Chapter 17
Language Report Galician

José Manuel Ramírez Sánchez, Laura Docío Fernández, and Carmen García-Mateo

Abstract This chapter reports on the current state of Language Technology (LT) for Galician. The main conclusion is that there are a limited number of resources, products, and technologies for the Galician language with text-based technologies and services being more mature than those based on speech processing. We start with general facts about Galician, followed by a high-level qualitative description of the LT situation for Galician, and conclude with recommendations for bridging the gap between Galician LT with Spanish and the other co-official languages of Spain.

1 The Galician Language

Galician is part of the Romance family of languages, closely related to Portuguese, and it is one of the co-official languages of Spain. The linguistic rights of Galician speakers are guaranteed and regulated under the Linguistic Normalisation Act, especially those related to administration, education, and media. Galician has about 1,926,000 speakers. There are still large Galician-speaking communities outside Spain (mainly in Europe and America). Their total size is unknown due to the variety and complexity of these communities.

The online presence of Galician is limited, with less than 0.1% of websites using it.[1] Nevertheless, some initiatives try to increase the presence of Galician on the web (PuntoGal[2] and Galipedia[3] are good examples). The official survey *Enquisa estrutural a fogares. Coñecemento e uso do galego* shows a generally low internet penetration and use by European standards, but between the ages of 15 and 44, the numbers are very similar to other European regions.[4]

José Manuel Ramírez Sánchez · Laura Docío Fernández · Carmen García-Mateo
Univ. of Vigo, Spain, jmramirez@gts.uvigo.es, ldocio@gts.uvigo.es, carmen.garcia@uvigo.es

[1] https://w3techs.com/technologies/details/cl-gl-

[2] https://dominio.gal

[3] https://meta.wikimedia.org/wiki/List_of_Wikipedias

[4] http://www.ige.gal/estatico/html/gl/OperacionsEstruturais/PDF/Resumo_resultados_EEF_Galego_2018.pdf

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_17

A substantial amount of digital content in the Galician language is generated by public institutions of the Autonomous Community of Galicia. In the last few years, the number of products and services developed has increased considerably, aimed at incorporating Galician into the digital society. The web portal of the Real Academia Galega and the Xunta de Galicia translator are noteworthy examples. Although some large enterprises (Microsoft, Apple, Google, Meta) offer a few products with support for Galician, many others do not (TikTok, Twitch, Adobe). However, there is a total lack of support for Galician in the virtual assistants market, where none of the popular solutions allows interaction via Galician.

2 Technologies and Resources for Galician

The 2012 META-NET White Paper on Galician (García-Mateo and Arza 2012) was moderately optimistic about the state of LT support for the language. Ten years later, the LT status for Galician has changed a bit (Sánchez and García-Mateo 2022). In our analysis, we noticed an increase in the resources and corpora created between 2018-2021 (67.7% of those indexed). However, tools and services developed in the same period have not increased to the same degree (37.3% of those indexed). There is a significant imbalance in the distribution of resources and corpora by technologies. Data in text format are the most common (more than 90%), whereas corpora for other technologies are very few (5% are multi-modal, and almost 2% are audio only).

Most of the resources come from three origins: non-Galician universities and research centres, Galician public institutions, and non-Galician private companies or public institutions. It is important to note that most of the resources, services, and tools created by non-Galician entities tend to belong to multilingual projects or products that include Galician as one of several languages. However, most of the resources, services, and tools created by Galician entities tend to focus on Galician, offering high-quality results.

Regarding the accessibility and use of resources for Galician, most of them have been developed by open source projects, research centres, or universities under GNU/GPL licences. Around 20% of the indexed items are not available for commercial purposes, and more than 10% of resources are under a proprietary licence.

The situation of Galician in terms of data and resources is optimistic for most of the technologies that process and use text. However, regarding multimedia data, there is an enormous gap. In that sense, speech processing technologies seem less mature than technologies based on text processing.

For Galician, key results regarding technologies and resources include:

- There are large reference text databases in modern and historical Galician with a balanced mix of various domains (economics, technology, or the legal field) (Piñeiro 2019; García-Mateo et al. 2014).
- There are some databases annotated with syntactic, semantic, or discursive information. However, the number and size of these resources decrease as more complex linguistic and semantic information is needed.

