Check for updates

Natural Language Processing (2026), **32**, pp. 555–588
doi:10.1017/nlp.2026.10040

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Named Entity Recognition Across Datasets and Domains: Resources and Models for Galician

Albina Sarymsakova[1,2], Ettore Mariotti[2], Helena Pérez Puente[2] and Marcos Garcia[2]

[1]Instituto de Estudios Gallegos Padre Sarmiento (IEGPS-CSIC), Santiago de Compostela, Galicia, Spain and [2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Galicia, Spain
Corresponding authors: Albina Sarymsakova; Email: albina.sarymsakova.es@gmail.com; Marcos Garcia; Email: marcos.garcia.gonzalez@usc.gal

(Received 21 March 2025 UTC; revised 23 January 2026 UTC; accepted 16 August 2026 UTC;
first published online 1 October 2026 UTC)

Abstract

Automatic named entity recognition (NER) is essential for many natural language processing applications, particularly in low-resource languages and those with corpora restricted to a single domain, where the lack of diverse data may hinder cross-domain generalisation. In this context, we attempt to contribute to the Galician NER field by extending manually annotated resources and exploring several approaches to improve the performance of automatic classifiers. To this end, we enrich two new datasets with NER annotations in Galician, which we release in this paper. Through experimental evaluation, we investigate the benefits of fine-tuning transformers on NER data by comparing various monolingual and multilingual BERT models and developing a new extra-small DeBERTa model for Galician. Furthermore, we assess the quality of the NER model by expanding the initial training corpus with additional data from related varieties, including Portuguese and Spanish. For evaluation, we use five Galician corpora, including our two newly annotated datasets and three adapted from existing resources, ensuring assessment across multiple domains. Our findings reveal (1) a positive impact of enhanced cross-lingual training data on improved model performance across five Galician test datasets; (2) similar performance of monolingual and multilingual models depending on the training data; and (3) the competitive effectiveness of extra-small BERT models for Galician NER tasks compared to larger models.

Keywords: Named entities recognition; Galician; BERT; natural language processing

1. Introduction

Named entity recognition (NER) is a fundamental task in information extraction (IE) and plays a central role in natural language processing (NLP). It involves identifying and classifying specific entity types within large volumes of unstructured text. As a key preprocessing step, NER underpins numerous NLP applications, including IE (Lu *et al.* 2022; Lezama-Sánchez and Vidal 2024) and machine translation (Babych and Hartley 2003; Nowakowski *et al.* 2022), among many others. Beyond traditional NLP domains, NER is also essential in areas such as anonymisation and pseudonymisation, where it helps protect sensitive data while preserving its contextual relevance (Mina *et al.* 2024). Additionally, it plays a significant role in Digital Humanities by enabling the extraction and analysis of textual insights (Blouin *et al.* 2021). These diverse applications highlight the pivotal role of NER across both technical and interdisciplinary fields.

A central challenge in NER for low-resource languages such as Galician is the scarcity of naturally occurring annotated training data (Torge *et al.* 2023; Khade *et al.* 2025). To mitigate this

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

limitation, the aforementioned works have explored various transfer learning strategies, for example, leveraging language families to enable knowledge transfer between related linguistic varieties. In such cases, the limited amount of training data available for a target language can be partially compensated for by shared linguistic features within the same language family.

However, and despite advances in transfer learning and pre-trained language models that have reduced the dependence on large manually annotated corpora, the scarcity of annotated NER datasets continues to pose a major challenge for low-resource languages (Yadav and Bethard 2018; Keraghel, Morbieu, and Nadif 2024; Bade, Kolesnikova, and Oropeza 2024). In the case of Galician, only a limited number of resources are currently available. An early contribution to NER for Galician by Garcia *et al.* (2012) introduced an annotated dataset. However, due to its limited sample size, it has been used exclusively to evaluate rule-based NER models. Additionally, Garcia and Gamallo (2014) released a potentially useful dataset, but, to the best of our knowledge, it has not been used for training or evaluating NER models, which has led to the limited number of existing Galician NER models. More recently, the Corpus Técnico do Galego (Galician Technical Corpus) for Named Entity Recognition (CTG-NER) dataset released by Agerri *et al.* (2018) has served as the only benchmark for assessing NER models in this language, resulting in the limited number of existing Galician NER models being tested exclusively on this dataset, which constrains a broader understanding of their performance across diverse contexts.

The absence of alternative datasets significantly limits the assessment of model robustness and generalisation across domains (Vania *et al.* 2019). This limitation underscores a critical gap in the Galician NER field, which remains a scarce resource and is insufficiently empirically validated. Consequently, there is a pressing need for (i) the creation of additional Galician NER datasets, (ii) the adaptation and reuse of existing linguistic resources for NER tasks, and (iii) systematic, cross-dataset evaluations of NER models to enable reliable performance comparisons and foster progress in this under-resourced language.

Based on these outcomes and to advance automatic NER solutions for Galician, this study presents the following contributions:

- Creation of two novel, manually reviewed datasets for the evaluation of existing and newly developed NER models for the Galician language.
- Adaptation and release of two supplementary datasets derived from previous research (Garcia *et al.* 2012; Garcia *et al.* 2014).
- Comprehensive evaluation and fine-tuning of nine pre-trained monolingual and multilingual transformer models, utilising five distinct datasets.
- Comparative evaluation of large language models (LLMs) with prompt-based strategies for Galician NER.
- Release of a new DeBERTa extra-small model for Galician, included in the mentioned evaluations.
- Incremental augmentation of training data with cross-lingual resources to assess its impact on the performance of Galician NER models.
- Release of a new state-of-the-art NER model for Galician.

Therefore, our research aims to contribute to the fields of computational linguistics and NLP by introducing new resources and models and advancing understanding of the lesser-studied language, Galician. Some of our findings provide further evidence supporting hypotheses previously proposed for other languages, while also offering new insights. For example, our initial experiments show that multilingual transformer architectures outperform their monolingual counterparts on the NER task in Galician. This result contrasts slightly with the findings reported by de Dios-Flores and Garcia (2022) for syntactic evaluation. However, after applying data augmentation with cross-lingual resources in our final experiments, a monolingual model achieved

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

the best performance. These and other findings from our study underscore areas warranting further investigation in future research across additional languages.

Our findings reveal several key insights: (i) the multilingual DeBERTa model fine-tuned with Galician data achieves better results compared to monolingual models, suggesting that its cross-lingual training enhances its ability to generalise across languages; (ii) however, for other architectures such as Bidirectional Encoder Representations from Transformers (BERT), monolingual models still outperform their multilingual counterparts, highlighting the advantages of language-specific models in certain cases; (iii) furthermore, the BERT-base-gl monolingual model (Garcia 2021) achieves the best performance benefiting from cross-lingual data augmentation; (iv) despite its smaller size, DeBERTa-xsmall demonstrates competitive performance, proving that more compact models can still achieve strong results; (v) the performance of the BERT-base-gl monolingual model declines when cross-lingual data in Spanish is added but improves with the addition of Portuguese data; in contrast, the mDeBERTa multilingual model experiences a performance decline when cross-lingual data from either the Spanish or Portuguese datasets is used for data augmentation; (vi) the performance of NER models across different datasets varies significantly, with factors such as dataset annotation guidelines influencing model effectiveness; (vii) medium-size LLMs (< 10 B parameters) with standard few-shot prompt-based strategies are not competitive with fine-tuning approaches; and (viii) our best models demonstrate statistically significant performance gained across multiple datasets, with consistent increases in F1-scores.

The subsequent sections of this paper are organised as follows. Section 2 provides a brief overview of existing resources and models for the NER task. Section 3 describes the creation of novel NER datasets for Galician and explains their relevance to our work. Section 4 presents the design of our experiments and their outcomes. Section 5 introduces the results of a qualitative analysis that assesses the outputs of the best Galician NER models for each entity type. Section 6 refers to the discussion that emerges from the main findings collected within Sections 4 and 5. Finally, the conclusions of our study are drawn in Section 7.

2. Related work

This section reviews prior work on NER, covering modeling approaches, datasets and domains, and language-specific studies, with emphasis on cross-lingual methods for low-resource settings with a focus on Galician.

2.1 NER methods and architectures

NER emerged in the 1990s, notably through evaluation campaigns such as the MUC (Message Understanding Conferences)[a] and later the Conference on Computational Natural Language Learning (CoNLL) shared tasks (Tjong Kim Sang 2002; Tjong Kim Sang and De Meulder 2003), which established the ENAMEX category (PER, person; LOC, geographical locations; ORG, organisations; and MISC, miscellaneous entities) alongside other entity types such as the NUMEX numerical expressions (Grishman and Sundheim 1996). These initiatives also released benchmark datasets that helped standardise the task and enabled systematic comparison of approaches.

Early NER systems relied on a combination of supervised learning methods and rule-based techniques, often incorporating handcrafted features and external linguistic resources such as gazetteers and lexicons (Keraghel *et al.* 2024). More recently, there have been significant advancements across various types of NER architectures and approaches, including knowledge-based systems (Wang *et al.* 2022), unsupervised strategies (Nadeau, Turney, and Matwin 2006), feature-engineered supervised systems (Agerri and Rigau 2016), and neural network models (Yan *et al.* 2019). While each approach offers its own advantages, they also have distinct limitations.

[a] https://www-nlpir.nist.gov/related_projects/muc/

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Knowledge-based NER systems rely on handcrafted rules, lexicons, and orthographic features so that do not require extensive annotated training data and typically leverage domain-specific resources to identify named entities. However, their dependency on manually crafted rules limits their adaptability to new domains. Unsupervised NER methods minimise the need for large labelled datasets by relying on patterns and statistical methods for data extraction. While effective at reducing annotation effort, these methods may be less accurate than supervised approaches. Feature-engineered supervised systems take a more holistic approach by employing machine learning algorithms with carefully crafted features. These systems integrate domain-specific knowledge, orthographic features, and other linguistic information. Before the rise of neural networks, feature-engineered systems were the popular method for the NER task.

As in other NLP tasks, NER has also benefited from advances in pre-trained word embeddings (Levy *et al.* 2015; Habibi *et al.* 2017). Therefore, modern neural architectures for NER classify words using representations based on different linguistic units, such as words, characters, sub-word units, or combinations of these Wang *et al.* (2021a). These models offer greater flexibility and adaptability compared to earlier approaches, making them the current standard in NER development.[b] Nevertheless, and despite its conceptual simplicity when compared to more elaborate architectures, fine-tuning transformer-based models remains a highly competitive approach, delivering strong performance that is often only marginally below the state of the art (Devlin *et al.* 2019; He, Gao, and Chen 2021a).

More recently, LLMs have been explored for NER due to their ability to operate in zero-shot or few-shot settings and to support new entity definitions without extensive annotation (Ashok and Lipton 2023; Keraghel *et al.* 2024). These properties make them attractive for low-resource scenarios, but they still underperform task-specific models in several cases (Wang *et al.* 2025). In addition, as generative rather than sequence-labelling models, LLMs may produce characteristic errors such as hallucinated entities, extraneous text, or incomplete coverage of the input (Ashok and Lipton 2023; Merdjanovska, Aynetdinov, and Akbik 2024).

2.2 Datasets and domains

While several unsupervised approaches exist, the major improvements in NER model development and evaluation have been driven primarily by accessible annotated corpora (Yadav and Bethard 2018). In this regard, recent advancements have led to the development of a wide variety of datasets covering multiple languages, domains, and entity types (Keraghel *et al.* 2024; Bade *et al.* 2024). Among the most prominent datasets are those created for the mentioned CoNLL 2002 and 2003, which released datasets derived from newswire articles in Spanish, Dutch, English, and German. NER datasets have also been created for social media platforms, particularly Twitter/X, where the data present unique challenges compared to well-formed text. These challenges include inconsistent spelling, informal and abbreviated language, creative punctuation, and incomplete or ungrammatical sentences. One of the earliest and most influential datasets in this domain is the Twitter NER corpus introduced by Ritter *et al.* (2011), which annotated ENAMEX entities in noisy microblog text. Subsequent efforts further expanded the scope of social media NER, notably the WNUT 2016 and WNUT 2017 shared task datasets on emerging and rare entities (Strauss *et al.* 2016; Derczynski *et al.* 2017), which include entity types that are less common in traditional newswire corpora. These datasets feature a broader range of entity types, including people, companies, facilities, bands, sports teams, movies, and TV shows, reflecting the dynamic, domain-specific nature of social media content. Together, these advancements underscore the growing diversity and specialisation of NER datasets designed to address the linguistic variability and domain shift inherent in social media platforms. The diversity of these datasets reflects

[b]Note, however, that performance varies significantly across datasets; state-of-the-art approaches for English exceed 94% F1 on CoNLL datasets but rarely reach 60% on social media data (Wang *et al.* 2021b).

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

the evolving nature of NER tasks and highlights the need for specialised approaches tailored to different textual contexts.

A particularly interesting case in the context of NER are BioNER tasks, which focus on the extraction of entities from biomedical texts, such as proteins, DNA, RNA, or cell attributes (Thiyagu *et al.* 2019). In this context, the 2010 i2b2 NER task (Uzuner *et al.* 2010) targets clinical data, focusing on entities related to clinical problems, tests, and treatments. Other notable examples include the Drug NER shared task (Segura-Bedmar, Martínez, and Herrero-Zazo 2013), which concentrates on entities like drugs, brands, chemicals, and related terms. Crucially, BioNER typically distinguishes between flat NER, which considers single, non-overlapping entity spans, and nested NER, which allows entities to overlap or be embedded within others (Park *et al.* 2024). While flat entities are relevant in biomedical scenarios, a substantial portion of the BioNER literature focuses on nested structures, which require specialised annotation schemes and modelling approaches. In this paper, however, we follow a flat annotation scheme, consistent with the standard paradigm for ENAMEX entity recognition.

