Automatic Metrical Scansion of Galician Poetry: First Results

Pablo Ruiz Fabo[1,2], Pauline Moreau[2], Anxo Alonso Pérez[1],
{pablo.ruiz.fabo,anxo.alonso.perez}@usc.gal, pauline.moreau2@etu.unistra.fr

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
[2]Université de Strasbourg, LiLPa UR 1339

Abstract

We present the first public, user-friendly system for Galician poetry scansion, a symbolic system derived from a well-performing mixed-meter Spanish scansion library. We adapted its resources to Galician and added a preprocessing module. The system achieves 88% per-line accuracy in exact stress-pattern match on data unseen during development, and has practical value: First, it helps create a large annotated corpus to train scansion systems. Second, its web interface can help engage a non-specialist public. Third, its current accuracy is helpful for annotating large volumes of poetry and studying metrical trends in Computational Literary Studies use cases.

1 Introduction

In the last decade, Natural Language Processing (NLP) has seen breakthroughs like neural attention, transformers and large language models (LLM), but some tasks still pose particular challenges. One of them is metrical scansion in poetry, or the detection of stress patterns in verse. For this task, recent literature suggests that symbolic methods are still competitive compared to LLMs. A reason may be that stress pattern analysis is far removed from models' pretraining task, based on the prediction of masked orthographic tokens rather than on (phonological) units more relevant for scansion, like syllables. Another NLP challenge is low-resource languages like Galician, a co-official language in the Galicia region (Spain), sometimes considered part of a continuum that includes Portuguese. Despite major Galician NLP projects (Gamallo et al., 2024), the language remains under-resourced for certain tasks, including the one addressed here.

The current work is part of a Computational Literary Studies (CLS) project[1] seeking to compare Galician 19th-century poetry with its Portuguese and Spanish counterparts, and thus assess how formal features of major literary traditions may have shaped canonical features in a peripheral one, such as Galician, which re-emerged in the 19th c. after a long period of reduced activity.

Nagy et al.'s (2025) computational work on the evolution of verse in classical Latin, the Renaissance, and 19th-century Europe showed the value of large metrically annotated corpora for comparing and assessing metrical distributions across traditions. As automatic scansion helps such annotation at scale, it is vital for projects like ours.

Symbolic and neural approaches and LLMs have been applied to scansion. All require annotated corpora for evaluation, and training if applicable. No such public electronic corpora existed in Galician prior to our project. In this context, we report on our work towards a first system for Galician scansion, based on a well-performing mixed-meter system for Spanish. The paper's contributions are:

- A first, symbolic, open-source **system for metrical scansion** in Galician, achieving ca. 88% accuracy in exact stress-pattern match per line. This has practical value, as it helps speed up the creation of metrically annotated corpora, for training potentially more flexible systems based on more recent NLP paradigms.
- **Manual annotations** for stress patterns and metrical phenomena in a Galician 19th-century poetry corpus, also available under an open license.

The paper is structured as follows: Section 2 reviews related work. Section 3 defines the task. Sections 4 and 5 describe the corpora and system, and Sections 6 and 7 present results and outlook.

2 Related Work

Focusing, for brevity, on Portuguese and Spanish only, a variety of scansion approaches have been implemented. An early rule-based system for Portuguese was created by Araújo and Mamede

[1]https://compellit.github.io/

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 994–1004

(2002). Mittmann (2016) created *Aoidos*, with a very complete 159 rule set to handle complex metrical phenomena. It was applied to canonical authors (mostly Brazilian from the 18th to 19th centuries) achieving 97.5% accuracy at distinguishing stressed vs. unstressed syllables (Mittmann, 2016, 141-2). Recently, Valenca and Calegario (2025) fine-tuned GPT 3.5, obtaining between 87.19% and 88.6% stress and syllable segmentation accuracy, depending on the number of training examples (3,520 vs. 7,200). Barbosa and Barbosa (2025) developed scansion methods within in a poetry generation system adapted to Brazilian Northeastern phonology and specific poetic styles.

In Spanish, several rule-based systems exist. Gervás (2000) developed an early one. Navarro-Colorado (2018) created a scansion system for sonnets, with 95% perfect stress-pattern match per line (per-line accuracy or *lacc*) in the ADSO 100 corpus (100 classical sonnets). Automatic syllabification is performed and parts of speech (PoS) are used to detect syllable tonicity, then rules resolve metrical ambiguities. De la Rosa et al. (2020) created *Rantanplan*, using syllabification, PoS and metrical heuristics. Besides the sonnet's fixed meter, where it attains 96.23% stress-match *lacc* on ADSO 100, it handles other meters, reaching 65.02% on the harder mixed-meter *Carvajal* corpus, with large metrical variety. *LibEscansión* by Sanz-Lázaro (2024) uses PoS and phonological transcription-based syllabification and gets 97.01% stress *lacc* on ADSO 100. Finally, *Jumper* by Marco Remón and Gonzalo (2021) identifies stress patterns without prior syllabification. It reaches 95% stress *lacc* on ADSO 100 and 82% on the harder, mixed-meter Carvajal corpus, outperforming other systems in Spanish mixed-meter scansion. We exploit it for scansion in Galician.

