Automatic Metrical Scansion of Poetry in a Low-Resource Setting

Pablo Ruiz Fabo[1,2], Anxo Alonso[1], Pablo Rodríguez Fernández[1], Paulo Gamallo[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
[2]Université de Strasbourg – LiLPa UR 1339
{pablo.ruiz.fabo,anxo.alonso.perez,pablorodriguez.fernandez,pablo.gamallo}@usc.gal

Abstract

We present the first neural systems for automatic metrical scansion of poetry in Galician, a Romance language close to Portuguese and Spanish. The task is threefold: First, identifying metrical syllables based on lexical ones; both syllable series may differ given metrical licenses modifying a line's syllable structure to enable stress-related rhythms. Second, identifying stress patterns, and third identifying the metrical syllable count, based on stressed positions. We manually annotated a corpus of 4,287 examples, a first in Galician, and fine-tuned an 8B-parameter LLM specialized in Galician and Portuguese, and two encoder–decoder models: ByT5, a token-free byte-to-byte model, and the multilingual mT5, which includes Galician. We also tested our recent symbolic scansion system. Several fine-tuning setups reached exact per-line accuracy above 90% on our test-set at all three scansion subtasks, using orthographic syllables with explicit stress marks as input. Encoder–decoders performed better than the LLM. The token-free ByT5 was best, particularly when adding the two surrounding lines to the input. The symbolic system (89.9% acc.) managed rare metaplasms infrequent in training data better than the neural ones, and the approaches can be seen as complementary.

Keywords: automatic metrical scansion, Galician poetry, transformers and LLM

1. Introduction

The automatic metrical analysis of poetry, or scansion, poses some challenges for NLP. The task requires identifying patterns such as the alternation between stressed and unstressed syllables. This is not trivial, as syllable boundaries and stress placement can be modified in poetry for rhythmic effects. Scansion automation can assist the development of large metrically annotated corpora, useful for comparing versification traditions and to help understand the distribution of metrical patterns across them. The value of such corpora was recently illustrated in Nagy et al. (2025), who performed a computational modeling of verse evolution in classical Latin, in the European Renaissance and in 19th-century Europe.

Both symbolic and neural approaches have been used for automatic scansion. Recently, large language models (LLMs) have started being used, particularly via fine-tuning and prompting proprietary models accessed via cloud platforms (Valença and Calegario, 2025; Kranti and Vajjala, 2025). Such work showed the feasibility of the task for LLMs. However, the reproducibility of cloud-based models is limited. Their long-term availability is not guaranteed and API changes may impact results. Besides, for large scale annotation, the cost may become unaffordable. A problem of the most capable LLMs, even those that can be deployed locally, is the computing demands posed by their large size.

Taking such limitations into account, we compare automatic scansion with models that can be run with the limited resources commonly available in Humanities teams: An 8B-parameter LLM (decoder), several smaller encoder–decoder models (0.3M parameters), and a symbolic baseline.

Some of the earlier studies on scansion-model training had access to generous preexisting training corpora. Here we focus rather on the case where no prior training data exists, and we need to create a first neural system from scratch. We work with Galician, a Romance language close to Portuguese and Spanish, co-official in the Spanish region of Galicia, and recognized in the European Charter for Regional or Minority Languages (Council of Europe, 1992). In the last decade, NLP projects have developed ample resources for Galician, covering the basic NLP pipeline (e.g. Gamallo et al., 2018), but also LLMs and specialized benchmarks to evaluate these (Gamallo et al., 2024; Rodríguez et al., 2025b). However, most NLP resources focus on contemporary Galician, whose orthography was normalized in the late 20th century. In the context of a Computational Literary Studies (CLS) project studying modern Galician poetry diachronically, we need to analyze 19th-century text, where orthographic practices vary given lack of a standard, decreasing the performance of available NLP models. In this sense, together with the prior unavailability of annotated data for scansion, the task corresponds to a low-resource scenario.

The paper's contributions are the following:

- First transformers-based systems for scansion in Galician evaluated on unnormalized 19th-century data, publicly available under an open license.[1] As a baseline for evaluation

[1]See Appendix A for model URLs.

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 114–125
11 May 2026.

we use our recent symbolic scansion system (Ruiz Fabo et al., 2026).

- A manually annotated corpus for Galician scansion (lexical and metrical syllables, metrical licenses, syllable count) with 4,287 examples.[2]
- A comparison between several transformers-based models, encoder–decoder and decoder-only, with each other and with the symbolic baseline, pointing out the strengths of each. This can be informative for research on developing a first neural scansion system in other languages: Smaller encoder–decoder models, particularly the token-free ByT5, obtained best results overall, and the symbolic system outperformed neural ones with rare metrical phenomena, harder to learn statistically.

Although very large decoders excel at challenging semantic tasks, scansion relies less on semantic information, making comparisons with smaller and alternative architectures particularly relevant.

The paper is structured as follows: Section 2 outlines the state of the art. Sections 3 and 4 define the task and describe the corpora. The models developed are presented in Section 5, and the results are discussed in Section 6; Section 7 concludes.

2. Related Work

Symbolic approaches, statistical ones based on classical machine learning, and neural approaches are all present in the state of the art.