2.3 Languages and cross-lingual NER

Most resources and studies for NER have been developed for English, covering domains such as medical texts, news, social media, and Wikipedia (Keraghel *et al.* 2024). For less-resourced languages, early approaches leveraged Wikipedia data to perform automatically annotation of multilingual data (Kim, Toutanova, and Yu 2012). Recent studies have shown the benefits of implementing cross-lingual knowledge transfer methods (Feng *et al.* 2018), while others simply rely on the fine-tuning of multilingual transformer models, obtaining competitive results even in zero-shot scenarios (Pires, Schlinger, and Garrette 2019). A widely adopted strategy for several downstream tasks involves exploiting annotated data from closely related languages, either directly via cross-lingual transfer or through adaptation methods that account for structural and lexical differences between the source and target languages (Garcia *et al.* 2018; Torge *et al.* 2023).

In the case of Galician, Portuguese and Spanish are the most commonly employed source languages and have been shown to yield the best results in cross-lingual tasks (Sarymsakova, Sánchez-Rodríguez, and Garcia 2024). Portuguese is particularly effective due to its strong linguistic proximity to Galician, while Spanish also proves beneficial given its long-standing coexistence with Galician within the same territory, which has resulted in significant linguistic influence (see Section 3.1 for details about Galician language).

Regarding NER studies for Portuguese, the earliest systematic evaluations were conducted within the HAREM shared tasks (Santos *et al.* 2006), which introduced a substantially finer-grained entity classification than the traditional ENAMEX scheme, making direct adaptation to standard NER settings not always straightforward. Subsequent evaluation campaigns, such as IberLEF 2019 (Collovini *et al.* 2019), as well as independent studies, have largely followed approaches comparable to those adopted for other languages. A comprehensive overview is provided by Albuquerque *et al.* (2023), who survey manually and automatically annotated Portuguese NER corpora spanning multiple domains and venues and report competitive results across a range of models, including BiLSTM+Conditional Random Field (CRF) and transformer-based architectures, in line with contemporary NER benchmarks for high-resource languages. Reported performance varies considerably across datasets and domains, with F1-scores ranging approximately from 67 to 97 for BiLSTM+CRF models and from 75 to 95 for BERT+CRF approaches, reflecting differences in annotation schemes, corpus size, and domain specificity.

Regarding NER research for Spanish, the CoNLL-2002 shared task dataset remains one of the most widely adopted benchmarks, consisting of manually annotated newswire texts labelled using the BIO scheme with the ENAMEX categories (Carreras *et al.* 2002). In addition, the AnCora corpus provides a large-scale manually annotated resource that, although originally designed as a linguistic treebank, has been successfully exploited for NER (Taulé *et al.* 2008).

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Recent Spanish NER systems predominantly rely on fine-tuning pre-trained transformer models on these datasets, achieving strong performance comparable to other high-resource languages (Gutiérrez-Fandiño *et al.* 2022). Reported results vary depending on the dataset and evaluation setup, with fine-tuned Spanish-specific and multilingual transformer models, such as BETO (Cañete *et al.* 2020) and XLM-RoBERTa (Conneau *et al.* 2019), typically reaching F1-scores in the 88–90 range on CoNLL-style benchmarks, confirming the effectiveness of transformer-based approaches for Spanish NER.

As for Galician, the object of our study, several studies have focused on developing NER resources. Garcia *et al.* (2012) presented a rule-based (knowledge-based) NER model, achieving a precision of 69.9, a recall of 62.05, and an F1-score of 60.42. Building on these results, Agerri *et al.* (2018) proposed a semi-supervised approach for NER, improving performance to 83.85 for precision, 84.54 for recall, and 84.19 for F1. More recently, Garcia (2024) explored the effectiveness of syntax-based vector models across various NLP tasks, including NER.

The aforementioned study evaluates several training techniques for NER models in Galician, comparing architectures like BiLSTM-CNN-CRF initialised with static embeddings, and fine-tuned BERT models such as mBERT (Devlin *et al.* 2019) and XLM-RoBERTa. Among these, fine-tuned BERT models demonstrated the best performance. Specifically, the multilingual XLM-RoBERTa-large achieved an F1-score of 86.93, outperforming multilingual mBERT and other approaches. However, monolingual BERT models yielded even better results for Galician, achieving a precision of 87.69, a recall of 89.7, and an F1-score of 88.68. This NER system was trained and evaluated using these methods and a comprehensive, manually reviewed corpus, SLI NERC (Agerri *et al.* 2018), which is described in detail in Section 3. These and other previously mentioned findings emphasise (i) the superior performance of NER models trained on large, meticulously curated corpora and (ii) the superior suitability of monolingual BERT models for Galician NER tasks compared to recurrent networks, with *fastText* obtaining the highest results among static embeddings. Nonetheless, apart from the SLI NERC dataset, no other datasets have been used to evaluate these recently proposed NER models for Galician. Therefore, a comprehensive assessment is essential to understand how these models perform across different Galician corpora.

In summary, this section has presented an overview of state-of-the-art methods and resources for NER, encompassing both monolingual and cross-lingual settings and accounting for different annotation schemes, including flat, nested, continuous, and discontinuous entities. English serves as a reference point, with consistently strong results, while transformer-based architectures also constitute the dominant and most effective approach for Portuguese and Spanish. On CoNLL-like datasets with flat annotations, state-of-the-art performance is typically high, often exceeding 94% F1, but degrades notably across datasets, domains, and more complex annotation schemes. In contrast, results for Galician remain inconclusive, highlighting the need for further evaluation across multiple corpora.

In light of these findings, our contributions, as formulated in Section 1, aim to (i) expand the resources available for training and evaluating NER systems in Galician using flat annotation schemes, (ii) evaluate the use of various pre-trained BERT-based word embeddings for training Galician NER systems on five different datasets, and (iii) analyze and confirm the impact of incorporating cross-lingual data into the NER system training process.

3. Datasets

To improve Galician NER, we annotated two new corpora and adapted existing datasets, significantly expanding labelled resources for the language.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

3.1 Galician

Modern Galician (ISO code `gl`) originated from mediaeval Galician-Portuguese and retains many features in common with Portuguese, being currently spoken by approximately 2 million people, primarily in the Spanish Autonomous Region of Galicia (Monteagudo *et al.* 2024). Although its dialects have historically been regarded as part of the same language as Portuguese (Lindley Cintra and Cunha 1984; Mira Mateus *et al.* 2003), a combination of sociolinguistic, political, and identity factors, along with the substantial linguistic influence of Spanish, is used to support its recognition as a separate language (Fernández Rei 1991).

Officially standardised since 1983, Galician now coexists with Spanish as an official language and employs a Spanish-based orthography (Samartim 2012). Therefore, in computational contexts, differences in spelling have posed challenges for resource sharing with Portuguese, yet several studies have successfully adapted Portuguese resources to develop NLP tools for Galician (Garcia *et al.* 2018; Sarymsakova *et al.* 2024). Although Galician remains a low-resource language (Mateo and Rodríguez 2012; Sánchez *et al.* 2024), initiatives such as the Proxecto Nós (de Dios-Flores *et al.* 2022) have contributed valuable datasets and models, further enriching the language's representation in NLP and complementing its prior inclusion in widely used multilingual models, such as mBERT or XLM-RoBERTa.

3.2 Galician NER resources

We have used the SLI NERC corpus as our primary training and evaluation resource (Agerri *et al.* 2018). We rely on splits of 161,853 tokens for training and 40,482 tokens for testing, released by Garcia (2024).[c] This corpus was encoded in CoNLL format, specifically created, annotated, and manually reviewed using Galician texts for NER. This dataset originates from the CTG Galician Technical Corpus,[d] which encompasses a wide range of domains, including GALEX (Galician legal texts), XIGA (Galician texts on computing and telecommunications), AUGA (Galician texts on ecology and environmental sciences), ACHEGA (Galician economy texts), SOGAL (Galician texts on sociology and social sciences), and MEDIGAL (Galician texts on medicine). It is important to note that the currently available Galician NER model released by Garcia (2024) was trained and evaluated on this dataset. We aim to enhance this model's performance by leveraging the same baseline resources.

Additionally, other annotated datasets, such as LREC (Garcia and Gamallo 2014) and CorNER (Garcia *et al.* 2012), were adapted and employed as evaluation resources in our study. The LREC dataset comprises journalistic and encyclopaedic (Wikipedia) texts in Galician, with approximately 35% journalistic and 65% Wikipedia content, totalling around 47,000 tokens. This dataset was primarily designed for person entity recognition tasks in journalistic and encyclopaedic documents. After processing the data, two linguists, both native Galician speakers, manually reviewed the NER annotation and added coreferential information to the corpus, ensuring linguistic accuracy and consistency.

The CorNER dataset (Garcia *et al.* 2012), consisting of around 20,000 tokens, is focused on named entities in Galician, including ENUMEX types such as dates, numerals, quantities, and proper nouns (ENAMEX). It was partially created through a random selection from the Gari-Coter project (Barcala *et al.* 2007), a textual corpus focused on economic press and supplements in daily newspapers. This corpus comprises 601 news articles totalling 235,175 tokens. Among others, the sources include newspapers and magazines such as 'O Correo Galego', 'Galicia Confidencial', and 'Tempos Novos'. Beyond the Gari-Coter project, the dataset also incorporates encyclopaedic texts (including Wikipedia articles) and articles from the 'Galicia Hoxe' newspaper, spanning various sections. The corpus was preprocessed using FreeLing (Padró and Stanilovsky

[c] https://github.com/marcospln/evaluation_splits_gl
[d] http://sli.uvigo.gal/CTG/

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

2012), and a subset of 20,000 tokens (containing nearly 1,000 proper names) was manually annotated to evaluate the system's ability to identify ENAMEX entities.

All of the resources described, covering a wide range of domains, have enabled researchers to tackle various NLP tasks in Galician. Many of these tasks have already been successfully implemented and reported in the previously cited works, yielding promising results and positive performance metrics. However, only a few of these corpora have been used for NER evaluation in Galician to date. Specifically, the SLI NERC corpus has been used as an evaluation resource in Agerri *et al.* (2018) and Garcia (2024). The other resources, such as LREC, CorNER, and others, have not been used in current NER models for Galician. As such, our goal is to adapt and include these resources for training and evaluating NER models.

3.3 Galician PUD and TreeGal NER datasets

In the following subsections, we detail the procedure for processing, annotating, and the results obtained with two newly developed NER corpora in Galician. Both corpora are released in the Universal Dependencies (UD) format (de Marneffe *et al.* 2021). These datasets aim to contribute to the Universal Named Entity Recognition (UNER) framework (Alves *et al.* 2020), addressing the multilingual challenges in NLP by providing high-quality NER datasets across multiple languages, all using unified guidelines.

3.3.1 Preprocessing

The first corpus, Galician-TreeGal v0.42 (Garcia *et al.* 2018), contains 1,000 sentences, totalling approximately 25,000 tokens. The TreeGal corpus was created by semi-automatically converting part-of-speech tags from the original resource[e] and assigning dependency labels using cross-lingual parsing techniques Garcia *et al.* (2018). TreeGal is derived from the XIADA Galician corpus (Rojo *et al.* 2015), annotated with lemmas and rich morphological information and thoroughly reviewed by experts. The full version of the XIADA corpus consists of 741,833 tokens. It is divided into four subsets, each representing a different typology: (i) generic press, (ii) economic press, (iii) short stories, and (iv) free, unrelated sentences.

The second corpus, Galician PUD, was recently developed and released by Sánchez-Rodríguez *et al.* (2023). It consists of 1,000 sentences and approximately 23,000 tokens. The source material, derived from the English PUD corpus (Zeman *et al.* 2017), comprises English sentences from the news domain and Wikipedia. Professional native Galician translators carried out the translations, following detailed guidelines to ensure consistency. The Galician PUD corpus was manually annotated, with every sentence reviewed by professional translators and aligned with 23 other PUD datasets in the UD framework.

Initial linguistic annotations for both corpora were generated using state-of-the-art NLP tools for Galician and subsequently refined by two experts, achieving a high level of inter-annotator agreement (IIA), as detailed in Section 3.3.4.

3.3.2 Annotation guidelines

In this study, we initially follow the UNER guidelines using the standard ENAMEX entities,[f] based on and its (proto)typical entities classes (PER: people; LOC: geographical locations; ORG: organisations; MISC: other entities). We leveraged a BERT-based model and tokeniser pre-trained for Galician Named Entity Recognition (NER) by Garcia (2024), specifically fine-tuned for token classification in Galician, to annotate TreeGal and Galician PUD tokens within the CoNLL-U framework, using the aforementioned tagset.

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

[f] As UNER does not use the MISC class, these annotations can be removed for full compatibility. Nonetheless, we include the MISC annotations to ensure alignment with the other Galician datasets used in this paper for model evaluation.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 1. Results of inter-annotator agreement (IAA) for the Galician PUD and TreeGal NER labelling

Modality	Dataset	IAA
A	TreeGal (400)	0.959
	PUD (200)	0.969
B	PUD (800)	0.958

Our NER annotation involves marking all names of people (both real and fictional), organisations, locations, products, and events mentioned in the texts. The primary principles for identifying and annotating named entities have already been established within the UNER framework and other related works cited in Section 2. However, we encountered some challenging annotation cases related to ambiguity in our datasets and, as a result, adapted the following guidelines specifically for the Galician corpus:

1. Contextual abbreviations, such as "o catalán inventor de ZP **MISC** " referring to " Zapatero Presidente **PER** " are annotated as `MISC`, as they represent a phenomenon rather than a person.
2. Coordinated elements like " PSA **ORG** e a Zona Franca **ORG** acabarán entendéndose" should be annotated with the same label, in this case, `ORG`, although *Zona Franca* could also be interpreted as a location.
3. Metonymic cases, such as "toda a vida sen tocar Bach **MISC** ou Scarlatti **MISC** " should be annotated as `MISC` entities; or other metonimic cases of `LOC` and `ORG` entities, such as "O apetito insaciable de China **ORG** "and a "cidade máis grande ao oeste de. China **LOC** "
4. Non-lexicalised foreign words (according to Real Galician Academy (*Real Academia Galega*)), such as "best-séller" or "ikebana" have not been annotated.