Moving on to neural approaches, Agirrezabal et al. (2017) used LSTM networks, reaching 90.84% stress *lacc* on ADSO 100. De la Rosa et al. (2021) fine-tuned encoders using 8,748 lines, getting 93.43% stress *lacc* on ADSO 100.

Based on the references above, symbolic systems so far outperform neural ones in Spanish. The same seems to hold for Portuguese, though the systems' evaluation corpora are not directly comparable.

3 Task Description and Challenges

We defined scansion as a twofold task. First, **stress pattern detection** takes a series of verse lines as input, and outputs the positions of each line's stressed metrical syllables. Second, **meter detection** outputs the number of metrical syllables for each line. Table 3 shows an example for a single verse-line.

Stress pattern detection presents several challenges. In Romance metrics, *lexical* syllables (obtained by applying the language's phonological constraints) need not match *metrical* ones. To allow stressed syllables to appear in particular positions, enabling specific stress-based rhythms, so-called *metaplasms* or metrical licenses may occur. The main ones for Galician are:

Synalepha: The final, vowel-final, syllable of a word is merged with the first, vowel-initial, syllable of the following word. E.g. *E*, *o* in ***e_os*** *or-na-tos*. Several syllables may be merged if the segments with higher sonority (low vowels) are in the middle part of the sequence.

Syneresis: Within a word, vowels from two syllables, which do not form a diphthong, are merged into a single syllable. E.g. *o* and *a* in *che-gues* ***voa****n-do li-xei-ra*.

Dialepha: Across the word boundary, vowels which could be pronounced as a diphthong are pronounced in separate syllables, like *a* and *ó* in *coa a-gui-lla-d****a ó*** *lom-bo*.

Dieresis: Within a word, a diphthong is split into two syllables. E.g. *a* and *i* in ***va-i*** *li-xe-i-ro*

Meter detection is affected by metaplasms and by the line's last stressed syllable position. We use Spanish-style conventions for meter definition (*contagem grave*), commonly applied in Galician metrics. How they differ from Portuguese ones is discussed in Mittmann (2016, 13) and Appendix A: If the line ends in a stressed syllable, a syllable is added to the meter. If the line's last stressed syllable is the antepenult, a syllable is deducted.

For each line, a scansion system must decide whether to apply metaplasms and determine its stress patterns and its meter, respecting the rules above. A line may be ambiguous between two or more meters and stress patterns. For a human, unambiguous lines in the context help decide what meter to target, and use metaplasms accordingly. An automatic system must address this challenge.

4 Corpora

We annotated two 19th-century Galician poetry corpora, based on web resources and on Santamarina et al. (2026). The **development corpus** (*dev*) consists of 2,065 lines by 7 authors. The system's