Systems we consider **symbolic** may include, besides rule-based methods, a statistical component that is not driven by machine learning (e.g. to disambiguate among scansion alternatives). For Portuguese, one of the traditions closest to Galician, an early rule-based system was created by Araújo and Mamede (2002), and Mittmann (2016) created *Aoidos*, a rule-based system with a thorough set of 159 rules, which achieves 97.5 exact line accuracy (higher with some test corpora), as tested on canonical authors, mostly Brazilian from the 18th-19th cts. In Spanish, Gervás (2000) created an early symbolic system. Navarro-Colorado's (2018) system specializes on hendecasyllable verse, reaching 95% perfect stress-pattern match per line, as tested on the ADSO 100 corpus, with 1,404 classical sonnets. The system performs automatic syllabification, using parts of speech (PoS) to determine syllable tonicity; rules resolve metrical ambiguities. De la Rosa et al.'s (2020) system (*Rantanplan*), also uses PoS, syllabification and metrical disambiguation heuristics. It achieves 95% exact stress-pattern match per line in ADSO 100, reaching, 65.02% on the more challenging *Carvajal* corpus (Pérez Venegas, 2015), with mixed-meter poems and a much larger metrical variety. *LibEscansión* by Sanz-Lázaro (2024) uses PoS and syllabification based on a phonological transcription, reaching 97.01% exact per-line stress match on ADSO 100. A final symbolic tool for Spanish is *Jumper* (Marco Remón and Gonzalo, 2021). It identifies stress patterns without prior syllabification. It reaches 95% perfect stress-match per line on ADSO 100 and 82% on the harder, mixed-meter Carvajal corpus. It is thus the best available Spanish mixed-meter scansion tool according to this benchmark. Our symbolic baseline for Galician is derived from this tool.

Symbolic systems have also been implemented for languages further removed from our task: For French, Delente and Renault (2015), and Bobenhausen and Hammerich (2015) for German. Several exist for English, to name just two, *Prosodic* (Anttila and Heuser, 2016), a metrical phonology parser (which also has some trainable components), and *ZeuScansion* (Agirrezabal et al., 2016b), based on Finite State Technology.

Statistical approaches, deploying classical machine learning (ML), were used by Estes and Hench (2016), using Conditional Random Fields (CRF) for Middle High German. Agirrezabal et al. (2016a) experimented with sequence labeling (CRF and Hidden Markov Models) and other classical ML models, for English scansion. Statistical methods appear as well in Plechac (2016), for Czech. Barbosa and Barbosa (2025) developed statistical scansion methods targeting Brazilian Northeastern phonology, not addressed in earlier Portuguese systems.

Several **neural systems** have been developed. Among those based on recurrent networks, Agirrezabal et al. (2017) and Agirrezabal (2017) used bidirectional LSTM-CRF for Basque, English and Spanish, in the latter case reaching 90.84% exact per-line accuracy on ADSO 100. Haider (2021) used a Bi-LSTM-CRF to predict syllable stress and other prosodic features, with above 83.1% per-line accuracy in English and 87.7% in German, using ca. 3,500 annotated examples in each language (Haider, 2023, 217). The same implementation was applied to Czech by Klesnilová et al. (2024), training with ca. 59,000 poems from the Corpus of Czech Verse (CCV) (Plecháč and Kolár, 2015), which totals ca. 2.3 million lines (ca. 66,500 poems). Several configurations were tested, e.g. giving the entire poem or single lines as the unit to tag. With the best configuration, results were above 99% exact-line accuracy, improving upon Plechac's (2016) statistical model, which achieved 81.9% exact-line accuracy. Koziev (2025) combined neural and other paradigms for Russian scansion.

[2] See project repository at https://github.com/compellit/gama-trf

Transformers have also been used. De la Rosa et al. (2021) fine-tuned encoders for Spanish scansion, with 8,748 examples. In the best setup (RoBERTa large and 100 epochs), perfect stress match per line was 93.43% on the ADSO 100 corpus. Glaser (2025) trained BERT (encoder) and T5 (encoder–decoder) for the scansion of 18th-century English iambic pentameter using >100K training examples, achieving 96% per-line accuracy.

Still within transformers, **LLMs** were used by Valença and Calegario (2025), who fine-tuned GPT 3.5 for Portuguese scansion, obtaining 88.6% per-line accuracy using 7,200 examples (87.19% if using 3,520). Kranti and Vajjala (2025) performed scansion via prompting, in Telugu. The results suggest that, without fine-tuning, quality is low: 60% accuracy for GPT-5 and 20% for Gemini-2.5-Pro (syllable classification task, Table 2 in their work).

Discussing transformer models for poetry-related tasks, Rosa et al. (2025) point out that, for tasks where manipulating syllables is important, the models' pre-trained tokenizers, which learn a subword vocabulary efficient for modeling, rather than targeting units like syllables, can underperform compared to a syllable-based or character-based tokenization. This informed our choice of a token-free option among our compared models (Section 5).

Some of the systems above showed remarkable accuracy, above 95%. However, in some cases, this involved massive training data (like the Czech neural tagger), or was achieved for a specific period or meter, like the T5 examples or the symbolic hendecasyllable taggers. Observing that extraordinary accuracy was only possible under specific conditions suggests that the task can have difficulties, regardless of the technological paradigm chosen.

3. Task Definition

We defined scansion as articulated into three subtasks, that are interrelated but can be evaluated individually, described below.

