Cross-Lingual and Cross-Domain Transfer Learning for POS Tagging in Historical Germanic Low-Resource Languages

Irene Miani[1] and Sara Stymne[2] and Gregory Darwin[3]
Department of Linguistics and Philology[1, 2] and Department of English[3], Uppsala University
(irene.miani[1], sara.stymne[2])@lingfil.uu.se, gregory.darwin@engelska.uu.se[3]

Abstract

Although Part-of-Speech (POS) tagging has been widely studied, it still presents several challenges, particularly reduced performance on out-of-domain data. While increasing in-domain training data can be effective, this strategy is often impractical in historical low-resource settings. Cross-lingual transfer learning has shown promise for low-resource languages; however, its impact on domain generalization has received limited attention and may remain insufficient when used in isolation. This study focuses on cross-lingual and cross-domain transfer learning for POS tagging on four historical Germanic low-resource languages in two literary genres. For each language, POS tagged data were extracted and mapped to the Universal Dependencies UPOS tag set to establish a monolingual baseline and train three multilingual models in two dataset configurations. The results were consistent with previous findings, indicating that structural differences between the genres can negatively influence transfer learning. The poetry-only multilingual model showed improvements within that domain compared to the baseline. In contrast, multilingual models trained with all available data had lower performance caused by substantial structural differences in the corpora. This study underlines the importance of investigating the domain-generalization abilities of the models, which may be negatively influenced by substantial structural differences between data. In addition, it sheds light on the study of historical low-resource languages.

1 Introduction

Part-of-Speech (POS) tagging is a fundamental task in Natural Language Processing (NLP) that builds the syntactic foundations for accurate linguistic analysis (Baruah and Jyoti Goutom, 2025). Although this task has been extensively studied over the years (Behzad and Zeldes, 2020; Arai, 2021; Hansen and van der Goot, 2023; Miani et al., 2025), several studies have shown that the performance of POS tagging tools is not consistent across domains, such as diverse literary genres, dialects, or time periods. This performance drop is most likely caused by the structural differences between domains that can lead to uneven distribution of linguistic patterns in the training data (Arai, 2021; Miani et al., 2025). Robust POS tagging performance across domains is fundamental to support the study of genre variation, syntactic patterns, and language use across corpora. In addition, it can also facilitate large-scale analysis and support automatic annotation. Among the solutions that have been proposed to strengthen the performance of these tools, the most common and successful approaches involve increasing the size of in-domain training data. However, it is not always a suitable solution as low-resource languages do not have enough data to increase training sizes; thus, other solutions have been taken into consideration. Cross-lingual transfer learning has been proven to be an effective approach in low-resource settings (de Vries et al., 2022; Rice et al., 2025; Schöffel et al., 2025a). Linguistic similarity — such as belonging to the same language family, similar writing systems, and common word orders — appeared to facilitate the learning process, leading to improved performance. Nevertheless, once model performance is strengthened through cross-lingual transfer, the domain-switching problem often receives less attention. As a result, these tools may be able to achieve high performance in the low-resource context, but they may still struggle to generalize across domains. For this reason, the present study focuses on both cross-lingual and cross-domain transfer learning in low-resource settings.

This study focuses on low-resource historical languages, which suffer from severe data scarcity and a lack of reliable POS tagging tools. Moreover, these languages are often overlooked in technological advancements in NLP, leading to unreliable

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 542–558

tools and, consequently, flawed linguistic analyses for linguistic and philological research. Following findings from previous studies, which highlighted the benefits of selecting languages from the same linguistic family to facilitate cross-lingual transfer learning, we focused on four historical languages from the Germanic family: Old English, Old Norse, Old High German, and Old Saxon. Another factor influencing the selection was the availability of data in the domains investigated in this study, specifically two literary genres: poetry and prose. The substantial structural differences between these genres make them particularly suitable for investigating transfer learning approaches, and the corpora selected for this study contain data for both genres.

From the corpora, POS annotated data were extracted and used to establish a monolingual baseline and train three multilingual models: one model trained only on poetry data, one only on prose data, and one trained on all available data for all languages and both genres. The models were all trained with two dataset size settings. The results were consistent with previous studies: the structural differences between the genres influence the learning process, lowering the performance of the models. In the poetry domain, the one-genre approach showed improvements compared to the baselines. The all-data approach highlighted the negative influence of the structural distributions.

This study underlines the importance of investigating the domain-generalization abilities of the models, that may be poor because of substantial structural differences between the data. In addition, it sheds light on the study of historical low-resource languages, usually overlooked, but that require tools to be able to perform accurate linguistic analysis.