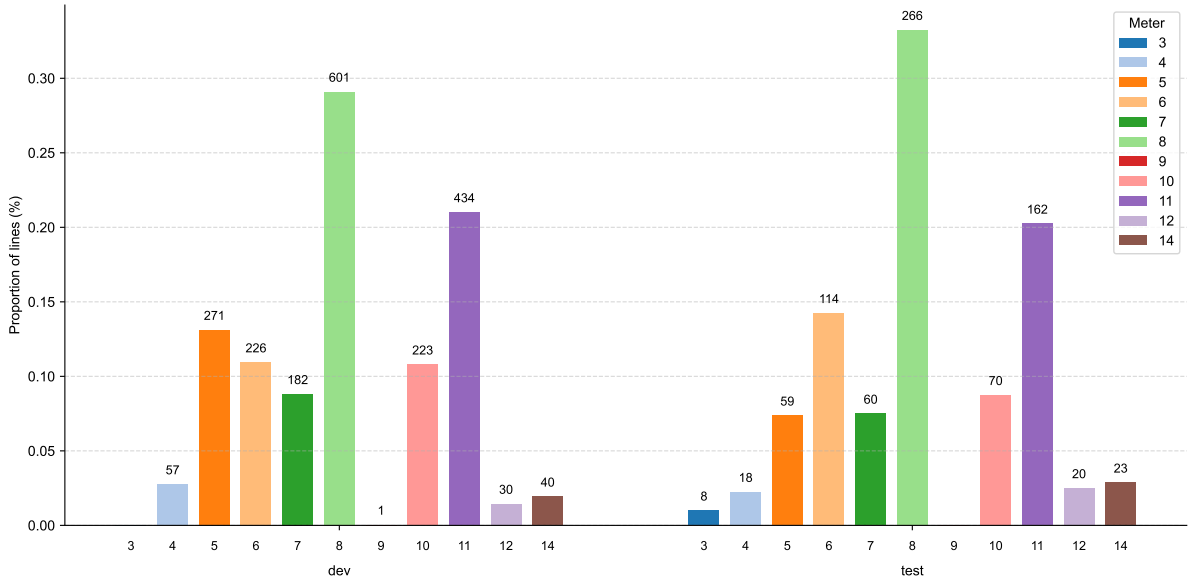

Figure 1: Meter distribution in *dev* and *test*. Syllable count follows the *contagem grave* or *contagem espanhola* convention, e.g. a line with the last metrical stress on syllable 7 is considered an *octosyllable* (column 8), and alexandrines are column 14. Column labels indicate the number of lines in the corpus for each meter.

lexical resources and preprocessing were gradually refined based on scansion evaluations with it. The **test corpus** (*test*) was not seen during development. It contains 800 lines (16 authors, 5 also in *dev*). A variety of meters occur: 3 to 14 syllables, with a comparable distribution in both corpora (Figure 1). In both, ~33% of lines come from mixed-meter poems, where at least two meters co-occur (rarely more than three), increasing scansion difficulty.

The *dev* corpus was manually annotated for stress patterns and meter by the first author, who triple checked his annotations, while the *test* was annotated by two of the authors (see inter-annotator agreement details below). Heuristics were implemented to decrease error chances: Impossible combinations of meters and stress patterns (like a meter smaller than the last stressed syllable position) were automatically flagged and corrected manually.

Lexical and metrical syllables were annotated in both corpora, including metaplasms (Appendix B shows an example). Metaplasm annotation supports system evaluation, helping asses how errors relate to specific metrical licenses. It also allows us to compare the difficulty of different corpus splits, clarifying how *dev* and *test* compare. Syllabification and scansion annotations are also useful for model training, work currently in progress.

The test corpus was manually annotated by two of the authors. We measured inter-annotator agreement (IAA) as the percentage of lines exactly matching for both annotators. IAA was 98.63% for stress patterns, 99.63% for meter (metrical syllable counts) and 97.63% for exact metrical syllabification (including metaplasm annotation).

For 18 lines in our 800-item test set, human annotators found more than one scansion plausible. The small proportion suggests that a clear solution exists for human annotators in most cases.

Poem titles, authors, and publication dates were stored as metadata. Original 19th c. spelling was kept, which is challenging in NLP terms.

5 System Description

The core algorithm is *Jumper*, fully described in Marco Remón and Gonzalo (2021), with their Python implementation. So far it had been applied to Spanish only, and we exploited it for Galician.

Jumper contains a **knowledge base** with lexically unstressed words (e.g. most prepositions), and possible diphthongs, in Spanish. We modified these for Galician, based on Carballo Calero (1966), Freixeiro Mato (1998) and Regueira Fernández (2010), also adding Galician triphthongs and logic to manage them. Jumper also has an exhaustive list of stress patterns in Spanish metrics (Jauralde Pou, 2020). Metrically relevant phonetic characteristics of Galician, like a clearer pronunciation of weak vowels compared to standard European Portuguese (Freixeiro Mato, 1998, 228), make Galician metrics close to Spanish, and we consider

Jauralde's stress pattern list accurate for Galician.

Determining a line's stress pattern and its meter requires **metrical ambiguity resolution**. For this, Jumper generates all possible candidate stress patterns for a line, based on metaplasm contexts (vowel sequences that could be merged or split), formalizing them as a vector representing the tonicity of syllable positions (±stressed). It then compares candidates to Jauralde's stress pattern inventory. Based on vector similarity to those patterns and to the candidates chosen for other lines in a configurable context window, it makes a final selection of the metaplasms that apply and of the resulting stress pattern. Meter is assigned based on this.