We followed a systematic process to ensure accurate token labelling and format compatibility, as detailed in the following Subsection 3.3.3.

3.3.3 Annotation of named entities

The TreeGal and Galician PUD corpora were annotated by two experts: a native speaker with a strong linguistic background and a postdoctoral researcher in Linguistics with advanced proficiency in Galician. To ensure consistency and alignment with the annotation guidelines, both annotators initially worked on a test dataset (400 sentences for TreeGal, then 200 for Galician PUD) from the original split of each corpus. To ensure quality in the annotation process, the initial annotated sets underwent a manual review by both annotators, who cross-examined each other's results through exchanges and comparisons of annotations. The results of the inter-annotator agreement for this review task (Modality A) are presented in Table 1.

For the further annotation process, the training datasets (along with the Galician PUD development set) were divided and assigned to each expert for semi-automatic annotation. After predicting the NER tags for each token using the aforementioned BERT-based pre-trained model, the outputs for both TreeGal and Galician PUD were reviewed by the annotators (Modality B).

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

During the annotation phase, regular follow-up meetings with additional language experts were held to resolve any ambiguities or questions that emerged during the review and curation phases. It is worth noting that, for the machine-assisted annotation review, the training dataset from Galician PUD was chosen and evaluated by both annotators, since the inter-annotator agreement for this corpus in the Modality A review task is higher than that of TreeGal.

This enabled us to enhance the quality of the manually annotated NER corpus for Galician at the final stage. By implementing this methodology, we mitigated the potential for mutual influence or bias among annotators during the review process conducted under Modality A. Consequently, we were able to evaluate whether the agreement for the machine-generated annotations remained consistent. As demonstrated by Modality B in Table 1, the results confirm that the annotation process was devoid of bias among annotators.

3.3.4 Inter-annonator agreement

After the NER annotation review process was completed, we calculated the annotators' agreement using all tagged entities (including O for "out") from the TreeGal and Galician PUD datasets. To do so, we used a framework proposed by Artstein and Poesio (2008) to calculate weighted Cohen's κ (Cohen 1960) for both A and B Modalities Table 1.

Within the Modality A, which involves the review of manually annotated tokens, the inter-annotator agreement values ranged from 0.96 (Cohen's κ for TreeGal, as it was annotated first, as noted in the previous Subsection 3.3.3) to 0.97 for the subsequently annotated Galician PUD. These scores are relatively high, particularly given that this represented the initial phase of annotation for both annotators. As for IIA for Modality B, consisting of the review of machine-generated annotation, these values slightly descend to $\kappa = 0.96$. Nevertheless, we consider this a positive dynamic in the annotation process, as the Modality B corpus encompassed a greater number of reviewed sentences than the Modality A corpus.

Overall, we observe a very high level of agreement, highlighting (i) the consistency between the two annotators in their annotations and (ii) the effectiveness of both types of annotation review (Modality A and B) and follow-up meetings throughout the annotation and curation phases. The enhancement in annotation quality is clear, with total agreement values exceeding 0.95 within both modalities. As a result, discussions of the discrepancies and a review of the initial annotations during the curation phase enabled the annotators to identify inconsistencies in the labelling of specific entities, leading to a comprehensive revision of the final annotations for the TreeGal and Galician PUD datasets.

3.4 Results

Building on the previous works cited in Section 3 and leveraging our two newly annotated Galician NER corpora, we compiled five datasets for the NER task. The named entity statistics for each dataset are provided in Tables 2 and 3. The per-token distribution of annotations ranges from 1.82% to 7.02%, meaning that across datasets, between 1.82% and 7.02% of all tokens are annotated as named entities. The TreeGal dataset has the fewest annotated entities (488, or 1.82% of the corpus). It is worth mentioning that the distribution of the four named entity classes in TreeGal and Galician PUD is roughly similar, indicating a consistent annotation approach across both datasets, which are nearly identical in size.

The combined total of annotated entities across the five datasets is 11,039 (see Table 16 in the Appendix). The majority of these are classified as *organisation* entities, accounting for 31.25% (#3,450), followed by *location* entities at 28.05% (#3,096). *Person* entities make up 22.29% (#2,461), and *miscellaneous* entities constitute 18.41% (#2,032). The distribution of named entities across the datasets is relatively balanced, with a slight predominance of organisation entities. The proportional frequency of these entities is beneficial for the fine-tuning and improvement process of the existing Galician NER model, as it provides a solid representation of one of the most frequent entity types.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 2. Distribution of entity classes in Galician PUD, TreeGal, CorNER, and LREC datasets

	Galician PUD			TreeGal			CorNER			LREC		
Tag	#	%	len	#	%	len	#	%	len	#	%	len
PER	420	33.71	1.44	141	28.89	1.88	121	13.10	1.83	1184	35.68	1.88
LOC	399	32.02	1.44	110	22.54	1.15	280	30.30	1.14	543	16.37	1.42
ORG	276	22.15	1.89	150	30.74	1.57	438	47.40	1.66	805	24.26	1.94
MISC	151	12.12	2.28	87	17.83	2.08	85	9.20	2.08	786	23.69	2.74
Overall	1,246	**5.30**	–	488	**1.82**	–	924	**4.41**	–	3,318	**7.02**	–
Total tokens		23,509			26,817			20,932			47,254	

Percentages for each entity are calculated over the total number of tags (see "Overall"), and overall percentages for annotated entities per dataset are computed over a total number os tokens. The average token length per entity type is shown in the "len" column.

Table 3. Distribution of entity classes in SLI NERC dataset

	Train			Dev			Test		
Tag	#	%	len	#	%	len	#	%	len
PER	303	0.21	1.90	79	0.49	1.92	213	0.53	2.07
LOC	1247	0.86	1.44	78	0.48	1.34	439	1.09	1.49
ORG	1095	0.75	1.92	130	0.80	1.77	556	1.38	2.15
MISC	538	0.37	2.36	25	0.15	2.36	360	0.89	2.53
Overall	3,183	2.19	–	312	1.92	–	1,568	3.88	–
Total tokens		145,582			16,254			40,405	

4. Experiments and results

This section describes the experiments conducted to train and evaluate several transformer-based NER models in Galician. We first introduce the data and methods, followed by the assessment of the pre-trained models, including a new DeBERTa_v3-xsmall for Galician. We then show the evaluation results and further experiments using data augmentation techniques as well as a comparison to prompting strategies using LLMs for NER.

4.1 Data

To enhance the performance of the existing Galician NER model (Garcia 2024), we conducted a series of experiments using datasets comprising the SLI NERC, TreeGal, Galician PUD, CorNER, and LREC corpora. We used the SLI NERC splits provided by Garcia (2024), which divided the SLI NERC test dataset into 'dev' and 'test' splits. Therefore, SLI NERC was divided into training (70%, 145,582 tokens), development (10%, 16,254 tokens), and test (20%, 40,405 tokens) sets, as shown in Table 3.

In line with our contributions outlined in Section 1—particularly the evaluation of a newly trained model with additional data integration—we incorporated the TreeGal training split into the training dataset. The TreeGal training split comprises 60% of TreeGal, amounting to 16,705 tokens. As Table 4 shows, we obtained the final splits to perform the evaluation. Consequently, the final training set combined the SLI NERC and TreeGal datasets, resulting in a total of 162,287

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 4. Distribution of entity classes in final training, development, and test dataset

Set	Corpus	Tokens
Train	SLI NERC (train)	145,582
	TreeGal (train)	16,705
Dev	SLI NERC (dev)	16,254
Test	Galician PUD	23,509
	TreeGal (test)	10,112
	CorNER	20,932
	LREC	47,254
	SLI NERC (test)	40,405

tokens. The development set, exclusively from the SLI NERC dataset, remained at 16,254 tokens. The test splits consist of the following datasets: the Galician PUD corpus (23,509 tokens), the TreeGal test set (10,112 tokens), the CorNER corpus (20,932 tokens), the LREC corpus (47,254 tokens), and the SLI NERC corpus (40,405 tokens).

4.2 Architecture selection

To train the NER models, we used the standard fine-tuning approach of pre-trained transformer models. To do so, we took advantage of the *transformers* library[g] provided by HuggingFace (Wolf *et al.* 2020). For each model, we explored the different batch sizes {8, 16, 32} and learning rates $\{1e^{-5}, 2e^{-5}, 3e^{-5}, 4e^{-5}, 5e^{-5}\}$ and selected the best checkpoint by its F1-score in the development set after five epochs. To assess the performance across datasets, and taking into account their different sizes, we compute both micro and macro average F1-scores.

We compare the performance of the following monolingual and multilingual pre-trained models (see Table 17 in Appendix A.2 for the details of each model):

Monolingual. We compare both "small" and "base" variants of Bertinho (Vilares, Garcia, and Gómez-Rodríguez 2021) and BERT-gl Garcia (2021) models. Additionally, we evaluate the new DeBERTa-xsmall model, described in Section 4.3.

Multilingual. Regarding multilingual models, we tested the official multilingual BERT base (mBERT, Devlin *et al.* (2019)), the 'base' and 'large' versions of XLM-RoBERTa (Conneau *et al.* 2019), and the multilingual DeBERTa_v3 model (mDeBERTa_v3-base, He *et al.* (2021a)).

Baseline. As a baseline, we use the best model provided by Garcia (2024), which consists of a BERT-base-gl fine-tuned with the CTG-NER corpus and available at the HuggingFace Hub.[h] Therefore, this is a strong baseline with state-of-the-art performance on NER for Galician, which only differs in having used slightly less training data than the models we are fine-tuning in this study.

4.3 A new Galician DeBERTav3 xsmall

Available monolingual encoder models for Galician (Vilares *et al.* 2021; Garcia 2021) take advantage of the BERT architecture (without the next sentence prediction training objective) and appear

[g] https://github.com/huggingface/transformers
[h] https://huggingface.co/marcosgg/bert-base-gl-SLI-NER.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 5. Accuracy of the baseline model trained with custom splits (70/10/20%) of SLI NERC

Model	Test dataset	F1	P	R	Micro F1	Macro F1
Baseline	SLI NERC test	88.68	87.69	89.69	**84.70**	**85.24**
	Treegal test	84.36	85.06	83.67		
	PUD	79.84	79.43	80.26		
	CorNER	**90.13**	90.87	89.39		
	LREC	83.20	81.88	84.57		

at 'small' and 'base' sizes, having 6 and 12 hidden layers, about 44 and 86 million backbone parameters, and occupying approximately 300 and 600 megabytes, respectively. The standard BERT architecture was improved by DeBERTa models (He *et al.* 2021b), which use disentangled attention (differentiating content and position embeddings) instead of self-attention and, in its version 3, substitutes masked language modelling with replaced token detection (He *et al.* 2021a). This sample-efficient pre-training task allows smaller models to obtain better results than larger models from previous architectures.

To verify how well this architecture performs on NER for Galician and to compare it with larger available models, we trained a new extra-small (xsmall) DeBERTa_v3 model for Galician. To do so, we did a continual pre-training of the original English DeBERTA-xsmall[i] on CorpusNÓS, an open Galician corpus composed of 2 billion words (de Dios-Flores *et al.* 2024). We trained the model using the original tokeniser on a server with two 80 GB NVIDIA Hopper H100 GPUs for one epoch over the corpus. The final model, here dubbed *deberta-gl-xsmall*, has 22 M backbone parameters, occupies 142 megabytes, and is freely available on HuggingFace.[j]

4.4 Results

We present the results on the five evaluation datasets, starting with the baseline and then the monolingual and multilingual models.

Baseline. As shown in Table 5, the baseline model achieves a micro average accuracy of 84.70% and macro average of 85.24%, with its highest prediction score (90.13%) observed on the CorNER test dataset, and the lowest in the Galician PUD corpus (79.84%). The F1-scores reported in Table 5 are averaged across datasets. This difference of more than 10 points across datasets reinforces the need for improved cross-domain NER systems for Galician.

4.4.1 Monolingual models

We started by searching for the best hyperparameters of the BERT-base-gl model, whose results are plotted in Figure 1. A comparison of both results (Table 5 and Figure 1) reveals an improved performance (84.70% vs. 85.07% in micro average F1). The red bar in Figure 1 highlights the best performance, obtained with a learning rate of 4e-05 and a batch size of 16. This preliminary finding, combined with the studies discussed in Section 2, supports the further experiments comparing additional models and data to enhance accuracy for the NER task in Galician.

A summary of the results for the best checkpoint of each monolingual model is shown in Table 6. These results show that BERT-base-gl outperformed all other models at both micro and macro F1 levels. This model, trained with 5 epochs, a batch size of 16, and a learning rate of $4e^{-5}$, achieved the best results among the monolingual models and delivered the highest F1-scores

[i]https://huggingface.co/microsoft/deberta-v3-xsmall
[j]https://huggingface.co/proxectonos/deberta-xsmall-gl

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

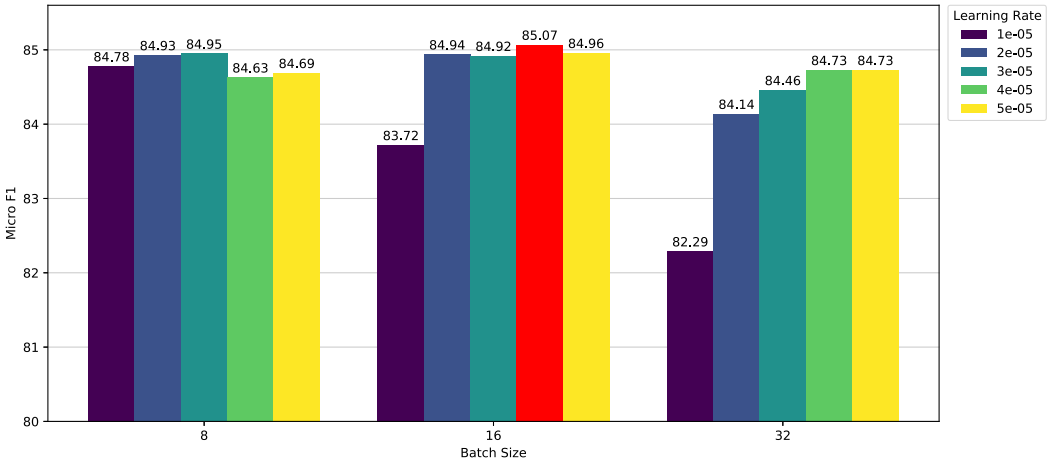


Figure 1. Results for NER Galician model trained with BERT-base-gl embeddings and diverse hyperparameters.

across all five test datasets. Among these, the best accuracy was observed in the SLI and CorNER evaluation datasets, and the lowest metrics were observed in the PUD and LREC corpora.

