CorpusNÓS: A massive Galician corpus for training large language models

Iria de-Dios-Flores[1,2] and **Silvia Paniagua Suárez**[1] and **Cristina Carbajal Pérez**[1] and **Daniel Bardanca Outeiriño**[1] and **Marcos Garcia**[1] and **Pablo Gamallo**[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2] Department of Translation and Language Sciences, Universitat Pompeu Fabra
{iria.dedios, silvia.paniagua.suarez, cristina.carbajal.perez, danielbardanca.outeirino, pablo.gamallo,marcos.garcia.gonzalez}@usc.gal

Abstract

CorpusNÓS is a massive Galician corpus made up of 2.1B words primarily devised for training large language models. The corpus sources are varied and represent a relatively wide range of genres. CorpusNÓS is, to the best of our knowledge, the largest collection of openly available Galician texts. This resource was created under the auspices of the Nós Project, and emerges as a fundamental prerequisite for developing language technologies in the era of deep learning.

1 Introduction

This work presents CorpusNÓS, a massive Galician corpus made up of 13.95GB of text (2.1B words) primarily devised for training large language models (LLMs). It represents, to the best of our knowledge, the largest collection of Galician texts openly available to date. This resource was created under the auspices of the Nós Project, and emerges as a fundamental prerequisite for developing language technologies in Galician in the era of deep learning. The corpus is divided into two subcorpus depending on how the texts were obtained (either via transfer agreement from the text owners or from publicly available sources). CorpusNÓS, as well as the cleaning pipeline developed to process the texts, is made available via the project's official GitHub repository: https://github.com/proxectonos/corpora.

The paper is structured as follows: by way of introduction, we present The Nós Project (section 1.1) and provide some notes on Galician LLMs that help situate the present contribution (section 1.2). The bulk of the work is concentrated in section 2, which presents the corpus structure, statistics, and a detailed description of the data sources. The processing and cleaning strategies are described in section 3. To conclude, section 4 discusses the applications of the resource and the future work we plan to carry out.

1.1 The Nós Project

The Nós Project (*Proxecto Nós*) is an initiative by the Universidade de Santiago de Compostela aimed at providing Galician with openly licensed resources and tools in the area of language technologies. Galician is a low-resource Romance language with around 2M speakers and very weak technological support (Sánchez and Mateo, 2022; García and de Dios-Flores, 2023). The project has been set up to address key challenges in several NLP areas (see de Dios-Flores et al. (2022) for further details), and has two cross-cutting objectives: (i) the compilation of high-quality linguistic resources, and (ii) the training of large language models. It is against this backdrop that we have compiled the resource reported here, which is a necessary step towards training state-of-the-art autoregressive and autoencoding LLMs -an endeavor that is already in progress.

1.2 Galician LLMs in context

Training LLMs using state-of-the-art architectures presents a critical challenge for low-resource languages, as they require the availability of huge amounts of text. This was already true a few years ago upon the publication of the first BERT model (Devlin et al., 2019), which was trained on 3.3B words of English text (notably, not so far from the size of the corpus presented here). Yet, further architectural developments, and particularly generative models, have become even much more data-hungry, as illustrated by GPT3 (Brown et al., 2020), which was trained on 181B words of English text.[1]

Several multilingual models have included Galician texts in their training data by making use of massive crawled corpus, although this inclusion is mostly anecdotal (and sometimes difficult to estimate). For instance, multilingual BERT (Devlin

[1]https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

et al., 2019), trained on 104 languages using the largest Wikipedias, included roughly 40M words of Galician text, and GPT3, trained on 118 languages (including programming ones) using a version of the C4 corpus, included 6M words of Galician text. Yet, to our knowledge, there was no available model with a performance in Galician that was at least somehow comparable to that of moderately-resourced languages until the release of Galician BERT (small and base) by Garcia (2021)[2], trained on a corpus of 550M words, which included the Galician Wikipedia plus a variety of web contents crawled by the authors - which are published for the first time as part of CorpusNÓS. In this context, CorpusNÓS represents a very noticeable improvement with respect to the data available thus far for LMM training, and it is paving the way for the creation of better models.

2 Corpus structure, statistics, and data sources

CorpusNÓS is a collection of many heterogeneous sources comprising 2.1B words and 9.7M documents. The corpus sources are varied and represent a relatively wide range of genres. It is published in plain text, and divided into different files for each of the sources. Within each file, the documents (e.g. different books, pieces of news, etc.) are separated by two line breaks. The materials are published under the CC BY 4.0 license, except for already published materials, which are released under their original license. The corpus is released in a partially deduplicated version to enable the use of the separate files for different purposes (see section 4 for details). Table 1 presents the structure of the corpus, the data sources grouped by genre, and size statistics. Importantly, the corpus is organized in two subcorpus containing data obtained via different processes. These are described in detail in the next sections.