The orthographic representation of phonological stress in Spanish is unambiguous. Stressed vowels bear a stress mark or can be inferred from a word's orthographic final consonant: a word ending in <*s*> or <*n*> has penult stress unless the final syllable bears a stress mark. Jumper relies on this to find stressed syllables. In current Galician (following the ILG/RAG norm), the orthographic representation of stress is more ambiguous. The final <*n*>/<*s*> rule applies, but not when a final tonic syllable has a falling diphthong: *compás* ([kom.ˈpas]) vs. *demais* ([de.ˈmajs]). Another challenge in our corpus is that 19th c. Galician had no written norm and orthographic practices vary; an accent mark can indicate stress or vowel aperture.

Our scansion approach requires reliably identifying lexical stresses to find metrical ones. For this reason, we developed a **preprocessing** module, which reduces some of the ambiguity in the way stress is represented orthographically in Galician, by inserting or removing stress marks in some contexts. The resulting text is used solely as an intermediate representation that helps apply Jumper's metrical ambiguity resolution.

The **preprocessing** module contains an invocabulary (IV) lexicon for current Galician based on resources from the LinguaKit and Apertium libraries (Gamallo et al., 2018; Forcada et al., 2011). For out-of vocabulary items, besides regex-based corrections, it generates IV candidates via weighted edit distances with empirically determined weights (cf. Ruiz Fabo et al. 2014). For instance, adding a stress mark to a vowel is less costly than replacing it with an entirely new vowel. It then ranks the candidates up to a configurable edit distance given their probability in the line's context, using a 5-gram language model trained on 126 million Galician tokens from *CorpusNÓS* (De-Dios-Flores et al., 2024) using KenLM (Heafield, 2011); Appendix C provides preprocessing configuration details. This workflow restores most orthographic stress marks of current Galician and makes stress orthographically explicit in some extra contexts, helping improve scansion. While the approach relies on classical techniques, it achieved good results, required no training data and demands few computational resources.

Some tonicity ambiguity cases can be resolved with parts of speech (PoS), e.g. interrogative pronouns (stressed) vs. relative ones (unstressed). To define PoS-based rules, we integrated the Stanza tagger (Qi et al., 2020), using its Galician model trained on the TreeGal corpus (Garcia et al., 2018).

Finally, a public **web interface** (Appendix D) makes the tool accessible to non-specialists.

6 Results and Discussion

Preprocessing was evaluated on *test*. With a manual correction of the preprocessed text as reference, Word-Error Rate (WER) was 0.0071, computed with *werpy* (v3.1.0, Armstrong, 2025). Errors were found on 3% of lines (24 lines). We only counted *metrically relevant* errors, affecting syllable count or stress placement (see Appendix C).

	Configuration	dev		test	
		sm	**mm**	**sm**	**mm**
(1)	Original Spanish	38.6	84.1	48.5	92.0
(2)	(1) + Galician lexical resources	58.9	85.4	68.0	94.1
(3)	(2) + Preprocessing	**86.0**	**98.4**	**88.1**	**97.4**

Table 1: Exact stress-pattern match (*sm*) and meter match (*mm*). Per-line accuracy by configuration.

Stress and meter detection were evaluated on *dev* (2,065 examples) and *test* (800 examples). In stress match (*sm*), a line's stress pattern must match the reference exactly to count as correct. Meter detection is also assessed with an exact meter match (*mm*). All results in this section and its tables are **exact per-line matches**.

Table 1 shows the positive impact of adapting the tool to use Galician diphthongs and unstressed words. When coupled with preprocessing, 86% exact per-line accuracy was achieved in stress-pattern match (*sm*) in the *dev* corpus. In the test corpus, unseen during development, exact stress-pattern match per line was 88.1%.

Difficulty Source	dev		test	
	sm	**N**	**sm**	**N**
Synalepha	84.5	993	84.9	397
Dialepha	70.9	31	81.3	32
Syneresis	64.5	62	75	36
Dieresis	70	10	50	6
Synalepha > 2 syllables	76.9	26	83.3	12

Table 2: Exact stress-pattern match (*sm*) per-line accuracy in lines with analysis challenges, for the best configuration, (3) in Table 1. *N* is number of lines with each challenge.

ID		Text	Stress	Meter
1	input:	renovese a vida	2 5	6
	preprocessing:	renóvese a vida		

Table 3: Scansion result for one verse line. See Appendix E for a complete poem.

M.	dev			test		
	sm	**mm**	**N**	**sm**	**mm**	**N**
3	—	—	—	100	100	8
4	96.5	98.2	57	100	100	18
5	94.1	99.6	271	91.5	98.3	59
6	87.2	98.7	226	91.2	100	114
7	83.5	97.8	182	91.7	98.3	60
8	87.2	98.5	601	87.2	96.2	266
9	—	—	1	—	—	—
10	83.4	96.9	223	82.9	98.6	70
11	81.3	98.6	434	85.2	96.9	162
12	73.3	90	30	80	90	20
14	75	100	40	95.7	95.7	23
Avg.	**86**	**98.4**	**2065**	**88.1**	**97.4**	**800**

Table 4: Per-line accuracy in exact stress-pattern match (*sm*) and meter match (*mm*) by meter (*M.*). Meters follow *contagem grave ou espanhola*. Results for meters with less than 5 occurrences omitted. *N* is the total number of lines per meter. *Avg.* is micro-averaged.