3.1. Metrical syllabification

In the context of our fine-tuning experiments, we defined metrical syllabification as obtaining metrical syllables based on lexical ones. Lexical syllables depend on general and language-specific phonological constraints. In Romance metrics, metrical syllables need not match lexical ones, which can be merged or split to allow stressed syllables to fall into specific positions, which helps create rhythmic patterns. For Galician, the main metaplasms are the following (see Table 2 for distribution):

• **Synalepha:** The final syllable of a word ending in a vowel merges with the initial syllable of the following vowel-initial word, forming a single metrical syllable across the word boundary. In Galician it is very frequent and is the default realization for vowel sequences across the word boundary. E.g. *a* and *o* in *Sin mi-rar, fi-x**a_o**s o-llos*.[3]

• **Syneresis:** Within a word, two vowels belonging to separate syllables and not constituting a diphthong are merged into a single syllable. It is relatively rare in Galician metrics. E.g. *e* and *o* in *Dé-ches-me fi-d**eo**s con gre-los*.[3]

• **Dialepha:** Takes place when the last syllable of a word and the first one of the following word could be pronounced as a single syllable, but are pronounced as separate ones. It can be seen as an exceptional absence of synalepha: *í* and *a* in *a-qu**í a**-que-las vei-gas*.[3]

• **Dieresis:** Within a word, a diphthong is split into two syllables, adding a metrical syllable. It is rare in Galician, e.g. *ia* pronounced as two syllables in *do sil-ves-tre_ar-bo-re-do su-b**i-a**n-do*.[3]

3.2. Stress-pattern detection

Stress-pattern detection consists in identifying which of the metrical syllables are stressed. The output can be formalized in several ways, like a syllable tonicity boolean vector, with as many dimensions as syllables in a line, or as a list with the positions of the line's stressed metrical syllables.

In current Galician (using the official ILG/RAG norm), orthographic cues for syllable tonicity are more ambiguous than in Spanish: Stressed interrogative pronouns do not bear an accent mark and are homographic with unstressed relative pronouns and conjunctions. Word-final stressed syllables containing a falling diphthong do not bear an accent. In our 19th-century corpus, the challenge increases because there was no written norm, practices to represent stress vary, and an accent mark can represent vowel aperture or stress.

3.3. Syllable count

Following Spanish-style counting practices, which apply to Galician (cf. Carballo Calero, 1981; Fer, 1991), syllable count is affected by the position of the last stressed metrical syllable in the line. If stressed, one syllable is added to the count. If the last metrical syllable in the line is the antepenultimate, a syllable is deducted. When the line is divided into hemistichs, the same rules apply at the end of the first hemistich. For instance, an alexandrine (14 metrical syllables) is divided into hemistichs of 7 metrical syllables. These can

[3]English glosses: 1. Without looking, she fixes her eyes (R. de Castro). 2. You gave me noodles with rapini (J. M. Posada). 3. Here those plains (X. M. Cabada). 4. Whistling in the wild grove (F. Vaamonde).

have only 6 lexical syllables, if their 6th syllable is stressed.[4]

3.4. Challenges and applications

As a first challenge, lines can have metrical ambiguities. It may be possible to apply different sets of metaplasms, which would result in different stress patterns and syllable counts. For humans, lines in the context, particularly metrically unambiguous ones, can help decide how to scan a given line: what syllable count to target, which metaplasms to choose and for which syllables. This all poses challenges for an automatic scansion system, which must resolve ambiguities identifying possible solutions and excluding unlikely or impossible ones.

In some ambiguous cases, human experts accept more than one scansion, or even disagree as to the correct one. This poses a challenge for automatic evaluation of scansion, as several scholars have commented (recently Cuéllar, 2025, p. 10, Martin, 2025, p. 128). We discuss how this manifested in our corpora in Section 4.

As regards the subtasks' relative importance, the correct detection of stress-patterns is arguably more important than exact syllabification. It encodes metrical prosody more directly and, unlike exact syllable match, it is largely unaffected by mismatches at the character level unrelated to prosody. Besides, syllable count as defined can be derived deterministically from the stress pattern.

In terms of downstream applications, syllable count can provide a coarse overview of the form of a large versified corpus. Stress patterns have richer applications, and have been used to cluster corpora across languages and traditions (Nagy et al., 2025), or as a stylometric signal for authorship studies (e.g. Plecháč, 2021; Cuéllar, 2025).

Concerning alternative task definitions, it would be possible to define metrical syllabification as taking orthographic words as input instead of lexical syllables, as in Valença and Calegario (2025). Similarly to their experiment, GPT-4.1 fine-tuned on our corpus succeeded, with ca. 75% accuracy. Still, a smaller locally deployed decoder (Sec. 5) did not manage the task so defined. We thus use lexical syllables as input, as do most studies reviewed (Sec. 2). It would also be possible to use phoneme-based syllabification instead of orthography-based (as Klesnilová et al., 2024; Sanz-Lázaro, 2024 among others), a future work possibility.

[4]For the names of meters (in the sense of line types based on their syllable count) we follow the Spanish/Italian convention (termed *contagem grave ou espanhola* in Chociay, 1974) rather than French convention (*contagem aguda ou francesa*). The latter is currently more common in Portuguese versification studies but not in Galician based on our sources. Accordingly, in our descriptions, a hendecasyllable is a line with the last stress on the 10th metrical syllable, called *decassílabo* in Portuguese convention. Likewise, an alexandrine has 14 metrical syllables in our descriptions rather than the 12 it has under French/Portuguese-style syllable counts.

4. Corpora

We manually annotated a corpus of 19th-century poetry in Galician, totalling 4,287 lines from 98 poems by 29 authors. We used 3,487 lines for training (among which 697 lines for validation), and 800 lines as a held-out test-set. An example of the corpus format is given in Table 1.

The corpus contains a variety of meters representative of metrical poetry in modern Galician (see Fig. 1). About 34% of lines belong to mixed-meter poems, where scansion is harder because at least two meters (in the sense of syllable counts) appear (rarely more than three). The original orthography was largely preserved, but we performed a lightweight typographical and orthographic normalization which did not alter any metrically relevant features (see 5.1). The corpus covers authors from the mid 19th century and the Galician Renaissance (*Rexurdimento*) in the second half, when sustained literary production in the language reemerged, after centuries of decreased activity.