The performance of the other available models follows relatively similar trends as those reported in other studies, here with better results for both Bertinho models than for BERT-small-gl. Regarding the new DeBERTa-xsmall-gl, it performs better than the 'small' models, despite having half their backbone parameters (22M vs. 44M) and size.

4.4.2 Multilingual models

For the multilingual models, the results in Table 7 indicate that the mDeBERTa_v3-base NER model outperforms the monolingual BERT-base-gl model (85.47 vs. 85.07 μ F1). Trained with 5 epochs, a batch size of 16, and a learning rate of $4e^{-5}$, the model achieved the highest F1-scores (micro F1 of 85.47 and macro F1 of 86.34) across all five test datasets, outperforming other multilingual models: 87.78% for the SLI test set, 85.28% for TreeGal, 85.01% for PUD, 90.40% for CorNER, and 83.22% for LREC. Among these, the highest accuracy was observed in the SLI and CorNER evaluation datasets. At the same time, the lowest scores were observed for the PUD and LREC corpora, consistent with the results from the monolingual BERT-base-gl model.

Among the other multilingual models, XLM-RoBERTa-large achieved competitive performance, matching the best monolingual model. At the same time, both mBERT and XLM-RoBERTa-base performed similarly to Bertinho-base, as also noted by previous research (Vilares *et al.* 2021).

In conclusion, regarding transformer architectures, our initial results on monolingual corpora differ slightly from trends reported in other evaluations across different tasks, for example, de Dios-Flores and Garcia (2022). While monolingual models trained on smaller, language-specific datasets perform well in our study, multilingual models trained on larger datasets seem to outperform them providing better accuracy metrics. These results suggest that multilingual models may be more effective for training NER models in contrast to other downstream tasks and intrinsic evaluations, as NER is less dependent on linguistic structure than, for example, PoS-tagging or targeted syntactic evaluation. Crucially, for both best monolingual and multilingual models, and despite using the same development set, the variation across datasets was lower than that of the baseline model, making them more suitable for generic NER labelling in Galician.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 6. The results on NER task trained with BERT-based monolingual models and evaluated on five different datasets

Model	Test dataset	F1	P	R	MicroF1	MacroF1	BS	LR
BERT-base-gl	SLI NERC test	88.08	86.46	89.76	**85.07**	**85.88**	16	4e-05
	Treegal test	85.98	85.63	86.33				
	PUD	82.66	81.94	83.39				
	CorNER	89.45	89.93	88.96				
	LREC	83.22	81.39	85.14				
BERT small gl	SLI NERC test	86.99	85.18	88.87	82.33	82.86	8	3e-05
	Treegal test	82.70	81.55	83.88				
	PUD	76.19	75.12	77.29				
	CorNER	87.45	87.45	87.45				
	LREC	80.98	79.22	82.82				
DeBERTa xsmall gl	SLI NERC test	87.41	86.11	88.74	83.01	83.49	16	3e-05
	Treegal test	82.86	82.86	82.86				
	PUD	78.75	78.84	78.65				
	CorNER	87.03	86.93	87.12				
	LREC	81.42	79.82	83.09				
Bertinho base	SLI NERC test	86.97	85.86	88.10	83.33	83.58	8	5e-05
	Treegal test	81.87	81.71	82.04				
	PUD	80.67	80.13	81.22				
	CorNER	86.44	86.30	86.58				
	LREC	81.96	80.32	83.66				
Bertinho small	SLI NERC test	85.44	83.97	86.96	82.45	83.08	8	5e-05
	Treegal test	82.63	81.80	83.47				
	PUD	79.63	78.94	80.34				
	CorNER	86.81	86.72	86.90				
	LREC	80.88	79.11	82.73				

BS—batch size; LR—learning rate. All the provided models were trained using a configuration of 5 epochs.

4.5 Augmentation of training data

To achieve our contributions as formulated in Section 1, we aim to incorporate supplementary data from closely related linguistic varieties to evaluate their effect on the performance of Galician Named Entity Recognition models. For this purpose, we have incorporated the Spanish AnCora dataset (Taulé *et al.* 2008) converted to IOB format from its UD version,[k] and the Portuguese Bosque dataset[l] as released by Pires (2017).[m]

Using data from languages related to Galician allows us to test the model's sensitivity to subtle linguistic contrasts, its capacity to generalise across related varieties, and its robustness in handling language continua rather than clearly separated systems.

[k] https://github.com/UniversalDependencies/UD_Spanish-AnCora
[l] https://www.linguateca.pt/floresta/corpus.html\#bosque
[m] https://github.com/arop/ner-re-pt/tree/master/datasets/harem/nltk

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 7. The results on NER task trained with BERT-based multilingual models and evaluated on five different datasets

Model	Test dataset	F1	P	R	MicroF1	MacroF1	BS	LR
mDeBERTa_v3-base	SLI NERC test	87.78	86.48	89.12	**85.47**	**86.34**	8	2e-05
	Treegal test	85.28	84.85	85.71				
	PUD	85.01	84.24	85.79				
	CorNER	90.40	90.11	90.69				
	LREC	83.22	81.67	84.84				
mBERT	SLI NERC test	86.68	85.42	87.98	83.53	84.09	8	3e-05
	Treegal test	83.13	82.79	83.47				
	PUD	80.49	79.39	81.62				
	CorNER	88.22	88.12	88.31				
	LREC	81.96	80.59	83.36				
XLM-RoBERTa-base	SLI NERC test	86.71	85.29	88.17	83.05	83.76	16	4e-05
	Treegal test	84.17	84.25	84.08				
	PUD	82.30	81.77	82.83				
	CorNER	84.58	84.86	84.31				
	LREC	81.03	79.40	82.73				
XLM-RoBERTa large	SLI NERC test	88.53	87.46	89.63	85.28	85.54	8	2e-05
	Treegal test	81.99	81.74	82.24				
	PUD	83.78	82.99	84.59				
	CorNER	89.86	90.01	89.72				
	LREC	83.53	82.20	84.90				

BS—batch size; LR—learning rate. All the provided models were trained using a configuration of 5 epochs.

Table 8. Distribution of entity classes in Spanish AnCora, Portuguese Bosque, and PUD datasets

AnCora				Portuguese NER			
Tokens	Tag	#	%	Tokens	Tag	#	%
589,548	PER	8,450	1.4	240,215	PER	4,230	2
	LOC	5,408	1		LOC	2,768	1
	ORG	11,522	2		ORG	3,396	1.5
	MISC	9,046	1.5		MISC	1,438	0.5
Total		34,426	6			11,832	5

These datasets are comparable in size and annotation standards to SLI NERC, which constitute most of our training data. Specifically, the Portuguese NER dataset contains 11,832 annotated entities, while the Spanish Ancora dataset contains 34,427 NER-labelled tokens. However, the Spanish AnCora corpus is significantly larger, with 589,548 tokens. Table 8 provides an overview of the named entity distribution across these datasets.

To explore the impact of the data augmentation, we have selected the best monolingual and multilingual models evaluated in the previous section. In each case, we performed three

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 9. Results for cross-lingual training data augmentation using BERT-base-gl embedding model, compiling Spanish and Portuguese NER corpus

Train dataset	Test dataset	F1	P	R	MicroF1	MacroF1	BS	LR
AnCora	SLI NERC test	87.48	86.50	88.49	**84.65**	**85.18**	8	2e-05
	Treegal test	85.54	85.37	85.71				
	PUD	85.11	84.84	85.39				
	CorNER	83.04	81.51	84.63				
	LREC	84.65	83.63	85.70				
Portuguese NER	SLI NERC test	88.90	87.88	89.95	**86.30**	**86.72**	8	3e-05
	Treegal test	84.81	84.73	84.90				
	PUD	82.07	81.87	82.26				
	CorNER	92.71	92.56	92.86				
	LREC	85.09	84.02	86.20				
Mixed	SLI NERC test	88.81	87.76	89.89	**86.30**	**87.03**	32	4e-05
	Treegal test	86.38	86.03	86.73				
	PUD	83.82	83.45	84.19				
	CorNER	91.51	91.46	91.56				
	LREC	84.60	83.29	85.96				

BS—batch size; LR—learning rate. All the provided models were trained using a configuration of 5 epochs.

experiments: (1) fine-tuning of the model with the Galician and Spanish data; (2) fine-tuning using Galician and Portuguese data; and (3) fine-tuning with the training data of the three varieties, Galician, Spanish, and Portuguese (Mixed).

The results of cross-lingual data integration are presented in Tables 9 and 10, for the monolingual and multilingual models, respectively. In the first cross-lingual training split, including Spanish data, the performance of the monolingual BERT-base-gl model decreases (84.65 μ F1, and 85.18 macro F1 vs. the previous 85.07/85.88). For the multilingual model, results also decline (from 85.47 to 83.91 for micro and from 86.34 to 84.27 for macro), indicating that including Spanish data negatively impacts the model's performance.

In the case of integrating a more closely related variety (Portuguese), the performance of both models, especially the monolingual one, improved. For BERT-base-gl, it achieved a micro F1-score of 86.30 and a macro F1-score of 86.72, while the mDeBERTa_v3-base model demonstrated improved performance in the micro F1 scenario (85.96) and a small drop of 0.02 in the macro F1-score. The multilingual model improves only when trained on the Portuguese NER corpus, whereas training on the cross-lingual dataset with Spanish AnCora slightly decreases performance. A deeper investigation into this case is left for future work.

Finally, in the Mixed scenario (with a total corpus of 991,436 tokens), Table 9 exhibits that the macro F1-score for BERT-base-gl increased to 87.03, while keeping the same values as for the previous experiment (86.30) In contrast, the performance of the mDeBERTa_v3-base model slightly declined, obtaining a micro F1-score of 85.35 and a macro F1-score of 85.70 (Table 10). Notably, across the evaluation datasets, both monolingual and multilingual models trained on the Mixed dataset achieved the highest accuracy on the SLI and CorNER datasets. At the same time, the lowest scores were recorded for the PUD and LREC corpora. This observation was previously noted in Section 4.4.

These final experiments show that labelled Portuguese data is beneficial for training NER models in Galician, while incorporating Spanish data has, in general, adverse effects (or little macro F1 in the Mixed scenario for the monolingual model). Furthermore, it is worth noting that when

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 10. Results for cross-lingual training data augmentation using mDeBERTa_v3-base embedding model, compiling Spanish and Portuguese NER

Train dataset	Test dataset	F1	P	R	MicroF1	MacroF1	BS	LR
AnCora	SLI NERC test	86.88	85.94	87.85	**83.91**	**84.27**	16	5e-05
	Treegal test	83.65	83.23	84.08				
	PUD	84.26	83.24	85.31				
	CorNER	84.19	83.65	84.74				
	LREC	82.34	81.03	83.69				
Portuguese NER	SLI NERC test	87.07	85.75	88.42	**85.96**	**86.32**	8	2e-05
	Treegal test	84.40	84.32	84.49				
	PUD	81.87	81.31	82.42				
	CorNER	93.00	92.51	93.51				
	LREC	85.24	84.27	86.23				
Mixed	SLI NERC test	86.77	85.13	88.49	**85.35**	**85.70**	8	3e-05
	Treegal test	84.72	85.15	84.29				
	PUD	84.08	84.04	84.11				
	CorNER	88.60	88.46	88.74				
	LREC	84.34	83.34	85.35				

BS—batch size; LR—learning rate. All the provided models were trained using a configuration of 5 epochs.

augmenting the training data with cross-lingual examples, improvements are higher in the monolingual model than in the multilingual one, with BERT-base-gl achieving the best macro and micro F1-scores.

4.6 Comparison with LLMs

Although not the primary objective of our study, we include a comparison between few-shot methods using LLMs and our best supervised model, BERT-base-gl fine-tuned with the Mixed cross-lingual data. It is worth noting that generative LLMs are known to present challenges when applied to sequence labelling tasks, as their outputs may exhibit structural mismatches with gold-standard annotations, including token misalignment, missing or duplicated tags, and invalid label sequences (Zhou *et al.* 2024).[n]

We compared four state-of-the-art open-weight instruction-tuned models with coverage for Galician: (i) Qwen2.5-7B-Instruct (Qwen Team 2024),[o] selected as the best-performing LLM for Galician in the IberBench leaderboard (González *et al.* 2026)[p]; (ii) Qwen3-4B-Instruct-2507[q] and (iii) Qwen3-8B[r] (Qwen Team 2025), included as successor models; and (iv) Llama-3.2-3B-Instruct[s] as a representative of the Llama 3.2 family (Grattafiori *et al.* 2024). Regarding the prompting strategy, we evaluated both zero-shot and few-shot ($k = 3$) settings and report the latter, as it consistently yielded better results (see Section A.3 in the Appendix for further details).