On exact meter match (*mm*), the results are above 97% on both *dev* and *test*. Performance is higher than in *sm* because *mm* is less demanding: Determining the number of metrical syllables only requires correctly identifying syllable nuclei and the line's final stressed syllable.

Reporting exact stress-pattern match per line for lines that present **metrical analysis difficulties** (Table 2) is important because around 50% of lines had no metaplasms. This is no corpus flaw and reflects their natural distribution in Galician poetry. Actually, in the test corpus, we included poems with a somewhat higher variety of metaplasms than would be found at random, to get more solid evaluation statistics, so *test* may slightly overestimate the task's difficulty. Measuring a system specifically on lines that present a challenge gives a more complete picture of its value and it is helpful to report such data, plus the meter distribution in the corpus, to help compare evaluations across systems and corpora. Dialepha and syneresis are less frequent and posed more challenge than synalepha, although the system behaved reasonably, at >64%. Dieresis was too infrequent to draw conclusions.

In view of the related literature (Section 2), the results are encouraging, particularly for an initial system, evaluated on annotations that we needed to create from scratch. The system's upper bound is above the gains that perfect preprocessing could bring, because there is also room to refine its stress-pattern selection heuristics.

Results for individual meters are show in Table 4. The system performs at >81% exact stress match per line (*sm*) in the best represented meters (octosyllables and hendecasyllables) and performance is balanced in general.

7 Conclusion and Outlook

Starting from a well-performing scansion library previously applied to Spanish only, we adapted its resources to Galician. We added a preprocessing that reduces the ambiguity about syllable tonicity in Galician orthography, helping scansion. The system's accuracy gives it practical value as a first scansion system for Galician that helps create training and test corpora and can engage the general public via a web tool. It also reflects the practical methods that can sometimes help achieve results in low-resource settings for uncommon tasks.

Future work will exploit the ongoing creation of corpora, aided by system-provided annotations, coupled with annotations for lexical and metrical syllabification (all manually corrected), to train models based on different NLP paradigms and compare their performance. This would be enlightening, as existing literature does not clearly establish how different neural methods and LLMs compare with each other and with earlier paradigms.

Limitations

As mentioned in Section 4, in a small number of cases, annotators considered more than one scansion possible. These alternatives were recorded, but the system was evaluated against the first annotation recorded only. This may slightly underestimate per-line accuracy, as system outputs matching one of the other alternatives are not counted as correct.

Ethics Statement

We are aware of a **gender bias** in the corpus, that is biased towards male authors. One of the most important authors in Galician literature and a major figure of the 19th-century Galician Renaissance (*Rexurdimento*) is a woman, Rosalía de Castro, and her work is well represented in our corpus. Other 19th-century women authors are nevertheless lesser known and less present in available corpora. In our corpus, we included the only other woman author found in our sources: Filomena Dato Muruais. We would be interested in adding more women authors, and this would require deeper search and the creation of new electronic resources.

Data and Code Availability

At `https://github.com/compellit/gama-sym`.

Acknowledgments

This work was supported by the European Union, under the Marie Skłodowska-Curie Actions, HORIZON MSCA-2023-PF, Grant ID 101149659, COMPEL – Computational Analysis of Peripheral Literatures.

The work was also supported by Xunta de Galicia – Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024–2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022–2026, ED431C 2022/19) and by the European Union's European Regional Development Fund – ERDF.

We are grateful to Elisa Fernández Rei (Instituto da Lingua Galega, ILG), for full-text access to 19th-century electronic sources in *Tesouro informatizado da lingua galega* (Version 4.1), directed by Antón Santamarina, Ernesto González Seoane, and María Álvarez de la Granja (http://ilg.usc.gal/TILG/).

References

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A Comparison of Feature-Based and Neural Scansion of Poetry. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 18–23, Shoumen, Bulgaria.

Paulo Alexandre Araújo and Nuno J Mamede. 2002. Classificador de Poemas. In *CCTE conference*, Lisbon.

Ross Armstrong. 2025. werpy - Word Error Rate for Python. Accessed: 2026-03-14.