As a departure point for manual annotation, we carried out an automatic pre-annotation of syllabification, stress and metaplasms, thanks to heuristics that combine two sources: First, the output of our recent symbolic scansion system, which identifies stress patterns, syllable count and metaplasms (see 5.1). Second, the output of an automatic lexical syllabification tool we developed.

All these automatic pre-annotations were corrected manually, yielding a human-validated corpus annotated with lexical and metrical syllabification, stress patterns, metaplasms and metrical syllable counts for each line. To promote data quality, some error patterns that can be common in manual annotation were identified algorithmically and corrected manually: misalignments between lexical and metrical syllables given differences at the segment-level (rather than in stress), or impossible stress patterns, where the series of stressed positions is not compatible with the syllable count.

The entire corpus was annotated by the first author, and the test corpus was annotated manually by two of the authors. We computed inter-annotator agreement (IAA) between both for all subtasks defined in Section 3. IAA was 97.63% for exact metrical syllabification match, 98.63% for stress-pattern match, and 99.63% for syllable counts. We consider agreement substantial and in line with values reported in the literature. Navarro-Colorado (2018) report 96% IAA with three annotators in their 100-sonnet test corpus (fixed meter).

Line Text	Lexical Syllables	Metrical Syllables	Stress Pattern	Syllable Count
co eco das harpas	co / *e- / co / das / *har- / pas	c**o / *e**- / co / das / *har- / pas	2 5	6
renóvese a vida	re- / *nó- / ve- / se / a / *vi- / da	re- / *nó- / ve- / s**e a** / *vi- / da	2 5	6
Hoxe o meu eido	*Ho- / xe / o / *meu / *ei- / do	*Ho- / x**e o** / *meu / *ei- / do	1 3 4	5
que onte blanqueaba	que / *on- / te / blan- / que- / *a- / ba	qu**e *o**n- / te / blan- / qu**e- *a**- / ba	1 4	5

Table 1: Two groups of two manually annotated lines. A slash delimits syllables, stars indicate stress. Metaplasms are bolded: Dialepha applies in first line and synalepha in the second.
Note: First group: Lines by F. M. de la Iglesia (1880). Gloss: *with the echo of the harps / may life be renewed*. Second group: Lines by E. Martelo Paumán (1893). Gloss: *today my field / which yesterday was whitening.*

Metaplasm	train		test	
	N	%	N	%
Synalepha	1661	47.63	397	49.62
Complex (>2 syllables)	50	1.43	12	0.75
Syneresis	121	3.47	36	4.50
Dialepha	107	3.07	32	4.00
Dieresis	25	0.72	6	0.75

Table 2: Number and percentage of lines with metaplasms in the corpus splits.

As said in 3.4, there can be difficulties in establishing the reference scansion for some lines. In our test corpus, of the 11 lines where both annotators' stress patterns did not match, none of the cases was due to conceptual disagreement. Two cases would match if we consider alternative patterns proposed by annotators, one case was due to a missing criterion in the annotation guidelines, and 8 cases were due to errors by one of the two annotators (which were corrected after computing agreement). Regarding alternative patterns, these were provided for 18 lines in total, suggesting that a clear solution existed for humans in most cases.

Concerning corpus metadata, poems' titles, authors, and year of publication were recorded.

5. Experiments

This section presents the fine-tuning experiments, including data preprocessing workflow, baselines and experimental conditions tested.

5.1. Preprocessing and Lexical Syllabification

We use lexical syllabification with stress marks as the input for fine-tuning (Table 1). Syllable segmentation is largely deterministic in Galician, with some exceptions like cases of full-vowel vs. glide variation (Freixeiro Mato, 1998; Regueira Fernández, 2010), that can be managed with lexical resources. Syllable tonicity detection, however, presents some challenges because there is ambiguity in orthographic cues, more so in the unnormalized 19th-century variants from our corpus (see 3.2).

To tackle the task, we created a rule-based syllabification tool. This also implements a lightweight preprocessing, aimed at resolving syllable tonicity ambiguity in Galician (historical) orthography, helping detect stressed syllables by restoring stress marks where they are absent in 19th-century text, or otherwise assigning syllable stress to ambiguous syllables based on parts-of-speech (PoS) and context information. The workflow is fully described in Ruiz Fabo et al. (2026). It relies on candidates generated via regex and weighted edit distances, ranked in context with an n-gram language model. The in-vocabulary (IV) lexicon is based on resources from the LinguaKit and Apertium libraries (Gamallo et al., 2018; Forcada et al., 2011) and the 5-gram language model was trained on 126 million tokens in Galician from *CorpusNÓS* (De-Dios-Flores et al., 2024), with KenLM (Heafield, 2011). Although the approach is based on classical techniques, it allowed good results without the need to develop any training data and with few computational resources.

Preprocessing errors in lexical stress detection affected 24 of 800 lines in the test-set (3%). The remaining 2 errors were irrelevant for metrics, not affecting stress placement or syllable count.

The goal of the experiments was to assess the models' performance at learning metrical syllabification based on lexical syllables, rather than evaluating lexical syllabification based on orthographic words. The latter was implemented as a preprocessing step. To isolate lexical-to-metrical syllabification, we corrected preprocessing errors in the training and test data prior to fine-tuning.

5.2. Baselines

The literature (Section 2) suggests that prompting alone is not sufficient for scansion. To assess this on our language and corpus, we used prompting with GPT-5.2 and GPT-5 mini as a first baseline, in zero-shot and few-shot modes (20 examples).