2.1 Data obtained via a transfer agreement

Since the launch of the Nós Project, great efforts have been made to engage cultural agents, institutions, associations, and companies from the Galician society to generously donate their textual production to enable our language modeling enterprise. This is an ongoing initiative that has been carried out with the support of a legal team to ensure all guarantees in terms of copyright. The data obtained via transfer agreement total up to over 400M words and represent roughly 19% of CorpusNÓS. Despite being the smallest subcorpus, it is the collection with the highest quality in terms of language and curation. The texts included have been produced by professionals who can be attributed with a high native language competence (e.g. journalists, writers, civil servants, etc.). Furthermore, the vast majority of the documents included in this section have been obtained in markup languages which allowed us to extract clean plain text. Some PDFs have been included after a thorough processing (see section 3). The data sources in this subcorpus are organized by genres, as described in the next subsections.

2.1.1 Books

This collection contains 104 books originally written in other languages and translated into Galician by professional translators. These include 10 fiction novels donated by the publishing house Hugin & Munin, 51 fiction novels donated by the publishing house Urco, and 43 books that make up the collection *Classics of universal thinking* released by Universidade de Santiago de Compostela, which contains translations of works of scientific or humanistic thought.

2.1.2 Research articles

This collection contains 664 research articles originally written in Galician and published in different journals managed by the Universidade de Compostela's publishing services. Although the topics are varied, most articles belong to the fields of social sciences and humanities (e.g. linguistics, economy, and sociology).

2.1.3 Press

This collection contains 223.133 pieces of news that comprise the entire archive of several general and specialized online journals (*Nós Diario*, *Praza Pública*, *Código Cero*, *Tempos Novos*, and *Que Pasa na Costa*) up to 2022. Additionally, we have included the newscast ladders from the Galician public TV channel (CRTVG) between 2019 and 2022.

2.1.4 Governmental

This collection contains 654.505 documents extracted from three sources: (i) the Official Gazette of Galicia between 2000 and 2023, (ii) the Official Gazette of Coruña's Provincial Council between

[2]It should be noted that it was preceded by the release of Bertinho by Vilares et al. (2021), trained on 45M words from Wikipedia.

Subcorpus	Genre	Nº tokens	Nº documents
1. Data obtained via transfer agreement	Books	7.255.784	104
	Research articles	2.665.351	664
	Press	124.253.084	224.419
	Governmental	245.897.880	654.505
	Web contents	15.946.686	44.165
	Encyclopedic	4.799.214	47.396
	Subtotal	**400.817.999**	**971.253**
2. Public data	Press and blog	153.497.883	665.265
	Encyclopedic	57.164.848	184.628
	Web crawls	1.384.015.664	3.366.449
	Translation corpora	133.726.004	4.745.799
	Subtotal	**1.728.404.399**	**8.777.514**
	Total	**2.129.222.398**	**9.748.767**

Table 1: Corpus statistics.

2009 and 2022, and (iii) the Galician Parliament's Journal of Sessions between 2015 and 2022. The first two sources contain documents disclosing legal regulations and other acts of the administration (announcements, calls, competitions for public office, etc.). Galician Parliament's Journal of Sessions is a summary of the speeches and addresses made in the parliamentary chambers.

2.1.5 Web contents

This collection contains 44.165 documents extracted from two types of online sites. On the one hand, it contains the web repositories donated by three cultural institutions (i.e. Consello da Cultura, IGADI, and Editorial Galaxia) that share cultural information online via their archives (e.g. reports, news, etc.). On the other hand, it contains the entire web repository of two institutional domains: `xuntal.gal` by Xunta de Galicia, and `depo.gal` by Deputación de Pontevedra.

2.1.6 Encyclopedic

This collection contains 47.396 entries of the *Universal Galician Encyclopaedia*, an encyclopedia of reference of Galician culture which includes universal themes contemplated from the Galician perspective as well as Galician themes (e.g. personalities, architecture, geography, etc.).

2.2 Public data

This subcorpus, amounting to 81% of CorpusNÓS, contains a variety of public data published or extracted by third parties, which were either not available in corpus usable format or which were available but with insufficient quality. Among other sources, such as the Galician Wikipedia, it includes our version of existing web crawls (e.g. mC4 or OSCAR). Our goal was to produce cleaner versions of these, since Galician is often intermingled with Spanish (and other languages) in these datasets. Despite containing less controlled or curated texts (hence, with a language quality that is difficult to estimate for some sources), these datasets represent a fundamental resource without which it would not be possible to train large architectures. They are described in the following two subsections.