Bryan K S Barbosa and Marcela Y A Barbosa. 2025. CordelSextilha.BR: A Benchmark for Poetic Form in Brazilian Cordel Verse Generation. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 736–747, Fortaleza/CE, Brasil.

Ricardo Carballo Calero. 1966. *Gramática elemental del gallego común*. Galaxia, Vigo.

Ricardo Carballo Calero. 1981. *Historia da Literatura Galega Contemporánea*. Galaxia, Vigo. 2019 facsimile reprint.

Rogério Chociay. 1974. *Teoria do Verso*. McGraw-Hill do Brasil, São Paulo.

Iria De-Dios-Flores, Silvia Paniagua Suárez, Cristina Carbajal Pérez, Daniel Bardanca Outeiriño, Marcos Garcia, and Pablo Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 593–599, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Javier De la Rosa, Álvaro Pérez, Mirella de Sisto, Laura Hernández, Aitor Díaz, Salvador Ros, and Elena González-Blanco. 2021. Transformers analyzing poetry: multilingual metrical pattern prediction with transformer-based language models. *Neural Computing and Applications*, 35(25):18171–18176.

Javier De la Rosa, Álvaro Pérez, Laura Hernández, Salvador Ros, and Elena González-Blanco. 2020. Rantanplan, Fast and Accurate Syllabification and Scansion of Spanish Poetry. *Procesamiento del Lenguaje Natural*, 65:83–90.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Xosé Ramón Freixeiro Mato. 1998. *Gramática da lingua galega I: Fonética e fonoloxía*. A Nosa Terra, Vigo.

Pablo Gamallo, Marcos Garcia, César Piñeiro, Rodrigo Martinez-Castaño, and Juan C Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Pablo Gamallo, Pablo Rodríguez, Silvia Paniagua, Daniel Bardanca, José Ramom Pichel, and Marcos Garcia. 2024. Open Generative Large Language Models for Galician. *Procesamiento del Lenguaje Natural*, 73:259–270.

Marcos Garcia, Carlos Gómez-Rodríguez, and Miguel A Alonso. 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.

Pablo Gervás. 2000. A logic programming application for the analysis of Spanish verse. In *Computational Logic—CL 2000*, pages 1330–1344. Springer.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Pablo Jauralde Pou. 2020. *Métrica española.* Cátedra.

Guillermo Marco Remón and Julio Gonzalo. 2021. Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural*, 66:77–87.

Adiel Mittmann. 2016. *Escansão automática de versos em português*. PhD Thesis, Universidade Federal de Santa Catarina.

Ben Nagy, Artjoms Šeļa, Mirella De Sisto, and Petr Plecháč. 2025. Metronome: tracing variation in poetic meters via local sequence alignment. *Computational Humanities Research*, 1.

Borja Navarro-Colorado. 2018. A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

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

Xosé Luís Regueira Fernández. 2010. *Dicionario de pronuncia da lingua galega*. Real Academia Galega; Instituto da Lingua Galega, A Coruña, [Santiago de Compostela].

Claudio Rodríguez Fer. 1991. *Arte literaria*. Xerais, Vigo.

Pablo Ruiz Fabo, Montse Cuadros, and Thierry Etchegoyhen. 2014. Lexical Normalization of Spanish Tweets with Rule-Based Components and Language Models. *Procesamiento del Lenguaje Natural*, 52:45–52.

Antón Santamarina, Ernesto González Seoane, and María Álvarez de la Granja. 2026. Tesouro informatizado da lingua galega (Version 4.1). Instituto da Lingua Galega.

Fernando Sanz-Lázaro. 2024. libEscansión: A Recursive Precedence Approach to Metrical Scansion. *Digital Humanities Quarterly*, 18(3).

Andre Valenca and Filipe Calegario. 2025. Experimenting with Large Language Models for Poetic Scansion in Portuguese: A Case Study on Metric and Rhythmic Structuring. In *Proceedings of ICCC, the 16th international conference on computational creativity*, Campinas, Brasil. Association for Computational Creativity.

A Syllable Count Conventions

Contagem grave* vs. *aguda

There are two main systems to compute a line's meter in Romance metrics (Chociay, 1974, 11-13). The first (***contagem aguda*** or *contagem francesa*) ignores syllables after the last stressed one, and the meter (in the sense of syllable count) is defined by that syllable: A line with a final stress on the 7th syllable is *heptasyllable*. This system is commonly used in Portuguese and French metrics.

The second, ***contagem grave*** *ou espanhola*, used in this paper, counts syllables after the last stressed one towards the meter: A line with a final stress on the 7th syllable is considered as *octosyllable*. The stress pattern of the line's and hemistich's last word also plays a role (see rules in Section 3). This system is used in Spanish and Italian metrics; it was used for Portuguese in the past, becoming less popular since the late 18th century (Chociay, 1974, 12).