[n] Several strategies have been proposed to improve the reliability of LLM outputs for NER, including advanced prompt engineering techniques (Ashok and Lipton, 2023; Wang *et al.* 2025) as well as model distillation and adaptation approaches (Zhou *et al.* 2024). We leave a systematic exploration of these methods for future work.
[o] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[p] https://huggingface.co/spaces/iberbench/leaderboard
[q] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507
[r] https://huggingface.co/Qwen/Qwen3-8B
[s] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 11. LLM performance (F1-score) across NER datasets under fuzzy span evaluation

Dataset	Qwen2.5-7B	Qwen3-4B	Qwen3-8B	Llama-3.2-3B	BERT-gl
CORGA	62.3	60.0	**65.2**	18.0	*92.3*
LREC	**63.7**	0.620	62.4	16.1	*88.1*
PUD	60.6	58.7	**60.8**	19.2	*88.3*
SLI	**58.6**	55.7	53.1	15.3	*93.2*
TreeGal	58.0	**58.7**	55.0	16.3	*91.8*
Macro avg	**60.6**	59.0	59.3	17.0	90.7

Bold values mark the highest score achieved by the evaluated LLMs on each of the test datasets.

Due to their generative nature, LLMs can produce outputs that are incompatible with the gold standard, so an evaluation strategy must be defined to maximise the alignment between both annotations. Following recent studies on sequence labelling (Paolini *et al.* 2021) and LLM-based NER (Santos *et al.* 2024), we first apply the Needleman–Wunsch global sequence alignment algorithm (Needleman and Wunsch 1970) between each LLM output and the corresponding test set. Then, we project the predicted tags onto the gold tokens as follows: gold tokens aligned to gaps are assigned the tag O, LLM output tags are repeated when multiple gold tokens align to a single LLM token, and tokens inserted by the LLM that do not align with any gold token are discarded. Finally, the performance at the entity level is computed using a fuzzy span-matching criterion, in which any overlap between predicted and gold entity spans is considered a match (Zhou *et al.* 2024).

Table 11 reports the F1-scores of the evaluated LLMs (see Table 18 in Section A.3 of the Appendix for the complete results). As shown, the performance of the LLMs is substantially lower than that of our best model, the baseline, and the previously proposed supervised models. These findings are consistent with recent work showing that, even in controlled settings, LLMs tend to produce noisy NER annotations (Merdjanovska *et al.* 2024). However, as mentioned above, it should be noted that one of the main factors underlying this low performance is the nature of the models' outputs and their varying degrees of adherence to the provided instructions. In particular, the Llama model exhibited especially poor performance, largely due to frequent deviations from the input sentences, such as generating tags for empty tokens and assigning a large number of named entity labels to common words (e.g., determiners, prepositions, and verbs). More extensive prompt engineering could potentially mitigate these issues. The reported values for Llama are therefore included for completeness and should not be interpreted as a fair or direct comparison.

4.7 Significance tests

To assess whether our best-performing models perform significantly better than the baseline, we have carried out a series of statistical significance tests.

First, we compared the overall performance of BERT-base-gl and mDeBERTa_v3-base, both trained on monolingual data, using the Friedman Aligned Ranks test. This choice was made because the data were not normally distributed and did not meet the homoscedasticity assumption. Following the approach of Rodriguez-Fernandez *et al.* (2015), we obtained significant results with $p = 0.028$, which is below the significance threshold of $\alpha = 0.05$.

Next, we examined whether our best-performing model (BERT-base-gl trained with mixed data) significantly outperformed the baseline across datasets. To do so, we applied the Wilcoxon Signed-Rank Test (Wilcoxon 1992) using F1-scores from the five datasets. The results confirmed statistically significant performance gains, with a consistent improvement validated by the Wilcoxon Signed-Rank Test ($z = -2.023$, $p < 0.05$).

Overall, the new models trained in this study demonstrate statistically significant improvements over the previous state-of-the-art NER model for Galician.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 12. Qualitative analysis micro average results for Galician NER models trained with BERT-based embeddings and evaluated on five Galician datasets. Values in bold are the best F1 results for each category and for each metric (bottom row)

	Baseline			BERT-base-gl			mDeBERTa_v3-base			BERT-base-gl (cross-lingual)		
Tags	F1	P	R	F1	P	R	F1	P	R	F1	P	R
PER	95.34	95.30	95.39	95.45	94.62	96.30	95.75	95.12	96.40	**96.07**	94.60	97.60
LOC	86.11	83.06	89.39	**87.25**	85.37	89.22	86.58	84.95	88.26	86.90	86.01	87.81
ORG	81.41	79.66	83.24	82.36	80.17	84.67	82.55	79.01	86.43	**84.01**	81.29	86.92
MISC	72.95	75.64	70.45	72.09	72.51	71.67	73.82	76.34	71.47	**74.82**	77.45	72.35
Total	84.70	84.02	85.54	85.07	83.91	86.41	85.47	84.32	86.69	**86.30**	**85.42**	**87.23**

5. Qualitative analysis

To gain a better understanding of how the models work, this section presents the results of a qualitative analysis that assesses the behaviour of the best models across entity types and explores their performance on datasets where results are more dissimilar.

5.1 Performance across entity types

In order to obtain qualitative data on the performance of the best models, we applied the approach proposed by Tjong Kim Sang and Buchholz (2000) to calculate the F1-score for each named entity type, with the results compiled in Table 12. As observed, the highest accuracy among the annotated named entities was achieved for person entities, with scores of 95.34 for the baseline model, 95.45 for BERT-base-gl, 95.75 for mDeBERTa_v3-base, and 96.07 for BERT-base-gl trained with cross-lingual data integration. In contrast, the lowest metrics were recorded for miscellaneous entities, with scores of 72.95 for the baseline model, 72.09 for BERT-base-gl, 73.82 for mDeBERTa_v3-base, and 74.82 for BERT-base-gl trained with cross-lingual data integration.

The numbers in Table 12 show consistent improvements across entity types for both monolingual and multilingual models trained with Galician and cross-lingual data. As shown in the final row of Table 12, the overall F1-score for named entities reached 86.30% on the mixed evaluation dataset, which compiles five Galician corpora, using the BERT-base-gl monolingual model and an augmented cross-lingual training dataset. This represents an improvement over the initial 84.70% F1-score achieved by the baseline model, and our BERT-base-gl NER model outperforms the previous state of the art by nearly two points across all named entity classes.

As for the span-wise evaluation of the BERT-base-gl model evaluated on the Mixed scenario, Table 13 reveals substantial variation in performance across entity types. The model achieves the highest accuracy for person entities, with an F1-score of 96.2, and a minimal error rate of 7.74%. Locations show similarly robust performance (87.4 F1). In contrast, organisational and miscellaneous entities pose greater challenges, with F1-scores of 84.9 and 75.9, respectively, and error rates of 30.88% and 45.81%. The lower recall for MISC entities indicates frequent omissions, while the elevated false-positive rates for ORG suggest occasional over-prediction or boundary errors. Overall, the model attains a micro-averaged F1-score of 87%, corresponding to an aggregate error rate of 26.11%, underscoring strong generalisation capabilities in cross-lingual settings while highlighting the need for targeted improvements in specific categories, such as MISC and ORG.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 13. Span-wise metrics per entity type from Mixed dataset for the best-performing BERT-base-gl (cross-lingual) model

Entity	TP	FP	FN	Errors	Error%	Precision	Recall	F1
PER	2,026	107	54	161	7.74	94.98	97.40	96.18
ORG	1,933	395	292	687	30.88	83.03	86.88	84.91
MISC	1,062	266	407	673	45.81	79.97	72.29	75.94
LOC	1,553	232	217	449	25.37	87.00	87.74	87.37
Overall	6,574	1,000	970	1,970	**26.11**	86.80	87.14	**86.97**

Table 14. NE labels errors for BERT-base-gl predictions evaluated on five Galician datasets and trained with cross-lingual data

	SLI NERC		TreeGal		PUD		CorNER		LREC	
Tags	#	%	#	%	#	%	#	%	#	%
O	95	28	37	32	93	29	8	6	233	22
B-LOC	29	8	13	11	110	34	8	6	54	5
I-LOC	22	6	9	8	38	12	4	3	30	3
B-PER	10	3	6	5	23	7	7	6	50	5
I-PER	9	3	2	2	16	5	7	6	48	5
B-ORG	36	10	11	10	20	6	35	28	234	22
I-ORG	45	13	11	10	8	2	9	7	234	22
B-MISC	45	13	8	7	12	4	20	16	89	9
I-MISC	54	16	18	16	5	2	27	22	72	7
Total #	345	0.85	115	1.07	**325**	**1.38**	125	0.60	**1,044**	**2.21**
Total tokens	40,405		10,112		**23,509**		20,932		**47,254**	

5.2 Performance across datasets

5.2.1 Best-performing model outcomes

Across the evaluation datasets, a consistent trend is observed: the SLI and CorNER datasets achieve the highest accuracy, while the PUD and LREC corpora consistently yield the lowest results. This pattern was previously noted in Sections 4.4 and 4.5, during the experiments with monolingual and multilingual models and cross-lingual data augmentation.

Building on these findings, it becomes crucial to conduct a qualitative analysis of NER task performance for both the BERT-base monolingual model and the mDeBERTa multilingual model. This analysis should focus on their performance within the PUD and LREC datasets, which consistently exhibited the lowest accuracy scores. Additionally, it is essential to extend this evaluation to the Combined Galician NER datasets, which aggregate all five Galician corpora, to understand better how these models perform when trained on a more diverse and comprehensive dataset.

In order to address the disparity in F1-scores observed between the PUD and LREC datasets, a qualitative analysis was conducted using the BERT-base-gl model's predictions, which achieved higher performance (86.30% for micro and 86.72% for macro average F1) and was trained on cross-lingual data. The results are presented in Table 14.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Regarding the SLI corpus, comprising 40,405 tokens, 345 tokens (0.85% of the total dataset) were identified as erroneous. Among these errors, the most significant proportion (95 tokens, or 28% of the total errors) were O-tagged tokens, indicating they were either missed or not identified as named entities. As for the TreeGal corpus, consisting of 10,112 tokens, the model predicted 115 erroneous tokens (1.07% of the total corpus). Similarly, the majority of these errors were concentrated in missed or unidentified named entity tokens.

The PUD results show a slightly different trend, with a total of 23,509 tokens and 325 erroneous tokens (1.38% of the entire corpus). The majority of these errors are found in *location* entities, accounting for 110 cases, or 34% of the total errors. The CorNER corpus, consisting of 20,932 tokens, contains 125 erroneous tokens (0.60% of the total corpus), with most errors occurring within the *organisation* class, comprising 35 cases (28% of all errors). Finally, the LREC dataset, which includes 47,254 tokens, contains 1,044 erroneous tokens, representing 2.21% of the corpus—the highest error rate among the five datasets. In LREC, the most frequent errors (22%) occur across two classes: O-tagged tokens and *organisation* entities.[t]

Although errors related to O-tagged tokens are commonly observed across the datasets (with four out of five exhibiting this pattern), the notably higher error rate in *organisation*-tagged entities for the LREC dataset predictions is particularly striking, as shown in the heatmap presented in Figure 2. The performance of BERT-base-gl drops significantly on the LREC test dataset, with the highest number of erroneous tokens occurring in O-tagged tokens (233 cases) and *organisation* entities (234 cases each for B-ORG and I-ORG). These results are followed by the PUD predictions, which show 110 errors in *location* entities and 93 in unidentified named entities. These findings suggest that BERT-base-gl likely struggles to resolve ambiguity between *location* and *organisation* entities, as explained in Subsection 3.3.2.

5.2.2 Best-performing monolingual versus multilingual model outcomes

Additionally, to investigate the differences in micro and macro F1-scores observed in the monolingual and multilingual models trained with the cross-lingual data (Mixed scenario), we focused our analysis on the final training split. This split included the SLI and TreeGal datasets, as well as cross-lingual data from the AnCora and Portuguese NER corpora, and was evaluated on the Mixed test dataset. The Mixed test dataset comprises all five Galician corpora, as detailed in Tables 2 and 3.

As shown in Figure 3, the BERT-base-gl monolingual model has predicted 1,952 erroneous tokens, while the mDeBERTa-v3-base model shows a slightly higher total of 2,115 errors. This suggests that, despite its more advanced architecture, the mDeBERTa multilingual model is prone to more errors when performing NER in Galician. As for the highest error count across both models, it occurs in O-tagged tokens, which account for 24% of total errors in the BERT monolingual model and 21% in the mDeBERTa multilingual model. This may indicate that both models tend to miss named entities or fail to identify them accurately. Regarding other errors within NE classes, B-LOC and I-LOC failed tags are more prominent in mDeBERTa multilingual predictions, with 254 (12%) and 125 (6%) cases, respectively, compared to BERT-monolingual, with 214 (11%) and 103 (5%) errors. This suggests that mDeBERTa struggles slightly more with identifying the boundaries and continuity of location entities.

mDeBERTa also exhibits higher errors for B-PER and I-PER tags, with 149 (7%) and 106 (5%) errors, respectively, compared to BERT's 95 (5%) and 81 (4%) errors. This suggests that a multilingual transformer-based model may struggle to distinguish individual names accurately. The errors in B-ORG and I-ORG tags are substantial for both models, though the numbers are similar.