2.2.1 Press and blog

This collection contains 665.265 pieces of news and blog entries discontinuously crawled from publicly available sources between the years 2009 and 2020 which were used to train the state-of-the-art BERT models for Galician (Garcia, 2021). They include data from the *Blogomillo* blogsphere, and Galician press, including extinct newspapers (e.g. *Vieiros*). These texts had not been made public until now, as they have been donated by the model author to be included in CorpusNÓS. Critically, we have removed those newspapers whose data have been obtained via transfer agreement and are thus included in the former subcorpus.

2.2.2 Encyclopedic

This collection contains a clean dump of the 184.628 entries of the Galician Wikipedia available up to mid-2023. It is shared under a CC-BY-SA 4.0 license following the original resource's license.

2.2.3 Massive web crawls

This collection is composed of our clean version of the Galician dataset from the mC4 Corpus released by Xue et al. (2021) under an Apache 2.0 license, and the Oscar Corpus published under a CC0 license (see Ortiz Suárez et al. (2019) for details). Together, they amount to 3.366.449 documents after deduplication. The quality problems of these resources, particularly for low-resource languages, are well known in the machine learning community (e.g. Kreutzer et al. (2022)), which is why we have deemed it necessary to produce cleaner versions of these datasets (see section 3 for details on cleaning and deduplication, which are particularly relevant for these two resources).

2.2.4 Translation corpora

This collection is composed of data extracted from corpora originally devised for machine translation purposes. Specifically, we included the Galician texts from four corpora that contained documents rather than isolated sentences. These are: (i) TED2020 (Reimers and Gurevych, 2020), containing Ted talk transcriptions released under a CC BY–NC–ND 4.0 license; (ii) OpenSubtitles (Lison and Tiedemann, 2016), including TV and movie subtitles, to which we added extra files from OpenSubtitles not included in the original corpus; (iii) Linux-GL, which includes data from Linux corpora KDE and GNOME; and (iv) CC-Matrix (Schwenk et al., 2021), a web-based collection of automatically aligned texts pulled from the CommonCrawl.

3 Data processing and cleaning

The following procedures were designed to process and clean the texts giving way to CorpusNÓS:[3]

Plain text extraction: data obtained via transfer agreement were received in a variety of formats. Plain text from XML and HTML files were processed using the library `BeautifulSoup` (Richardson, 2007). PDF files were clean and deskewed using the library `ocrmypdf`. Then, the main body of the text was selected using `pdfCropMargins` and `openCV` in order to extract plain text using `ocrmypdf`.

Noise reduction: this was the central part of the cleaning process and encompassed three steps. First, encoding problems were solved by making sure that all non-UTF8 characters, invalid or binary characters, and odd symbols were not present in the texts while preserving as many original characters as possible (e.g. other alphabets, mathematical symbols, currencies, etc.). This intricate task was facilitated through the development of Python scripts, leveraging the capabilities of the libraries `ftfy`, `unicodedata`, `re` and `emoji` and complemented by manually curated lists for special characters and their equivalents. This process was applied to all the files in the corpus. Second, and most importantly, to get rid of noisy input (code, lists, boilerplates, etc.) present in the web crawls, we trained a Galician bigram model, which was incorporated in `pyplexity`, an unsupervised cleaning method based on perplexity Fernández-Pichel et al. (2023). We adapted the original software by implementing a document-based read of the input so that the original documents were tagged with a perplexity score. To adjust the perplexity threshold, three annotators revised several random files with results ranging from the lowest perplexity score to values up to 15.000. This analysis showed that most noise appeared when increasing the threshold beyond 2500. Hence. documents with higher perplexity values were deleted. Furthermore, due to the varied nature of the data included in the public data subcorpus (and particularly in the massive web crawlers), it was crucial to incorporate a language filter that could distinguish between Galician and Spanish to delete texts exclusively in Spanish or those containing small Galician fragments inside a mostly Spanish text. For this end, we used `Quelingua` (Gamallo et al., 2016), a multilingual n-gram based tool. We specifically tackled the Spanish-Galician contrast because it was very common to find Spanish in the Galician files of the web crawls (e.g. bilingual web pages).