The choice of ***contagem grave*** determines how to read Figure 1 in the paper. The **column for 11** represents a meter with 10 metrical syllables (usually known as ***decassílabo*** in Portuguese, but *endecasílabo* in Spanish). Likewise, the **column for 8** corresponds to ***heptassílabo*** in Portuguese, but *octosílabo* in Spanish. **Alexandrines** are represented in the **column for 14** (12 syllables under *contagem aguda*, which correspond to 14 under *contagem grave*). We used *contagem grave* because our sources for Galician poetry use it, see Carballo Calero (1981); Rodríguez Fer (1991).

B Manual annotation example

The table presents a simplified example of the manually annotated corpus, using a format inspired by Valenca and Calegario (2025). Corpus metadata are omitted. The example shows how lexical and metrical syllabification need not match, given metaplasms.

Line Text	Lexical Syllables	Metrical Syllables	Stress Pattern	Meter
Hoxe o meu eido	*Ho- / **xe** / **o** / *meu / *ei- / do	*Ho- / **xe o** / *meu / *ei- / do	1 3 4	5
que onte blanqueaba	**que** / ***on-** / te / blan- / **que-** / ***a-** / ba	**que *on-** / te / blan- / **que- *a-** / ba	1 4	5

Table 5: Simplified example of two manually annotated lines. Stars indicate stresses. Metaplasms are bolded: In the first line, synalepha applies. In the second one, both synalepha (across words) and syneresis (within a word) apply.

C Preprocessing Details

C.1 Edit distances and n-gram language model

Correction candidates were accepted up to a distance of 0.5 in our weighted edit distance scheme. Largely, this amounts to only accepting edits that affect an accent mark or homophonic and silent consonants; this is conservative but is meant to prevent modifying the text in ways that would hurt scansion by altering syllable count. To complement edit-distance-based corrections, preprocessing also uses regexes applying in very specific contexts.

For the 5-gram language model trained with KenLM, pruning was set as `--prune 0 1 1 2 3`, removing singleton 2-grams and 3-grams, 4-grams with 2 or less occurrences and 5-grams with 3 or less occurrences.

The system repository can be consulted for more details: `https://github.com/compellit/gama-sym`.

C.2 Preprocessing Evaluation

The examples below illustrate how preprocessing had an impact on scansion and how it was evaluated, i.e. which types of modifications made by preprocessing were considered as correct and as errors.

The table also compares preprocessing output with current orthography as per the ILG/RAG norm. The goal of preprocessing is not to match ILG/RAG, but to make metrically relevant edits. If it makes an edit that goes against ILG/RAG but helps scansion find the right stress pattern, the modification is counted as correct (see "Prepro Eval" column). Conversely, failing to make an edit that would help scansion is counted as an error (by omission) even when the output matches ILG/RAG.

Preprocessing made 293 edits in the 800-line test corpus, but many were of a "neutral" type, like (3) in the table, making corrections that do not affect scansion, as they do not affect syllable count or stress.

#	Line Text	Scansion Impact	Prepro Eval	Comment
1	input: renovese a vida prepro: renóvese a vida ILG/RAG: renóvese a vida	positive	correct	Positive impact on scansion since marks antepenultimate stress explicitly, helping obtain the correct scansion by preventing the word from being considered as having penult stress by default.
2	input: Eso que agora decias prepro: Eso que agora decías ILG/RAG: Iso que agora dicías	positive	correct	Preprocessing does not match ILG/RAG norm, but it adds the stress mark on *í*, helping scansion find the right stress pattern and meter. Other differences with ILG/RAG do not affect scansion.
3	input: N'esta vida de pasaxen prepro: Nesta vida de pasaxen ILG/RAG: Nesta vida de pasaxe	neutral	correct	Removing the apostrophe does not affect stress or syllable count. The edit is not needed, but it does not hurt scansion either.
4	input: ¡Ai! si, que a morrer van prepro: ¡Ai! si, que a morrer van correction: ¡Ai! sí, que a morrer van ILG/RAG: Ai! si, que a morrer van	negative	error	In current norm, the affirmative adverb *si*, like most monosyllables, does not bear an accent mark. However, preprocessing is intended to make tonicity explicit where it would help scansion, and here fails to do so, so it is a false negative counted as an error.

Table 6: Details about preprocessing and its evaluation.

D Web interface

A screenshot of the results view on the interface is below. The interface is publicly deployed at the URL indicated in the system repository: https://github.com/compellit/gama-sym.