[t]To facilitate the interpretation of these results, Table 15 in the Appendix presents the tag frequency data across the five Galician datasets.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

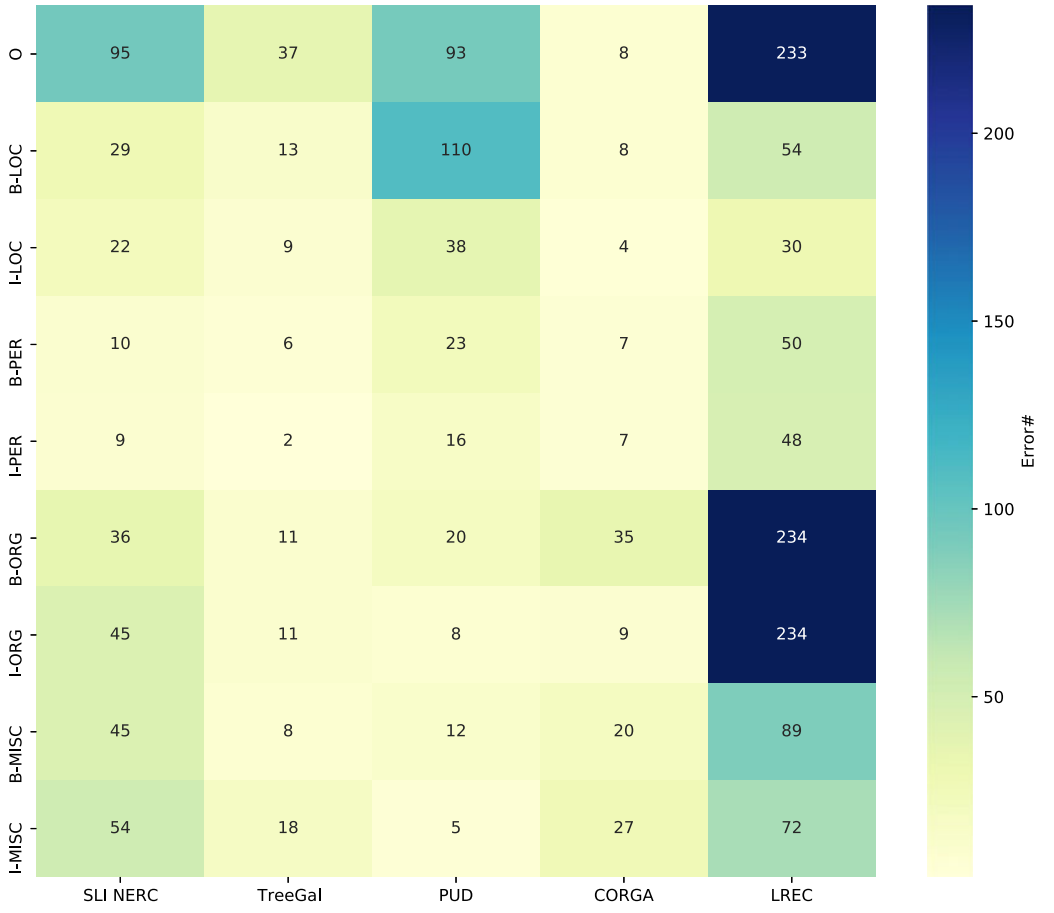


Figure 2. Distribution of the erroneously annotates NE for BERT-base-gl predictions among five Galician datasets.

For BERT-base-gl, there are 336 (17%) and 307 (16%) errors, while mDeBERTa has 347 (16%) and 326 (15%) errors.

This finding supports the observed tendency identified during our analysis of the distribution of erroneously annotated named entities in the BERT monolingual predictions across five Galician datasets. Specifically, it suggests that both monolingual and multilingual models may be having difficulty distinguishing between ambiguous cases involving location and organisation entities. Finally, errors for B-MISC and I-MISC tags are comparable across the models. BERT monolingual predictions exhibit 174 (9%) and 176 (9%) errors, while mDeBERTa multilingual shows 178 (8%) and 180 (9%) errors. Compared to other entity types, miscellaneous entities do not pose a significant challenge for both monolingual and multilingual BERT-based models in Galician.

Overall, both models exhibit the highest error rates on O-tagged tokens and organisation entities, suggesting that both the BERT-based monolingual and multilingual models may struggle to distinguish between ambiguous cases involving location and organisation entities, as previously discussed. The multilingual mDeBERTa model tends to have slightly higher error rates across most categories, particularly for person and location entities, compared to the BERT monolingual Galician model.

Regarding the annotation of named entity tags and their associated error rates, an analysis was conducted on the predictions made by the BERT monolingual Galician model across five datasets to explore potential discrepancies in relation to annotation errors, especially for the PUD and

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

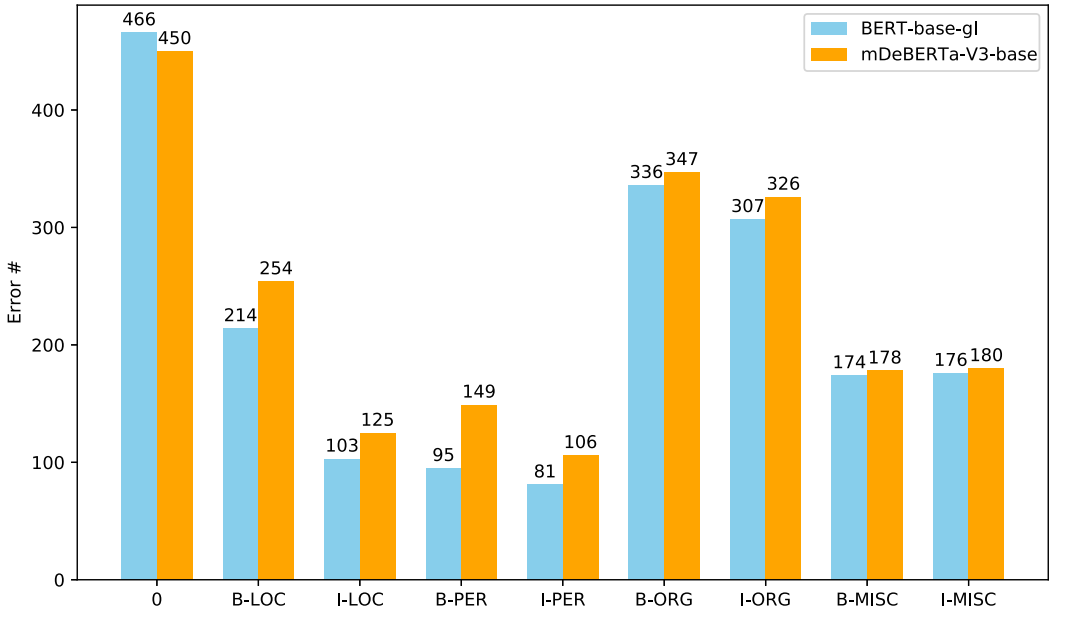


Figure 3. NE labels errors for BERT-base-gl fine-tuned with cross-lingual data evaluated on five Galician test dataset.

LREC datasets, which showed the highest error rates. The results, both in terms of error percentages and the most common NE tag errors, suggest that revising the NER tagging guidelines may be considered. This is particularly evident from the high rates of O-tagged token errors observed in four of the five datasets, as shown in Figure 2.

6. Discussion

The results obtained under the Mixed training scenario indicate the beneficial impact of cross-lingual data augmentation on Galician NER performance. As reported in Section 4.5, increasing the training corpus to 991,436 tokens led to a noticeable improvement for the monolingual BERT-base-gl model, which reached a macro F1-score of 87.03. In contrast, the multilingual mDeBERTa_v3-base model experienced a slight performance degradation, suggesting that additional heterogeneous data may not be as beneficial for multilingual models as it is for monolingual models. Across all evaluation datasets, both monolingual and multilingual models achieved their best performance on SLI and CorNER. In contrast, PUD and LREC consistently proved more challenging, corroborating earlier observations from Section 4.4. Overall, these findings confirm that labelled Portuguese data is particularly beneficial for training NER models in Galician. At the same time, including Spanish data tends to yield limited or even negative gains, especially in macro F1. Moreover, the fact that the monolingual model benefits more from cross-lingual augmentation than the multilingual one underscores the advantage of language-specific representations, with BERT-base-gl ultimately achieving the best overall results.

The performance metrics obtained for Galician are relatively comparable to the best fine-tuning transformers for English, Portuguese, and Spanish NER, as described in Section 2. Depending on the dataset and language, F1-scores for these languages are around 90%, achieving very high performance (≈97%) in some cases, but also lower values in different domains (e.g., 60% on social media data). More complex models, using cooperative learning, embedding concatenation, and external resources, achieve slightly better results (Wang *et al.* 2021a, 2021b). Unlike previous approaches trained exclusively on large monolingual datasets, the proposed Galician NER

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

model attains competitive accuracy using relatively small training data by exploiting cross-lingual resources.

As for the qualitative analysis results of the final training split in the Mixed cross-lingual scenario, the following insights into the differences in micro F1 performance between the monolingual BERT-base-gl and the multilingual mDeBERTa-v3-base models were observed. When evaluated on the Mixed test dataset, which comprises all five Galician corpora, BERT-base-gl produced fewer overall errors (1,952 erroneous tokens) than mDeBERTa (2,115), indicating a more robust behaviour for Galician NER despite the latter's multilingual architecture. For both models, the most significant proportion of errors occurs in O-tagged tokens, accounting for 24% in BERT-base-gl and 21% in mDeBERTa, suggesting a common tendency to miss or misrecognise named entities. Beyond this general pattern, mDeBERTa consistently shows higher error rates across several named entity classes. In particular, errors related to location entities (B-LOC and I-LOC) are more frequent in the multilingual model, pointing to greater difficulty in correctly identifying entity boundaries and continuity. A similar trend is observed for person entities (B-PER and I-PER), where mDeBERTa shows notably higher error rates, indicating challenges in distinguishing individual names. Errors for organisational entities (B-ORG and I-ORG) are substantial in both models and comparable in magnitude, suggesting that organisational entities may also pose a challenging category across models. Overall, these findings help explain the observed micro F1 differences and highlight that, in this cross-lingual setting, the monolingual BERT-base-gl model achieves more reliable performance for Galician NER.

Regarding the limitations of our study, the qualitative error analysis of the best-performing model suggests that the imbalance among NER categories in the training data may contribute to errors in certain entity classes. Although the class imbalance is not severe, as shown in Tables 3 and 8, additional experiments would be necessary to verify whether a more balanced training dataset could further improve Galician NER performance. Therefore, experiments using larger, more diverse training and evaluation datasets are required to assess the model's performance stability.

7. Conclusions

The study presented in this work aims to contribute to the field of NER for Galician. In summary, we emphasise the following key contributions, as outlined in Section 1.

We have examined the impact of various factors on NER for Galician, including training data, pre-trained Transformer models, hyperparameter selection, and cross-lingual data augmentation. Our study introduced two new manually annotated datasets with high inter-annotator agreement and leveraged five distinct evaluation datasets to robustly assess model performance across varying conditions.

Our results show that (1) among the models evaluated, the multilingual mDeBERTa-base demonstrated the best performance when using monolingual training data, highlighting the strength of multilingual representations in this scenario. However, the monolingual BERT-base-gl obtained similar results, and the new DeBERTa-xsmall-gl—released as part of this study—achieved competitive results despite its smaller size, underscoring its efficiency. (2) Augmenting the training data with related varieties, particularly Portuguese, shifted the performance landscape. In this scenario, the monolingual BERT-base model consistently outperformed mDeBERTa, achieving the best results in almost all cases. This underscores the value of leveraging cross-lingual data for low-resource language tasks using monolingual models. (3) The best-performing models demonstrated consistent improvements across all evaluation datasets, with significant increases in F1-scores and large effect sizes, and consistently outperformed few-shot LLMs on the NER task, reinforcing the robustness of the results. (4) Our qualitative analysis revealed that 'person' entities were identified with the highest accuracy. However, we also highlighted the need for more harmonised annotation standards across datasets to enhance model performance further.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Moreover, our work addressed several critical gaps in the Galician NER landscape, outlined in Sections 1 and 2. The lack of datasets has been a constraint on evaluating the Galician NER model's robustness and cross-domain generalisation, limiting progress in this under-resourced language. In response to this challenge, we contribute on multiple fronts. First, we support the expansion of Galician NER resources by creating two new manually annotated datasets. Second, we adapt and leverage existing Galician linguistic resources, promoting their reuse and applicability for NER tasks. Finally, we conduct systematic evaluations of NER models across five datasets, enabling more reliable performance comparisons and advancing the understanding of models' performance in Galician. Collectively, these contributions help mitigate long-standing resource and evaluation limitations and represent a step forward for Galician NER research.

Finally, it is worth to note that, by releasing the new annotated datasets and new state-of-the-art models,[u] we aim to provide valuable resources for the NLP community, fostering further research and development in NER for Galician and related linguistic varieties.

Acknowledgements. This work was funded by the Galician Government (ED431G 2023/04 and ED431B 2025/16) and by MCIN/AEI/10.13039/501100011033 (grants with references PID 2021-128811OA-I00, PID2024-161928OB-I00, CNS2024-154902, and AIA2025-163322-C62). The publication was also funded by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU—NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*.

This study is part of the research project of Momentum CSIC Programme (ref. MMT24–IEGPS–01) within the framework of the Generación D initiative, promoted by Red.es, an entity attached to the Ministry for Digital Transformation and the Civil Service, to attract and retain talent through scholarships and training contracts funded by the Recovery, Transformation, and Resilience Plan supported by the Next Generation EU funds.

Author contributions. **Albina Sarymsakova** and **Marcos Garcia** designed the experimental framework and processed and analysed the data. **Albina Sarymsakova** conducted experiments and evaluated BERT models. **Albina Sarymsakova** and **Helena Pérez Puente** annotated and revised two new Galician NER corpora. **Ettore Mariotti** trained the new extra-small DeBERTa model for Galician. **Albina Sarymsakova** wrote the manuscript with input from all authors. **Marcos Garcia** conceived the study and was responsible for overall direction and planning.

References

Agerri R., Gómez Guinovart X., Rigau G. and Solla Portela M.A. (2018). Developing new linguistic resources and tools for the Galician language. In Calzolari N., Choukri K., Cieri C., Declerck T., Goggi S., Hasida K., Isahara H., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S. and Tokunaga T. (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Agerri R. and Rigau G. (2016). Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence* **238**, 63–82.

Albuquerque H.O., Souza E., Gomes C., de C. Pinto M.H., Filho R.P.S., Costa R., de M. Lopes V.T., da Silva N. F. F., de Carvalho A.C.P.L.F. and Oliveira A.L.I. (2023). Named entity recognition: A survey for the Portuguese language. *Procesamiento de Lenguaje Natural* **70**, 171–185.