The literature shows that symbolic systems can achieve high scansion quality, in some cases competitive with statistical and neural ones. As a symbolic baseline, we use our recent system (Ruiz Fabo et al., 2026), an adaptation to Galician of Jumper (Marco Remón and Gonzalo, 2021), which performs stress pattern detection without syllabification in Spanish poetry. We adapted its lexical resources to work with Galician. It generates scansion candidates (syllable tonicity vectors) based on vowel sequences that could be merged or split. The candidates are ranked based on their similarity to well-attested patterns in a stress pattern inventory, also taking into account the candidates selected for lines in a context window. The algorithm requires orthographic input with unambiguous tonicity, for which we used the preprocessing in 5.1. We corrected preprocessing errors before evaluation to assess scansion independently of preprocessing accuracy. Sec. 6 reports results before and after corrections. The symbolic system does not perform exactly the same task as the fine-tuned systems: it operates on orthographic words to infer stress patterns and syllable count directly, without syllabified input. Nevertheless, it provides a useful baseline, allowing us to gauge to what extent neural models implicitly acquire representations relevant to solve a task that the symbolic system encodes through explicit expert knowledge.

5.3. Conditions: Varying Input Context

We structured data for fine-tuning according to two different conditions. In ***single-line***, the model input and output consist in lexical and metrical syllables for a single verse-line respectively. In ***context-lines***, the input contains lexical syllables for the previous, current, and following lines (within the same poem), with delimiters to mark structure clearly. The output contains the metrical syllables for the current line only.

The two conditions test the influence of added context in fine-tuning. Humans use context lines to decide on the parse for a metrically ambiguous line; we thus tested whether context was also beneficial for the models.

5.4. Neural Model Fine-Tuning

Fine-tuned LLMs (decoder only) can succeed at scansion, as shown by Valença and Calegario (2025) with Portuguese and GPT-3.5. We wanted an LLM that can be deployed locally with limited resources (e.g. a Colab session, a usual tool in humanities teams in our experience). We chose 8B-parameter Nos-PT/Llama-Carvalho-PT-GL, specialized in Galician and Portuguese, created via continual pretraining of Llama-3.1-8B with 232M Galician and 250M Portuguese tokens, along the lines of methods in Rodríguez et al. (2025a), also using English and Spanish text to prevent catastrophic forgetting.

Scansion can be seen as transforming the input sequence (lexical syllables in our case) into an output one (metrical syllables). The task can involve removing or adding tokens, to apply metaplasms which erase or insert syllable boundaries. This is a natural fit for an encoder–decoder architecture, fine-tuned for sequence-to-sequence generation. Glaser (2025) recently fine-tuned T5 for scansion with success, so we chose T5-variants for our experiments.

Our first encoder–decoder base model was mT5-small (Xue et al., 2021). This is a multilingual T5 variant which includes Galician among its pretraining languages. Our training data is about 4,000 examples, the small version is appropriate for such data volume.

Our second encoder decoder model was ByT5-small (Xue et al., 2022), a token-free T5 variant. This model does not rely on a sub-word tokenizer, learning to perform byte-to-byte generation. The literature has shown that pre-trained tokenizers can perform worse than syllable- or character-based tokenization at manipulating poetic form (Rosa et al., 2025). By testing a token-free model we wanted to see if our setup also shows benefits from character-to-character learning.

We applied supervised fine-tuning, 5 runs per model with the same set of seeds. T5 variants were trained for 30 epochs, selecting the best checkpoint based on exact metrical syllabification match per-line, which is the most demanding one of our evaluation subtasks and also improves results at the others. Effective batch-size was 16 and learning rate 5×10^{-5}. The decoder was fine-tuned with LoRA (Hu et al., 2022) loaded in 4-bit precision, for 3 epochs, with an effective batch size of 8 and a learning rate 10^{-5}. Other hyperparameters are in the project repository.[5] We used `transformers` (Wolf et al., 2020) and, for the decoder, Unsloth (Han-Chen et al., 2025).

6. Results and Discussion

We evaluated all subtasks defined in Section 3. For each, our metrics are based on exact match per-line. We use the following abbreviations to discuss results: *sym*, *spm* and *scm* refer to exact-match per line in metrical syllabification, stress patterns and syllable counts respectively. We report results for the baselines and neural models. For the latter, we performed inference using greedy decoding and we report results averaged over 5 seeds.

The prompting baselines do not show exploitable results, as Valença and Calegario (2025) reported.

[5] https://github.com/compellit/gama-trf

	cd	sym	spm	scm
Baselines				
gpt5m-zs	sg	47.5	52.38	58.13
gpt52-fs	cx	61.2	68.12	82.5
symbolic		dna	89.88	**97.38**
Fine-tuning				
carvalho	sg	86.43 (.41)	87.05 (.53)	87.46 (.37)
	cx	87.75 (.78)	88.60 (.77)	88.80 (.87)
mt5-sm	sg	89.48 (.14)	90.46 (.14)	90.38 (.13)
	cx	90.37 (.51)	91.12 (.46)	91.02 (.53)
byt5-sm	sg	*92.02* (.44)	*92.42* (.53)	92.42 (.53)
	cx	**92.95** (.19)	**93.50** (.20)	*93.48* (.18)

Table 3: Exact-match per-line accuracy (%) and its *std* (5 runs) in metrical syllabification (*sym*), stress patterns (*spm*) and syllable counts (*scm*), in single-line (*sg*) and context (*cx*) conditions (*cd*).

Table 3 shows the worst (GPT-5 mini, zero shot *single-line*) and best results (GPT-5.2, few shot *context-lines*), which only reached 61.2 *sym*.