The interface is intended to be user-friendly so that it can be used by the general public. A user can paste a poem or upload a zip with poems, optionally including metadata. Scansion results (stress patterns and syllable counts for each line) will be returned. The stress pattern without extrarhythmic stresses (that fall out of the main rhythmic trend in the line), calculated by the Jumper library, is also given. The results of preprocessing and scansion can be downloaded in delimited format. The interface was developed with the Django framework.[2]

Metadatos

Nome do corpus: Follatos **Autor·a:** Filomena Dato **Data:** 1891
Título: Defensa d'as mulleres - IV **Subtítulo:** —

Escansión

Amosar preprocesamento

Amosar 100 rexistros Buscar:

#	Texto orixinal	Silabas métricas	Silabas acentuadas	Sen extra-ritmicas
1.	Os que decís, qu' a muller	8	4 7	4 7
2.	non ten a cabeza feita	8	1 2 5 7	2 5 7
3.	pra soster unha coroa	8	3 4 7	3 7
4.	e que non sirve pra reina.	8	3 4 7	3 7
5.	É qu' esquencedes cicais	8	1 4 7	1 4 7
6.	que tod' a hestoria está chea	8	2 4 6 7	2 4 7
7.	de nomes qu' o brilo teñen	8	2 5 7	2 5 7
8.	de luminosas estrelas.	8	4 7	4 7
9.	Semíramis i-Artemisa:	8	2 7	2 7
10.	duas Aspasias y-a nena	8	1 4 7	1 4 7
11.	File, filla d' Antipatro,	8	1 3 7	1 3 7
12.	con quen iste s' aconsella;	8	2 3 7	2 7
13.	Livia que domin' á Augusto	8	2 5 7	2 5 7
14.	y-o mundo co-él goberna:	8	2 5 7	2 5 7

Figure 2: Screenshot of the results view of the scansion user interface.

[2] https://www.djangoproject.com/. We used version 5.2.4.

E Results for a complete poem

The table below shows the system's scansion results for a complete poem, Section IV in Filomena Dato's 1891 poem "Defensa d'as mulleres".

Meter (syllable count) was always correctly detected. The stress pattern is incorrect in lines 12 and 13; the column indicating stress pattern correctness was added manually.

Line ID	Line Text	Preprocessing	Stress Pattern	Correct?	Meter
1	Os que decís, qu' a muller	Os que decís, que a muller	4 7	✓	8
2	non ten a cabeza feita	non ten a cabeza feita	1 2 5 7	✓	8
3	pra soster unha coroa	pra soster unha coroa	3 4 7	✓	8
4	e que non sirve pra reina.	e que non sirve pra reina.	3 4 7	✓	8
5	É qu' esquencedes cicais	É que esquencedes cicáis	1 4 7	✓	8
6	que tod' a hestoria está chea	que toda a hestoria está chea	2 4 6 7	✓	8
7	de nomes qu' o brilo teñen	de nomes que o brilo teñen	2 5 7	✓	8
8	de luminosas estrelas.	de luminosas estrelas.	4 7	✓	8
9	Semíramis i-Artemisa:	Semíramis iArtemisa:	2 7	✓	8
10	duas Aspasias y-a nena	dúas Aspasias e a nena	1 4 7	✓	8
11	File, filla d' Antipatro,	File, filla de Antipatro,	1 3 7	✓	8
12	con quen iste s' aconsella;	con quen iste se aconsella;	2 3 7	✗	8
13	Livia que domin' á Augusto	liviá que domina a Augusto	2 5 7	✗	8
14	y-o mundo co-él goberna:	e o mundo coél goberna:	2 5 7	✓	8
15	Agripina qu' ô seu fillo	Agripina que o séu fillo	3 6 7	✓	8
16	de Roma o imperio lle dera:	de Roma o imperio lle dera:	2 4 7	✓	8
17	a ilustrada Amalasunta	a ilustrada Amalasunta	3 7	✓	8
18	qu' entende todal-as lengoas:	que entende tódalas lengoas:	2 4 7	✓	8
19	a gran reina anque croel	a gran reina anque croel	2 3 7	✓	8
20	Sabela d' Ingalaterra:	Sabela de Ingalaterra:	2 7	✓	8
21	a Catelina de Médeces	a Catelina de Médeces	4 7	✓	8
22	y-a Católeca Sabela,	e a Católeca Sabela,	3 7	✓	8

Table 7: Scansion results for Section IV of poem "Defensa d'as mulleres" (1891), by Filomena Dato. The meter was always correctly detected. We manually added a column to indicate if the stress pattern is correct.