Alves D., Kuculo T., Amaral G., Thakkar G. and Tadic M. (2020). UNER: Universal named-entity recognitionframework.

Artstein R. and Poesio M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* **34**(4), 555–596.

Ashok D. and Lipton Z.C. (2023). PromptNER: Prompting For Named Entity Recognition.

Babych B. and Hartley A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT Workshop on MT and Other Language Technology Tools, Improving MT through Other Language Technology Tools, Resource and Tools for Building MT at EACL 2003*.

Bade G., Kolesnikova O. and Oropeza J. (2024). The role of named entity recognition (ner): Survey. *International Journal of Computer and Organization Trends* **14**(3), 1–7.

[u]The datasets (https://github.com/albinasarymsakova/Named-Entity-Recognition-Resources-for-Galician), the DeBERTa model (https://huggingface.co/proxectonos/deberta-xsmall-gl), and its fine-tuned variant (https://huggingface.co/asarymsakova/bert-base-gl-ner) introduced in this study are publicly accessible, ensuring transparency, reproducibility, and broad usability for the research community.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Barcala F.M., **Domínguez E.M.**, **Gamallo P.**, **López M.**, **Moscoso E.M.**, **Rojo G.**, **Santalla M.P. and Sotelo S.** (2007). El proyecto gari-coter en el seno del proyecto ricoterm2. *Procesamiento del Lenguaje Natural* **39**, 295–296.

Blouin B., **Favre B.**, **Auguste J. and Henriot C.** (2021). Transferring modern named entity recognition to the historical domain: How to take the step? In Hämäläinen M., Alnajjar K., Partanen N. and Rueter J. (eds), *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. NIT Silchar, India: NLP Association of India (NLPAI), pp. 152–162.

Cañete J., **Chaperon G.**, **Fuentes R.**, **Ho J.-H.**, **Kang H. and Pérez J.** (2020). Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Carreras X., **Màrquez L. and Padró L.** (2002). Named entity extraction using AdaBoost. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement* **20**(1), 37–46.

Collovini S., **Santos J.**, **Consoli B.**, **Terra J.**, **Vieira R.**, **Quaresma P.**, **Souza M.**, **Claro D.B. and Glauber R.** (2019). IberLEF 2019 Portuguese named entity recognition and relation extraction tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*, pp. 390–410.

Conneau A., **Khandelwal K.**, **Goyal N.**, **Chaudhary V.**, **Wenzek G.**, **Guzmán F.**, **Grave E.**, **Ott M.**, **Zettlemoyer L. and Stoyanov V.** (2019). Unsupervised cross-lingual representation learning at scale. CoRR, abs/1911.02116.

de Dios-Flores I. and Garcia M. (2022). A computational psycholinguistic evaluation of the syntactic abilities of Galician BERT models at the interface of dependency resolution and training time. arXiv preprint arXiv:2206.02440.

de Dios-Flores I., **Magariños C.**, **Vladu A.I.**, **Ortega J.E.**, **Pichel J.R.**, **García M.**, **Gamallo P.**, **Fernández Rei E.**, **Bugarín-Diz A.**, **González González M.**, **Barro S. and Regueira X.L.** (2022). The nós project: Opening routes for the Galician language in the field of language technologies. In Aldabe I., Altuna B., Farwell A. and Rigau G. (eds), *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 52–61.

de Dios-Flores I., **Suárez S.P.**, **Pérez C.C.**, **Outeiriño D.B.**, **Garcia M. and Gamallo P.** (2024). CorpusNÓS: A massive Galician corpus for training large language models. In Gamallo P., Claro D., Teixeira A., Real L., Garcia M., Oliveira H.G. and Amaro R. (eds), *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, pp. 593–599.

de Marneffe M.-C., **Manning C.D.**, **Nivre J. and Zeman D.** (2021). Universal dependencies. *Computational Linguistics* **47**(2), 255–308.

Derczynski L., **Nichols E.**, **van Erp M. and Limsopatham N.** (2017). Results of the WNUT2017 shared task on novel and emerging entity recognition. In Derczynski L., Xu W., Ritter A. and Baldwin T. (eds), *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 140–147.

Devlin J., **Chang M.-W.**, **Lee K. and Toutanova K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein J., Doran C. and Solorio T. (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Feng X., **Feng X.**, **Qin B.**, **Feng Z. and Liu T.** (2018). Improving low resource named entity recognition using cross-lingual knowledge transfer. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, pp. 4071–4077.

Fernández Rei F. (1991). *Dialectoloxía da Lingua Galega*. Vigo: Xerais de Galicia.

Garcia M. (2021). Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In Zong C., Xia F., Li W. and Navigli R. (eds), *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)*. Association for Computational Linguistics, pp. 3625–3640.

Garcia M. (2024). Training and evaluation of vector models for Galician. *Language Resources and Evaluation* **58**(4), 1–44.

Garcia M. and Gamallo P. (2014). Multilingual corpora with coreferential annotation of person entities. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. and Piperidis S. (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3229–3233.

Garcia M., **Gayo I. and González López I.** (2012). Identificação e classificação de entidades mencionadas em galego. *Estudos de Lingüística Galega* **4**, 13–25.

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

González J., **Borrego Obrador I.**, **Ivaro Romo Herrero Á.**, **Sarvazyan A.M.**, **Chinea-Ríos M.**, **Basile A. and Franco-Salvador M.** (2026). IberBench: LLM evaluation on Iberian languages. *Computer Speech & Language* **96**, 101899.

Grattafiori A., **Dubey A.**, **Jauhri A.**, **Pandey A.**, **Kadian A.**, **Al-Dahle A.**, **Letman A.**, **Mathur A.**, **Schelten A.**, **Vaughan A.**, **Yang A.**, **Fan A.**, **Goyal A.**, **Hartshorn A.**, **Yang A.**, **Mitra A.**, **Sravankumar A.**, **Korenev A.**, **Hinsvark A.**, **Rao A.**, **Zhang A.**, **Rodriguez A.**, **Gregerson A.**, **Spataru A.**, **Roziere B.**, **Biron B.**, **Tang B.**, **Chern B.**, **Caucheteux C.**, **Nayak C.**, **Bi C.**, **Marra C.**, **McConnell C.**, **Keller C.**, **Touret C.**, **Wu C.**, **Wong C.**, **Ferrer C.C.**, **Nikolaidis C.**, **Allonsius D.**, **Song**

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

D., Pintz D., Livshits D., Wyatt D., Esiobu D., Choudhary D., Mahajan D., Garcia-Olano D., Perino D., Hupkes D., Lakomkin E., AlBadawy E., Lobanova E., Dinan E., Smith E.M., Radenovic F., Guzmán F., Zhang F., Synnaeve G., Lee G., Anderson G.L., Thattai G., Nail G., Mialon G., Pang G., Cucurell G., Nguyen H., Korevaar H., Xu H., Touvron H., Zarov I., Ibarra I.A., Kloumann I., Misra I., Evtimov I., Zhang J., Copet J., Lee J., Geffert J., Vranes J., Park J., Mahadeokar J., Shah J., van der Linde J., Billock J., Hong J., Lee J., Fu J., Chi J., Huang J., Liu J., Wang J., Yu J., Bitton J., Spisak J., Park J., Rocca J., Johnstun J., Saxe J., Jia J., Alwala K.V., Prasad K., Upasani K., Plawiak K., Li K., Heafield K., Stone K., El-Arini K., Iyer K., Malik K., Chiu K., Bhalla K., Lakhotia K., Rantala-Yeary L., van der Maaten L., Chen L., Tan L., Jenkins L., Martin L., Madaan L., Malo L., Blecher L., Landzaat L., de Oliveira L., Muzzi M., Pasupuleti M., Singh M., Paluri M., Kardas M., Tsimpoukelli M., Oldham M., Rita M., Pavlova M., Kambadur M., Lewis M., Si M., Singh M.K., Hassan M., Goyal N., Torabi N., Bashlykov N., Bogoychev N., Chatterji N., Zhang N., Duchenne O., Çelebi O., Alrassy P., Zhang P., Li P., Vasic P., Weng P., Bhargava P., Dubal P., Krishnan P., Koura P.S., Xu P., He Q., Dong Q., Srinivasan R., Ganapathy R., Calderer R., Cabral R.S., Stojnic R., Raileanu R., Maheswari R., Girdhar R., Patel R., Sauvestre R., Polidoro R., Sumbaly R., Taylor R., Silva R., Hou R., Wang R., Hosseini S., Chennabasappa S., Singh S., Bell S., Kim S.S., Edunov S., . . .Ma Z. (2024). The Llama 3 Herd of Models. *arXiv preprint*. arXiv:2407.21783. https://doi.org/10.48550/arXiv.2407.21783.

Grishman R. and Sundheim B. (1996). Message understanding conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Gutiérrez-Fandiño A., Armengol-Estapé J., Pàmies M., Llop-Palao J., Silveira-Ocampo J., Carrino C.P., Armentano-Oller C., Rodriguez-Penagos C., Gonzalez-Agirre A. and Villegas M. (2022). MarIA: Spanish language models. *Procesamiento del Lenguaje Natural* **68**(0), 39–60.

Habibi M., Weber L., Neves M., Wiegandt D.L. and Leser U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* **33**(14), i37–i48.

He P., Gao J. and Chen W. (2021a). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Eleventh International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda. arXiv preprint 2111.09543.

He P., Liu X., Gao J. and Chen W. (2021b). DeBERTa: Decoding-enhanced BERT with disentagled attention. In *International Conference on Learning Representations*.

Keraghel I., Morbieu S. and Nadif M. (2024). Recent advances in named entity recognition: A comprehensive survey and comparative study. https://arxiv.org/abs/2401.10825, 2401.

Khade O., Jagdale S., Phaltankar A., Takalikar G. and Joshi R. (2025). Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In Sarveswaran K., Vaidya A., Krishna Bal B., Shams S. and Thapa S. (eds), *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*. Abu Dhabi, UAE: International Committee on Computational Linguistics, pp. 217–222.

Kim S., Toutanova K. and Yu H. (2012). Multilingual named entity recognition using parallel data and metadata from Wikipedia. In Li H., Lin C.-Y., Osborne M., Lee G.G. and Park J.C. (eds), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 694–702.

Levy O., Goldberg Y. and Dagan I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* **3**, 211–225.

Lezama-Sánchez A.L. and Vidal M.T. (2024). Named entity recognition in scientific texts using long short-term memory networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024)*. Valladolid, Spain: CEURWS. org.

Lindley Cintra L.F. and Cunha C. (1984). *Nova Gramática do Português Contemporâneo*. Lisbon: Livraria Sá da Costa.

Lu Y., Liu Q., Dai D., Xiao X., Lin H., Han X., Sun L. and Wu H. (2022). Unified structure generation for universal information extraction. In Muresan S. Nakov P. and Villavicencio A. (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5755–5772.

Mateo C.G. and Rodríguez M.A. (2012). *Language Technology Support for Galician*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 50–67.

Merdjanovska E., Aynetdinov A. and Akbik A. (2024). NoiseBench: Benchmarking the impact of real label noise on named entity recognition. In Al-Onaizan Y., Bansal M. and Chen Y.-N. (eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 18182–18198.

Mina M., Rodríguez C., Gonzalez-Agirre A. and Villegas M. (2024). Extending off-the-shelf NER systems to personal information detection in dialogues with a virtual agent: Findings from a real-life use case. In Volodina E., Alfter D., Dobnik S., Lindström Tiedemann T., Muñoz Sánchez R., Szawerna M.I. and Vu X.-S. (eds), *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*. St. Julian's, Malta: Association for Computational Linguistics, pp. 44–53.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Mira Mateus M.H., **Brito A.M.**, **Duarte I.**, **Hub Faria I.**, **Frota S.**, **Matos G.**, **Oliveira F.**, **Vigário M. and Villalva A.** (2003). *Gramática da Língua Portuguesa*, 6 edn. Lisbon: Caminho.

Monteagudo H., **Loredo X. and Vázquez-Grandío G.S.** (2024). *Informe sobre a situación da lingua galega segundo os datos do Instituto Galego de Estatística de 2023: Análise da enquisa sobre coñecemento e uso do idioma galego (2023) e da evolución sociolingüística de Galicia (1992–2023)*. Real Academia Galega.

Nadeau D., **Turney P.D. and Matwin S.** (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Québec City, Québec, Canada, June 7–9, 2006. Proceedings 19*. Springer, pp. 266–277.

Needleman S.B. and Wunsch C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**(3), 443–453.

Nowakowski A., **Pałka G.**, **Guttmann K. and Pokrywka M.** (2022). Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In Koehn P., Barrault L., Bojar O., Bougares F., Chatterjee R., Costa-jussà M.R., Federmann C., Fishel M., Fraser A., Freitag M., Graham Y., Grundkiewicz R., Guzman P., Haddow B., Huck M., Jimeno Yepes A., Kocmi T., Martins A., Morishita M., Monz C., Nagata M., Nakazawa T., Negri M., Névéol A., Neves M., Popel M., Turchi M. and Zampieri M. (eds), *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 326–334.

Padró L. and Stanilovsky E. (2012). FreeLing 3.0: Towards wider multilinguality. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Moreno A., Odijk J. and Piperidis S. (eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2473–2479.

Paolini G., **Athiwaratkun B.**, **Krone J.**, **Ma J.**, **Achille A.**, **Anubhai R.**, **dos Santos C.N.**, **Xiang B. and Soatto S.** (2021). Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

Park Y., **Son G. and Rho M.** (2024). Biomedical flat and nested named entity recognition: Methods, challenges, and advances. *Applied Sciences* **14**(20), 9302.

Pires A.R.O. (2017). Named entity extraction from Portuguese web text. Master's thesis, Universidade do Porto.

Pires T., **Schlinger E. and Garrette D.** (2019). How multilingual is multilingual BERT? In Korhonen A., Traum D. and Màrquez L. (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001.