- Parallel databases with millions of tokens exist between Galician and other languages such as Spanish, Portuguese, and English (OPUS[5] is a good example). These databases have been used to develop machine translation systems in production and education environments for Portuguese or Spanish.
- A relevant model to highlight is Bertinho (Vilares Calvo et al. 2021), a monolingual BERT model for Galician. Bertinho implements state-of-the-art technology, and it is possible to use it in many NLP tasks. However, its developers state that Bertinho does not reach the size or performance of other monolingual versions, such as BETO for Spanish.
- Available multimedia resources are relatively limited, with little domain variability and usually recordings of readings. The acoustic quality is excellent though.
- Another gap is related to human-computer interaction, where the necessary tools and resources to put together chatbots, virtual assistants, and similar systems are poor or outdated.

Spain has national plans for both Artificial Intelligence (AI, Gobierno_de_España 2020a) and LT (specifically for NLP, Gobierno_de_España 2020b). These plans focus more on the potential, opportunities, and needs of Spanish LT, putting less emphasis on co-official languages such as Galician. Two national associations bring together the community of researchers on issues related to LT: *Sociedad Española de Procesamiento del Lenguaje Natural* with a focus on NLP, and the *Red Temática en Tecnologías del Habla* with its focus on speech processing.

The Autonomous Community of Galicia has its own strategy for AI.[6] This document describes the current environment of AI in Galicia and provides a roadmap for public investments and developments until 2030. There is also an initiative called Proxecto Nós, a regional LT plan for Galician focused on digital challenges promoted by the Galician regional government. Furthermore, there are many more projects related to LT in the Galician university environment, both from a linguistic and technological point of view. Another interesting fact is that from the number of companies in the Galician ICT industrial environment that use AI, only 21% are focused on cognitive assistants and just 12% on NLP. The Galician LT industry is very small, but a very active environment of spin-offs and public programmes exists dedicated to transferring knowledge from universities to the market.

3 Recommendations and Next Steps

The main goal of LT for Galician is to reach the level of other co-official languages of Spain, such as Catalan or Basque. In this sense, increasing the use of LT in Galician public services and institutions could be a necessary line of action to support and stimulate research and development of new resources and better tools. Galician

[5] https://opus.nlpl.eu

[6] https://amtega.xunta.gal/sites/w_amtega/files/20210608_estrategia_ia_gl.pdf

institutions are the producers of high-quality resources and tools for Galician. However, there is a lack of standardisation and dissemination of these products. An office that centralises and standardises all the LT resources and tools created for Galician could be a significant contribution to unifying all efforts.

Support for open source solutions (data and software) would be a good long-term strategy for small-market languages. These solutions allow the development and research of new technologies without having to face an initial investment barrier. Furthermore, an open-source policy encourages the creation of strong communities and guarantees some technological sovereignty from the interests of global markets and multinational corporations.

References

García-Mateo, Carmen and Montserrat Arza (2012). *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/galician.

García-Mateo, Carmen, Antonio Cardenal López, Xosé Luis Regueira, Elisa Fernández Rei, Marta Martinez, Roberto Seara, Rocío Varela, and Noemí Basanta (2014). "CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis". In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2653–2657.

Gobierno_de_España (2020a). *Estrategia Nacional de Inteligencia Artificial 2020*. https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2020/201202_np_ENIAv.pdf.

Gobierno_de_España (2020b). *Estrategia Procesamiento del Lenguaje Natural 2020*. https://drive.google.com/file/d/1eXlFdRNTmOx4sm3FQ439Z8zaeNqEFGiK/view.

Piñeiro, Centro Ramón (2019). *Corpus de Referencia do Galego Actual (CORGA) [3.2]*. http://corpus.cirp.gal/corga/.

Sánchez, José Manuel Ramírez and Carmen García-Mateo (2022). *Deliverable D1.15 Report on the Galician Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-galician.pdf.

Vilares Calvo, David, Marcos García González, and Carlos Gómez Rodríguez (2021). "Bertinho: Galician BERT representations". In: *Procesamiento del lenguaje natural*, pp. 13–26.