The symbolic baseline was very strong. In *spm*, both encoder–decoder models improved upon the symbolic baseline, with ByT5 in the *context-lines* condition achieving the largest margin (3.6 percentage points). The symbolic system, however, scored 0.72 points higher than the best decoder in *spm*. In *scm*, the symbolic system achieved the highest score, 3.9 points above the best fine-tuned model. When evaluating the symbolic system without prior correction of preprocessing errors, *spm* decreased to 88.12%, and *scm* remained unaffected.

Regarding results per neural architecture, encoder–decoders were better than the decoder. ByT5 was best, agreeing with earlier literature on improved performance of token-free models at formal poetry-related tasks. Accuracy differences per model within the same experimental conditions were significant, based on (i) a non-parametric bootstrap test per-item and (ii) a sign-flip permutation test treating poems as the unit of analysis. Statistical significance was assessed at $\alpha = 0.05$. We controlled the false discovery rate using the Benjamini–Hochberg step-up procedure; adjustments were applied over the full set of pairwise model and condition comparisons within each metric.

In terms of experimental conditions, for all models, *context* outperformed *single*, suggesting that the previous and following lines are helpful for learning and inference. However, evaluating the differences between conditions with the statistical tests above showed a significant difference only for ByT5. Note that we had tested other ways of adding context, like grouping 2 or 4 lines as input-output pairs, but they did not consistently improve performance across models and could degrade results, likely due to the increased output complexity. Adding context lines to the input while predicting a single-line output proved more effective.

Results in lines with specific metrical difficulties (Table 4) show a contrast between the symbolic and neural paradigm. As the distributions in Table 2 showed, the test corpus contains almost 50% of lines with the frequent synalepha license (*slp* in Table 4), and 74 lines with less frequent metaplasms: syneresis (*srs*), dialepha (*dlp*), dieresis (*die*). The best fine-tuned model outperforms the symbolic system by 3.6 points in stress-pattern match, but examining results for lines containing specific metaplasm types shows that improvement takes place mainly in lines with synalepha, whether it involves 2 syllables (*slp*) or more (*slpc*). For the infrequent metaplasms, the symbolic model outperforms the neural ones, more clearly with dialepha and dieresis, likely because they are the least represented in training data (Table 2), posing challenges for statistical learning. Results for syneresis (erasing a word-internal syllable boundary between vowels) were similar in the symbolic and the encoder–decoder, suggesting the latter's capacity to model task-specific transformations, or perhaps reflecting that erasing the boundary is easier with byte-to-byte learning than with subword tokens.

Per-meter results (Figure 1) show that the average per-line accuracy decreases as syllable count increases, likely due to the higher complexity of predicting an exact per-line pattern when the number of positions to consider grows. Results for the symbolic system also illustrate the tension between symbolic modeling and data-driven learning: The symbolic system outperformed neural models with alexandrines (14 syllables, 2 hemistichs) by 2 lines. Error analysis shows that the symbolic system correctly updated syllable counts if the first hemistich's last stressed position required it (see 3.3) also avoiding synalepha across the hemistich boundary. The rarity of these configurations makes them harder for learning-based systems to capture.

In summary, the fine-tuned systems, particularly ByT5, managed frequent metrical licenses robustly. Rare ones (particularly dieresis) were managed more adeptly by the symbolic model. Both approaches have practical value to assist in large-scale metrical annotation, followed by human validation. It would be possible to devise heuristics to select cases likely to require revision. The symbolic system marks metaplasms in its output; this signal could be used to flag lines that potentially have rare metaplasms for human verification. Although LLMs have become dominant for challenging tasks, from our results it is unclear if they are the most

Model	cd	**slp** · n=397		**slpc** · n=12		**srs** · n=36		**dlp** · n=32		**die** · n=6	
		spm	N	spm	N	spm	N	spm	N	spm	N
symbolic		86.9	345	83.3	10	**75**	**27**	**78.1**	**25**	**50**	**3**
byt5	sg	92.1	365.8	**100**	**12**	67.8	24.4	28.1	9	0	0
	cx	**93.1**	**369.6**			61.7	22.2	36.2	11.6		
mt5	sg	88.8	352.6	**100**	**12**	60	21.6	35	11.2	0	0
	cx	90.5	359.2			49.4	17.8	34.4	11		
carvalho	sg	80.7	320.2	71.7	8.6	38.3	13.8	58.8	18.8	0	0
	cx	85	337.4	**100**	**12**	40.6	14.6	42.5	13.6		

Table 4: Exact-match per-line accuracy in stress-patterns (*spm*) in lines with metaplasms: Synalepha (*slp*; *slpc* if >2 syllables), syneresis (*srs*), dialepha (*dlp*), dieresis (*die*) per model, in context (*cx*) or single (*sg*) conditions. The total number of metaplasms of each type follows *n*= in the header. For fine-tuned models, results are averaged over 5 runs, *N* is the average of correct lines.