Qwen Team (2024). Qwen2.5: A party of foundation models.

Qwen Team (2025). Qwen3 technical report.

Ritter A., **Clark S.M. and Etzioni O.** (2011). Named entity recognition in tweets: An experimental study. In Barzilay R. and Johnson M. (eds), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK: Association for Computational Linguistics, pp. 1524–1534.

Rodriguez-Fernandez I., **Canosa A.**, **Mucientes M. and Bugarin A.** (2015). STAC: A web platform for the comparison of algorithms using statistical tests. In *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8.

Rojo G., **Martınez M.**, **Noya E.D. and Barcala F.** (2015). Corpus de adestramento do etiquetador/lematizador do galego actual (xiada), versión 2.6.

Samartim R. (2012). Língua somos: A construção da ideia de língua e da identidade coletiva na galiza (pré-) constitucional. In Rodríguez-González O., Carballo Piñeiro L. and Baltrusch B. (eds), *Novas achegas ao estudo da cultura galega II: enfoques socio-históricos e lingüístico-literarios*. A Coruña: Universidade da Coruña, pp. 27–36.

Sánchez J.M.R., **Fernández L.D. and García-Mateo C.** (2023). *Language Report Galician*. Cham: Springer International Publishing, pp. 143–146.

Santos D., **Seco N.**, **Cardoso N. and Vilela R.** (2006). HAREM: An advanced NER evaluation contest for Portuguese. In Calzolari N., Choukri K., Gangemi A., Maegaard B., Mariani J., Odijk J. and Tapias D. (eds), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA).

Santos J., **Cameron H. F.**, **Olival F.**, **Farrica F. and Vieira R.** (2024). Named entity recognition specialised for Portuguese 18th-century history research. In Gamallo P., Claro D., Teixeira A., Real L., Garcia M., Oliveira H.G. and Amaro R. (eds), *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, pp. 117–126.

Sarymsakova A., **Sánchez-Rodríguez X. and Garcia M.** (2024). Towards accurate dependency parsing for Galician with limited resources. *Procesamiento del Lenguaje Natural* **73**, 247–257.

Segura-Bedmar I., **Martínez P. and Herrero-Zazo M.** (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (∗ SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 341–350.

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

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

H.G. and Amaro R. (eds), *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, pp. 587–592.

Strauss B., **Toma B.**, **Ritter A.**, **de Marneffe M.-C. and Xu W.** (2016). Results of the WNUT16 named entity recognition shared task, Organizing Committee. In Han B., Ritter A., Derczynski L., Xu W. and Baldwin T. (eds), *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 138–144.

Taulé M., **Martí M.A. and Recasens M.** (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S. and Tapias D. (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 96–101.

Thiyagu T.M., **Manjula D. and Shruthi S.** (2019). Named entity recognition in biomedical domain: A survey. *International Journal of Computer Applications* **181**(41), 30–37.

Tjong Kim Sang E.F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tjong Kim Sang E.F. and Buchholz S. (2000). Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Tjong Kim Sang E.F. and De Meulder F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147.

Torge S., **Politov A.**, **Lehmann C.**, **Saffar B. and Tao Z.** (2023). Named entity recognition for low-resource languages - profiting from language families. In Piskorski J., Marcińczuk M., Nakov P., Ogrodniczuk M., Pollak S., Přibáň P., Rybak P., Steinberger J. and Yangarber R. (eds), *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1–10.

Uzuner Ö., **South B.R.**, **Shen S. and DuVall S.L.** (2010). 2010 I2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* **18**(5), 552–556.

Vania C., **Kementchedjhieva Y.**, **Søgaard A. and Lopez A.** (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In Inui K., Jiang J., Ng V. and Wan X. (eds), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 1105–1116.

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

Wang S., **Sun X.**, **Li X.**, **Ouyang R.**, **Wu F.**, **Zhang T.**, **Li J.**, **Wang G. and Guo C.** (2025). GPT-NER: Named entity recognition via large language models. In Chiruzzo L., Ritter A. and Wang L. (eds), *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 4257–4275.

Wang X., **Jiang Y.**, **Bach N.**, **Wang T.**, **Huang Z.**, **Huang F. and Tu K.** (2021a). Automated concatenation of embeddings for structured prediction. In Zong C., Xia F., Li W. and Navigli R. (eds), *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)*. Association for Computational Linguistics, pp. 2643–2660.

Wang X., **Jiang Y.**, **Bach N.**, **Wang T.**, **Huang Z.**, **Huang F. and Tu K.** (2021b). Improving named entity recognition by external context retrieving and cooperative learning. In Zong C., Xia F., Li W. and Navigli R. (eds), *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)*. Association for Computational Linguistics, pp. 1800–1812.

Wang X., **Shen Y.**, **Cai J.**, **Wang T.**, **Wang X.**, **Xie P.**, **Huang F.**, **Lu W.**, **Zhuang Y.**, **Tu K.**, **Lu W. and Jiang Y.** (2022). DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, pp. 1457–1468.

Wilcoxon F. (1992). Individual comparisons by ranking methods. In Kotz S. and Johnson N.L. (eds.), *Breakthroughs in Statistics: Methodology and Distribution*. New York: Springer, pp. 196–202.

Wolf T., **Debut L.**, **Sanh V.**, **Chaumond J.**, **Delangue C.**, **Moi A.**, **Cistac P.**, **Rault T.**, **Louf R.**, **Funtowicz M.**, **Davison J.**, **Shleifer S.**, **von Platen P.**, **Ma C.**, **Jernite Y.**, **Plu J.**, **Xu C.**, **Le Scao T.**, **Gugger S.**, **Drame M.**, **Lhoest Q. and Rush A.** (2020). Transformers: State-of-the-art natural language processing. In Liu Q. and Schlangen D. (eds), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45.

Yadav V. and Bethard S. (2018). A survey on recent advances in named entity recognition from deep learning models. In Bender E.M., Derczynski L. and Isabelle P. (eds), *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2145–2158.

Yan H., **Deng B.**, **Li X. and Qiu X.** (2019). Tener: adapting transformer encoder for named entity recognition, arXiv preprint arXiv:1911.04474.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

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

Zhou W., Zhang S., Gu Y., Chen M. and Poon H. (2024). UniversalNER: Targeted distillation from large language models for open named entity recognition. In *Proceedings the Twelfth International Conference on Learning Representations (ICLR 2024)*.

Appendix

A.1 Named entity tags per dataset

Table 15. Tag frequencies of five Galician datasets

	SLI NERC		TreeGal		PUD		CorNER		LREC	
Tags	#	%	#	%	#	%	#	%	#	%
O	37,198	92.06	9278	91.75	21,461	91.29	19,487	93.10	40,539	85.79
B-LOC	439	1.09	110	1.09	399	1.70	280	1.34	543	1.15
I-LOC	217	0.54	12	0.12	177	0.75	39	0.19	229	0.48
B-PER	213	0.53	141	1.39	420	1.79	121	0.58	1184	2.51
I-PER	229	0.57	125	1.24	185	0.79	100	0.48	1047	2.22
B-ORG	556	1.38	150	1.48	276	1.17	438	2.09	805	1.70
I-ORG	640	1.58	101	1.00	246	1.05	290	1.39	754	1.60
B-MISC	360	0.89	87	0.86	151	0.64	85	0.41	786	1.66
I-MISC	553	1.37	108	1.07	194	0.83	92	0.44	1367	2.89
Total #	40,405	100	10,112	100	23,509	10,000	20,932	100	47,254	100

Table 16. Total amount of annotated NE over five Galician datasets

Entity type	#	%
PER	2,461	22.29
LOC	3,096	28.05
ORG	3,450	31.25
MISC	2,032	18.41
Total	11,039	100.00

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

A.2 Pre-trained models

Table 17 details the main properties of the pre-trained models used for fine-tuning.

Table 17. Comparison of pre-trained models used for fine-tuning on Galician NER data. *Parameters* include both backbone and vocabulary parameters and are approximate values in millions. *Corpus* is the approximate size of the respective training corpus. Additional corpus size for BERT-base gl and DeBERTa_v3-xsmall-gl refer to the training corpus of each original model they are based on (mBERT and deberta-v3-xsmall)

Model	Lang	Param.	Pre-training Corpus	Hugging face handle
BERT-base-gl	gl	177	3.1 MB (+13 GB)	marcosgg/bert-base-gl-cased
BERT-small-gl	gl	83	3.1 MB	marcosgg/bert-small-gl-cased
Bertinho-base	gl	108	250 MB	dvilares/bertinho-gl-base-cased
Bertinho-small	gl	66	250 MB	dvilares/bertinho-gl-small-cased
mBERT	104	177	13 GB	bert-base-multilingual-cased
XLM-RoBERTa-base	100	277	2.5TB	FacebookAI/xlm-roberta-base
XLM-RoBERTa-large	100	559	2.5TB	FacebookAI/xlm-roberta-large
mDeBERTa_v3-base	100	278	2.5TB	microsoft/mdeberta-v3-base
DeBERTa_v3-xsmall-gl	gl	71	11 GB (+160 GB)	–

A.3 Comparison with large language models

This appendix provides a brief description of the prompting strategy employed for the LLMs and reports the corresponding results.

Prompting strategy

Input and output format. In preliminary tests, we evaluated several prompting strategies aimed at obtaining outputs which could be aligned with the gold-standard. We explored given the input both as plain text (e.g., 'Maruxa Souto foi incluída. . .') and tokenised ('Maruxa', 'Souto', 'foi', 'incluída'. . .). Regarding the output, we compared asking only for the tags (e.g., 'B-PER', 'I-PER', 'O', 'O') and for 'token TAG' pairs ('Maruxa B-PER', 'Souto I-PER', 'foi O', 'incluída O'). In every combination, we tested several format configurations, including processing the content of each sentence as a whole, token by token, etc. The best alignment results were obtained with the format shown in the example below.

Zero-shot versus few-shot. We evaluated both zero-shot and few-shot scenarios, including in the examples entities of the four ENAMEX classes. The best results were obtained with few-shot ($k = 3$), as exemplified in the following prompt.

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Few-shot prompt example:

```
Instruction:
    "Your task is to perform NER annotation on a text written in Galician."
    "Use the IOB2 annotation schema with the following classes: PER, ORG,
    LOC, MISC, and O."
    "Valid TAGs are O, B-PER, I-PER, B-LOC, I-LOC, B-MISC, I-MISC, B-ORG,
    I-ORG."
    "Input tokens are given ONE PER LINE."
    "Output must contain 'token TAG' pairs (EXACTLY ONE PER LINE), aligned
    with the input."
    "VERY IMPORTANT:DO NOT add extra text, punctuation or explanations."
    "IMPORTANT: THE NUMBER OF OUTPUT LINES SHOULD BE EXACTLY THE SAME
    AS THE NUMBER OF INPUT LINES."

Example 1 Input:
    "Maruxa", "Souto", ",", "de", "A", "Coruña", ",", "foi", "incluída",
    "na", "proposta", "da", "Federación", "Ecoloxista", "Galega",
    "(", "FEG", ")", "para", "a", "ampliación", "da", "Rede", "Natura",
    "2000"

Example 1 Output:
    "Maruxa B-PER", "Souto I-PER", ", O", "de O", "A B-LOC", "Coruña I-LOC",
    ", O", "foi O", "incluída O", "na O", "proposta O", "da O",
    "Federación B-ORG", "Ecoloxista I-ORG", "Galega I-ORG", "( O",
    "FEG B-ORG", ") O", "para O", "a O", "ampliación O ", "da O", "Rede B-MISC",

    "Natura I-MISC", "2000 I-MISC"

Example 2 Input: [...]
Example 2 Output: [...]
Example 3 Input: [...]
Example 3 Output: [...]

Input: [tokenised input]
```

Results

Table 18 displays the NER results of the LLMs compared to the best supervised model (BERT-base-gl fine-tuned with *Mixed* cross-lingual data, see Section 4.5). As mentioned, the Llama model had a very low performance mainly because it often produced outputs differing from the input sentences (generating labels tags for empty tokens, large number of NE tags in common words, etc.).

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.

Table 18. LLM performance (Precision, Recall, and F1-score) on NER under fuzzy span evaluation across datasets. Values in bold are the best P/R/F1 per dataset. BERT-gl, included for comparison, is the BERT-base-gl fine-tuned with *Mixed* cross-lingual data

	Qwen2.5-7B-Inst			Qwen3-4B-Inst-2507			Qwen3-8B			Llama-3.2-3B-Inst			BERT-gl		
Dataset	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
CORGA	0.527	0.762	0.623	0.483	0.790	0.600	**0.546**	**0.811**	**0.652**	0.181	0.179	0.180	0.926	0.920	0.923
LREC	**0.566**	0.728	**0.637**	0.526	**0.755**	0.620	0.537	0.744	0.624	0.191	0.139	0.161	0.875	0.886	0.881
PUD	**0.509**	0.750	0.606	0.467	**0.787**	0.587	0.498	0.781	**0.608**	0.187	0.196	0.192	0.886	0.880	0.883
SLI	**0.477**	**0.761**	**0.586**	0.445	0.743	0.557	0.416	0.735	0.531	0.145	0.163	0.153	0.930	0.934	0.932
TreeGal	**0.481**	0.730	0.580	0.478	**0.762**	**0.587**	0.453	0.699	0.550	0.164	0.162	0.163	0.923	0.912	0.918
Macro avg	**0.512**	0.746	**0.606**	0.480	**0.768**	0.590	0.490	0.754	0.593	0.174	0.168	0.170	0.908	0.906	0.907

Cite this article: Sarymsakova A, Mariotti E, Pérez Puente H and Garcia M (2026). Named Entity Recognition Across Datasets and Domains: Resources and Models for Galician. *Natural Language Processing* **32**, 555–588. https://doi.org/10.1017/nlp.2026.10040

Downloaded from https://www.cambridge.org/core. 05 Nov 2026 at 09:19:13, subject to the Cambridge Core terms of use.