sierra. 2011. On the development of the RST Spanish treebank. En *Proceedings of the 5th Linguistic Annotation Workshop*, páginas 1–10, Portland, Oregon, USA, Junio. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, y M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Fergadis, A., D. Pappas, A. Karamolegkou, y H. Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. En *Proceedings of the 8th Workshop on Argument Mining*, páginas 100–111.

Granger, S. y M. Paquot. 2015. Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica*, 31(1):118–141.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, y T. Mikolov. 2018. Learning word vectors for 157 languages. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, Mayo. European Language Resources Association (ELRA).

Guzzi, E., M. Alonso-Ramos, M. Garcia, y M. García Salido. 2023. Annotation of lexical bundles with discourse functions in a Spanish academic corpus. En *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, páginas 99–105, Dubrovnik, Croatia, Mayo. Association for Computational Linguistics.

Hardmeier, C. 2014. *Discourse in statistical machine translation*. Ph.D. tesis, Acta Universitatis Upsaliensis.

He, P., X. Liu, J. Gao, y W. Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. En *International Conference on Learning Representations*.

Hou, Y. 2020. Fine-grained information status classification using discourse context-aware bert. En *Proceedings of the 28th International Conference on Computational Linguistics*, páginas 6101–6112.

Huber, P. y G. Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. En *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 2376–2394.

Hyland, K. 2008a. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Hyland, K. 2008b. Genre and academic writing in the disciplines. *Language Teaching*, 41(4):543–562.

Iruskieta, M., M. J. Aranzabe, A. D. de Ilarraza, I. Gonzalez, M. Lersundi, y O. L. de Lacalle. 2013. The rst basque treebank: an online search interface to check rhetorical relations. En *4th workshop RST and discourse studies*, páginas 40–49.

Kishimoto, Y., Y. Murawaki, y S. Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. En *Proceedings of the Twelfth Language Resources and Evaluation Conference*, páginas 1152–1158.

Koto, F., J. H. Lau, y T. Baldwin. 2021. Top-down discourse parsing via sequence labelling. *arXiv preprint arXiv:2102.02080*.

Kurfali, M. y R. Östling. 2021. Probing multilingual language models for discourse. En *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Bangkok, Thailand, August 1-6, 2021*.

Kwon, J., N. Kobayashi, H. Kamigaito, y M. Okumura. 2021. Considering nested tree structure in sentence extractive summarization with pre-trained transformer. En *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, páginas 4039–4044.

Lauscher, A., V. Ravishankar, I. Vulić, y G. Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Leopold, H., J. Mendling, y A. Polyvyanyy. 2014. Supporting process model validation through natural language generation. *IEEE Transactions on Software Engineering*, 40(8):818–840.

Liu, J., S. B. Cohen, y M. Lapata. 2019. Discourse representation structure parsing with recurrent neural networks and the transformer model. En *Proceedings of the IWCS shared task on semantic parsing*.

Liu, Z., K. Shi, y N. Chen. 2020. Multilingual neural rst discourse parsing. En *Proceedings of the 28th International Conference on Computational Linguistics*, páginas 6730–6738.

Ma, X. y E. Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1064–1074, Berlin, Germany, Agosto. Association for Computational Linguistics.

Meyer, T. y B. Webber. 2013. Implicitation of discourse connectives in (machine) translation. En *Proceedings of the Workshop on Discourse in Machine Translation*, páginas 19–26.

Mikel Artetxe, Itziar Aldabe, R. A. O. P.-d.-V. A. S. 2022. Does corpus quality really matter for low-resource languages?

Mukherjee, S. y P. Bhattacharyya. 2012. Sentiment analysis in twitter with lightweight discourse analysis. En *Proceedings of COLING 2012*, páginas 1847–1864.

Nazar, R. 2021. Automatic induction of a multilingual taxonomy of discourse markers. *Electronic lexicography in the 21st century: postediting lexicography. Brno*, páginas 440–454.

Nie, A., E. Bennett, y N. Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 4497–4510.

Otegi, A., A. Agirre, J. A. Campos, A. Soroa, y E. Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. En *Proceedings of The 12th Language Resources and Evaluation Conference*, páginas 436–442.

Pan, B., Y. Yang, Z. Zhao, Y. Zhuang, D. Cai, y X. He. 2018. Discourse marker augmented network with reinforcement learning for natural language inference. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 989–999, Melbourne, Australia, Julio. Association for Computational Linguistics.

Pandia, L., Y. Cong, y A. Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint arXiv:2109.12951*.

Ru, D., L. Qiu, X. Qiu, Y. Zhang, y Z. Zhang. 2023. Distributed marker representation for ambiguous discourse markers and entangled relations. En *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 5334–5351, Toronto, Canada, Julio. Association for Computational Linguistics.

Salido, M. G., M. Garcia, M. Villayandre-Llamazares, y M. A. Ramos. 2018. A lexical tool for academic writing in spanish based on expert and novice corpora. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sileo, D., T. Van-De-Cruys, C. Pradel, y P. Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. *arXiv preprint arXiv:1903.11850*.

Simpson-Vlach, R. y N. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4):487–512.

Toro, A. G., J. P. Zamorano, y A. Moreno-Sandoval. 2022. A discourse marker tagger for spanish using transformers. *Procesamiento del Lenguaje Natural*, 68:123–132.

Villayandre, M. y others. 2018. "harta" de noveles: un corpus de español académico. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 5(1):131–140.

Xiao, W., P. Huber, y G. Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. *arXiv preprint arXiv:2104.07058*.

Zhou, Z.-M., Y. Xu, Z.-Y. Niu, M. Lan, J. Su, y C. L. Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. En *Coling 2010: Posters*, páginas 1507–1514, Beijing, China, Agosto. Coling 2010 Organizing Committee.