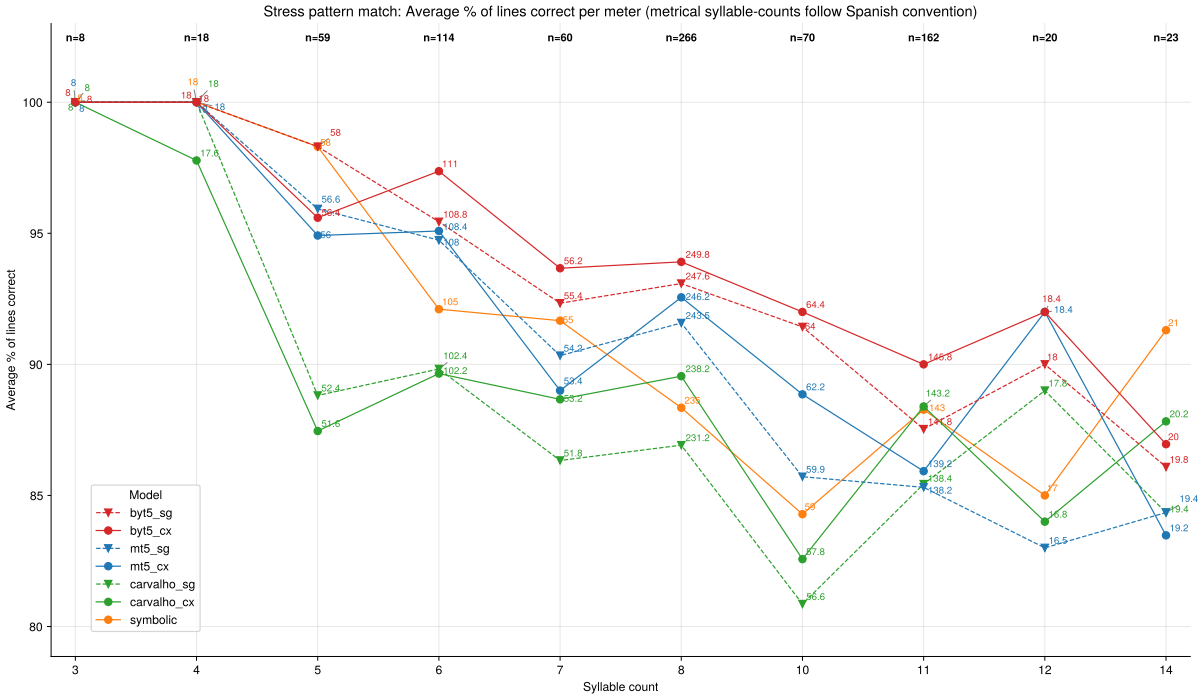

Figure 1: Stress pattern match per-meter (i.e. according to line syllable counts), using Spanish-style conventions (e.g. hendecasyllables are considered to have 11 metrical syllables and alexandrines 14).

efficient neural architecture for scansion, as pre-trained encoder–decoder models were consistently more successful at a fraction of the parameter size.

7. Conclusion and Outlook

We presented experiments on fine-tuning encoder–decoder models (ByT5, mT5) and an LLM (8B-parameter decoder-only Llama-Carvalho-PT-GL) for metrical scansion of poetry in Galician, including three subtasks: metrical syllabification, stress pattern detection and syllable count, evaluated with exact per-line accuracy. We also tested the influence of additional input context vs. single-line input. We created the first manually annotated training corpus for scansion in Galician for public release under an open license, including metaplasms or metrical license annotations. Using 3,487 examples for training and the remaining 800 for testing, several systems attained >90% on all three subtasks. The best model was the token-free ByT5, suggesting the usefulness of byte-to-byte learning for tasks involving units like syllables, which need not correspond to pre-trained tokenizers' subword tokens. Encoder–decoder models were better than the decoder-only. This has practical implications,

since the former are smaller and proved faster in inference, for which they also demanded less VRAM than the decoder. The influence of additional input context was positive, but only when asking to predict metrical syllables for a single line given its two surrounding lines of lexical syllables in the input. Asking to predict multiple lines at a time underperformed compared to single-line prediction with our data; this behaviour may differ with a larger training set. There was a contrast between the symbolic and the fine-tuned systems. The latter obtained better results overall. However, on lines that have rare metaplasms (particularly dialepha and dieresis), most infrequent in training data, the symbolic system did better. Both approaches can thus be seen as complementary. In a scansion workflow, the robust analysis of more general patterns by the fine-tuned models could be combined with symbolic input to identify lines likely to involve exceptional cases and require human inspection.

Future work may explore several directions. Regarding the results with rare metaplasms, oversampling or other data augmentation methods may be attempted. In this work we operated on orthographic input. It would be possible to compare this with phonetic-transcription-based training, as was done in some related works, and might contribute to generalization as it would neutralize orthographic differences. Another future topic might be assessing cross-language transfer in closely related traditions, such as Galician, Portuguese and Spanish.

Acknowledgements

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/).

Limitations

One constraint of the study is the size of the test set (800 items), which reflects the need to create annotations from scratch while maintaining a sufficiently large training set. This may limit the robustness of the findings and a larger test set would strengthen the conclusions.

In terms of evaluation, as noted in Section 3.4, certain verse lines admit more than one plausible analysis. Although alternative annotations were recorded for the small subset of such cases, only the first of the recorded annotations was used for system evaluation. This may slightly underestimate system performance, as outputs corresponding to valid alternative analyses are counted as incorrect.

Ethics Statement

We acknowledge a gender imbalance in the corpus, which predominantly reflects male authorship. While one of the central figures of 19th-century Galician literature and the *Rexurdimento*, Rosalía de Castro, is well represented, other women writers of the period are less visible in available digital resources and, consequently, in our dataset. Besides de Castro, the only additional woman author identified in our sources was Filomena Dato Muruais. Expanding the representation of women writers would require further archival research and the development of additional digital materials.

The study required the use of GPUs, which are associated with higher energy consumption than more modest computational setups. We compared resource-efficient models (e.g., 0.3M-parameter encoder–decoders) with an 8B-parameter LLM, as well as a symbolic system with minimal computational requirements. Since the encoder–decoder and symbolic systems achieved better results than the LLM, additional experimentation with larger models did not appear warranted given their higher computational cost.