Deduplication: to facilitate the use of the individual files for different purposes, we are not publishing the corpus in a fully deduplicated version. Only the massive web crawls included in the public data subcorpus (i.e. section 2.2.3) were deduplicated to avoid the same web material entering the corpus twice. This process was performed document by document. To do so, the texts were normalized by removing trailing spaces and collapsing multiple spaces into a single space. Furthermore, documents smaller than 15 tokens were removed from the corpus. The resulting data was then filtered by creating a hashmap that stored the final

[3] All the scripts and documentation are available in https://github.com/proxectonos/corpora.

collection of unique documents. Additionally, we performed a full deduplication of the entire corpus to investigate how much of the corpus was unique. When doing so, its size is reduced by 183K words, showing that 99.91% of the text is original.

Post-processing: all the resulting files from the two subcopora were visually inspected, and several regular expression patterns were created to eliminate or fix specific remaining noise, particularly from the crawls (e.g. tabulations and white spaces, punctuation issues, code, uncompleted tags, etc.).

4 Conclusions and future work

CorpusNÓS represents a substantial increase in the textual material available for the training of LLMs in Galician. Its division illustrates the two avenues we have explored to gather the largest amount of text possible within our reach. On the one hand, those texts obtained via transfer agreement and published for the first time in this resource constitute a very valuable contribution, as their compilation was underpinned by three important premises: the legal dimension, as all the texts were donated by their copyright owners to be included in this resource (considering the effort that this entails), the quality dimension, as all the texts were written by professionals who can be attributed with a high native language competence, and its heterogeneity of genres, as we strived to gather texts that represented as many domains as possible. On the other hand, the texts that make up the public data subcorpus had not been previously published in a thoroughly cleaned corpus usable format, and represent a fundamental resource for LLM training.

It should be emphasized that the publication of this resource is only a starting point, as the efforts to increase CorpusNÓS will be sustained in time. We plan to release future versions when additional donated materials are received or when improved versions of the data or cleaning pipeline are produced. All this will be found in the project's official repository.[4]

At the moment, CorpusNÓS is being used to train a 1.3B GPT3 model and several small DeBERTa models, and it will be used in the coming months to produce different autoencoder and autoregressive (pre-trained and fine-tuned) models. We hope that the release of this corpus also contributes to placing the Galician language in a better position for any LLM initiative beyond the Nós Project.

[4]https://github.com/proxectonos/corpora

Acknowledgements

This publication was produced within the framework of the Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336, and by the Xunta de Galicia through the collaboration agreements signed in with the University of Santiago de Compostela in 2021 and 2022.

Additionally, the authors of this article received funding from MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR (TED2021-130295B-C33), the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), by MCIN/AEI/10.13039/501100011033 (grants with references PID2021-128811OA-I00, PLEC2021-007662, and TED2021-130295B-C33, the latter also funded by the European Union Next Generation EU/PRTR), a Ramón y Cajal grant (RYC2019-028473-I), and a Juan de la Cierva Grant (JDC2022-049433-I) funded by MCIN/AEI/10.13039/501100011033 and the European Union Next Generation EU/PRTR.

We are deeply grateful to all the entities that have generously donated their texts to the Nós project via transfer agreement. These are: Axencia para a Modernización Tecnolóxica de Galicia, Código Cero, Consello da Cultura Galega, Corporación Radio e Televisión de Galicia, Deputación de Coruña, Deputación de Pontevedra, Galaxia, Huging & Munin, Instituto Galego de Análise e Documentación Internacional, Nós Diario, Parlamento de Galicia, Praza Pública, Que Pasa na Costa, Servizo de Publicación da Universidade de Santiago de Compostela, Tempos Novos, Urco e Xunta de Galicia.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack

Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Iria de Dios-Flores, Carmen Magariños, Adina Ioana Vladu, John E. Ortega, José Ramom Pichel, Marcos García, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín-Diz, Manuel González González, Senén Barro, and Xosé Luis Regueira. 2022. The nós project: Opening routes for the Galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Fernández-Pichel, Manuel Prada-Corral, David E. Losada, Juan C. Pichel, and Pablo Gamallo. 2023. An unsupervised perplexity-based method for boilerplate removal. *Natural Language Engineering*, page 1–18.

Pablo Gamallo, Jose Ramom Pichel, Inaki Alegria, and Manex Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 170–177, Osaka, Japan.

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

Sofía García and Iria de Dios-Flores. 2023. GL-BLARK – A BLARK for minoritized languages in the era of deep learning: expertise from academia and industry. Project deliverable; EU project European Language Equality (ELE2).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

José Manuel Ramírez Sánchez and Carmen García Mateo. 2022. Deliverable D1.15 Report on the Galician Language. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

David Vilares, Marcos García, and Carlos Gómez-Rodríguez. 2021. Bertinho: Galician BERT representations. *Proces. del Leng. Natural*, 66:13–26.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.