Code and Data Availability

The project repository is at `https://github.com/compellit/gama-trf`.

8. Bibliographical References

Manex Agirrezabal. 2017. *Automatic Scansion of Poetry*. Ph.D. thesis, University of the Basque Country.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016a. Machine Learning for Metrical Analysis of English Poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages

772–781, Osaka, Japan. The COLING 2016 Organizing Committee.

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. Incoma Ltd. Shoumen, Bulgaria.

Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016b. ZeuScansion: A Tool for Scansion of English Poetry. *Journal of Language Modelling*, 4(1):3–28.

Arto Anttila and Ryan Heuser. 2016. Phonological and Metrical Variation across Genres. *Proceedings of the Annual Meetings on Phonology*, 3.

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

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.

Klemens Bobenhausen and Benjamin Hammerich. 2015. Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme *Metricalizer2*. *Langages*, 199(3):67.

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

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

Council of Europe. 1992. European charter for regional or minority languages. European Treaty Series No. 148.

Álvaro Cuéllar. 2025. From Atoms to Waves: Rhythmic Stylometry for Authorship Studies of Early Modern Spanish Theatre. *Janus. Estudios sobre el Siglo de Oro*, (14).

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 transfomer-based language models. *Neural Computing and Applications*.

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*, pages 83–90.

Eliane Delente and Richard Renault. 2015. Traitement automatique des formes métriques des textes versifiés. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 116–122, Caen, France. ATALA.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Workshop on Computational Linguistics for Literature (CLfL)*, pages 1–8.

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

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.

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

Ben Glaser. 2025. TrochAIc: Metrical Tools for AI Interpretability. *Anthology of Computers and the Humanities*, 3:1438–1453.

Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725, Online. Association for Computational Linguistics.

Thomas Haider. 2023. *Computational Stylistics of Poetry*. PhD Thesis, Universität Stuttgart.

Daniel Han-Chen, Michael Han-Chen, and Unsloth AI. 2025. Unsloth. `https://github.com/unslothai/unsloth`.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kristýna Klesnilová, Karel Klouda, Magda Friedjungová, and Petr Plecháč. 2024. Automatic Poetic Metre Detection for Czech Verse. *Studia Metrica et Poetica*, 11(1):44–61.

Ilya Koziev. 2025. Automated Evaluation of Meter and Rhyme in Russian Generative and Human-Authored Poetry. ArXiv:2502.20931 [cs].

Chalamalasetti Kranti and Sowmya Vajjala. 2025. MetricalARGS: A Taxonomy for Studying Metrical Poetry with LLMs. ArXiv:2510.08188 [cs].

Guillermo Marco Remón and Julio Gonzalo. 2021. Escansión automática de poesía española sin silabación. *Procesamiento del Lenguaje Natural*, 66(0):77–87.

Meredith Martin. 2025. *Poetry's data: Digital humanities and the history of prosody*. Princeton University Press.

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.

Petr Plechac. 2016. Czech Verse Processing System KVĚTA – Phonetic and Metrical Components. *Glottotheory*, 7.

Petr Plecháč. 2021. *Versification and Authorship Attribution*. Karolinum Press.

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

Pablo Rodríguez, Pablo Gamallo, Daniel Santos, Susana Sotelo, Silvia Paniagua, José Ramom Pichel, Pedro Salgueiro, Vítor Nogueira, Paulo Quaresma, Marcos Garcia, and Senén Barro. 2025a. Enhancing large language models for underrepresented varieties: Pretraining strategies in the galician-portuguese diasystem. *Journal of the Brazilian Computer Society*, 31(1):1049–1062.

Pablo Rodríguez, Silvia Paniagua Suárez, Pablo Gamallo, and Susana Sotelo Docio. 2025b. Continued Pretraining and Interpretability-Based Evaluation for Low-Resource Languages: A Galician Case Study. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4622–4637, Vienna, Austria. Association for Computational Linguistics.

Rudolf Rosa, David Mareček, Tomáš Musil, Michal Chudoba, and Jakub Landsperský. 2025. EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 524–542, Albuquerque, USA. Association for Computational Linguistics.

Pablo Ruiz Fabo, Pauline Moreau, and Anxo Alonso Pérez. 2026. Automatic metrical scansion of galician poetry: First results. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese*. (Accepted, to appear).

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

Andre Valença 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.

9. Language Resource References

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.

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.

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.

Petr Plecháč and Robert Kolár. 2015. The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1):107–118.

Dionisio Pérez Venegas. 2015. *El endecasílabo y su combinatoria en "Extravagante jerarquía"*. PhD Thesis, Universidad de Granada. Appendix (metrical annotations) to the PhD thesis.

A. Fine-tuned Model URLs

Model IDs and URLs are provided in Table 5 below. Accuracy (*Acc*) may differ slightly from results reported in Table 3 for the same base-model and fine-tuning condition, because the earlier table reports averages across 5 seeds, and the table below reports accuracy for the specific checkpoint uploaded to Hugging Face.

Cd	ID	Acc
sg	compellit/llama-carvalho-scansion-gl-sg	86.75
cx	compellit/llama-carvalho-scansion-gl-cx	88.75
sg	compellit/mt5-scansion-gl-sg	89.5
cx	compellit/mt5-scansion-gl-cx	91
sg	compellit/byt5-scansion-gl-sg	92.5
cx	compellit/byt5-scansion-gl-cx	93.25

Table 5: Model IDs on Hugging Face. Input-type refers to the (*sg*) and context (*cx*) conditions (*Cd*) from Section 5.3). *Acc* refers to exact match per-line in metrical syllabification.