2 Related Work

Both cross-domain and cross-lingual transfer learning for POS tagging have been extensively studied over the years from different perspectives. Concerning cross-domain transfer learning, Behzad and Zeldes (2020) recently investigated the robustness of POS tagging models across genres using user-generated Reddit texts. The results showed that small amounts of in-domain data can outperform larger out-of-domain data, highlighting the influence of genre differences on models' performance. Arai (2021) addresses the domain shift problem for Modern English poetry. Since existing POS taggers' performances became worse when subjected to poetry data, data augmentation techniques were implemented to address the problem. Other studies, such as Abo Mokh et al. (2022), focused on out-of-domain POS tagging for four Arabic dialects, which was particularly challenging since most existing taggers are trained on Modern Standard Arabic and not on dialectal data. They investigated three approaches to improve the predictions: up-sampling target dialect data within a joint tagging model to address data imbalance, increasing annotation consistency across dialectal training data, and incorporating pre-trained word embeddings learned from large dialectal corpora. They increased the accuracy on out-of-domain dialectal text by circa 20% on average compared to the baseline. Hansen and van der Goot (2023) tested English POS taggers on the Wall Street Journal portion of the Penn Treebank on an out-of-domain dataset from the Elder Scrolls Fandom Wiki. Accuracy dropped with out-of-domain data, especially with unknown tokens, decreasing from 90% in-domain to around 78–80% out-of-domain, highlighting the influence of structural differences on the training.

Cross-lingual transfer learning in the low-resource scenario has been investigated by de Vries et al. (2022), who focused on zero-shot learning for POS tagging on 65 source languages and 105 target ones using the Universal Dependencies dataset. The analysis revealed that matching language family, writing system, word order, and inclusion of languages in pre-training can significantly strengthen the learning process. Recently, Rice et al. (2025) analyzed the factors that determine effective transfer language selection for zero-shot cross-lingual POS tagging. They proposed a holistic ranking approach that combines fine-grained typological features and dataset-dependent metrics to rank potential source languages for transfer.

For what concerns the selected historical low-resource languages, the majority of the studies have been conducted on Old English or Old High German or historically related languages, such as Middle English or Middle High German. To the extent of our knowledge, we are not aware of any research on POS-tagging for Old Saxon and Old Norse. For Old English, a POS tagging tool is available as part of the CLTK library (Johnson et al., 2021), which was trained on the data from the ISWOC Treebank

(Bech and Eide, 2014). The accuracy of this model is uncertain. Recent work has been conducted by Miani et al. (2025) who focused on cross-domain transfer learning for Old English poetry and prose. The study highlighted the importance of in-domain data for successful performance. For what concerns Old High German, some studies focused on evaluating existing POS taggers for Early Modern German corpora (Scheible et al., 2011; Ferreri Hanberry, 2015). Koleva et al. (2017) extended this work to Middle Low German, developing a dedicated POS tagger. Recent work focused on cross-lingual transfer learning approaches includes Nie et al. (2023), who demonstrated that delexicalized cross-lingual parsing from Modern German can effectively predict syntactic structures for Middle High German, circumventing the lack of large annotated treebanks.

3 Datasets

For Old English, two corpora were used: the York-Helsinki Parsed Corpus of Old English Poetry (YCOEP)[1], containing a selection of poems from the Helsinki Corpus of English Texts totaling 71.490 words; and the York Toronto Helsinki Parsed Corpus of Old English (YCOE)[2] comprising approximately 1.5 million words of Old English prose annotated for syntax and morphology. Since the YCOE documentation is the only one available, it was adopted for both datasets. The segmentation of both corpora is based on units called *tokens* which may represent one main verb with arguments and adjuncts, matrix inflectional phrases, complementizer phrases, or independent non-clausal utterances. The original textual form of each unit was extracted, together with the corresponding POS tag.

The Old Saxon poetry data were extracted from the HeliPaD[3] dataset, which comprises 5.968 lines from the C manuscript of the Old Saxon text *Heliand*. The annotation included textual and metrical information, lemmatization, POS tagging, morphology, and syntactic parsing. The annotation of HeliPaD follows the same guidelines as YCOEP and YCOE; therefore, segmentation and POS tags are consistent across these three corpora. As with the Old English corpora, the textual form of the units was extracted with their POS tags.

Both Old Norse poetry and prose data were extracted from the Menotec collection[4] available on the INESS platform (Rosén et al., 2012). The collection comprises seven treebanks with a total of 20.308 sentences, from which five Old Norwegian manuscripts were selected. For poetry, *Edda Regius* was used, containing 3665 lines. The prose data consist of the following texts, comprising up to 10.318 lines in total: *Pamphilus*, *The Old Norwegian Homily Book*, *the legendary saga of St Olaf*, *Strengleikar*. All the texts contain morphological and syntactical information. The annotation followed the guidelines for the annotation of Old Norwegian texts by Haugen and Øverland (2014). [5]

For Old High German, the Deutsch Diachron Digital, Referenzkorpus Altdeutsch (Version 1.2) dataset (Zeige et al., 2025) was used. The dataset contains approximately 650.000 words and covers the period from 750 to 1050 and includes both Old High German and Old Saxon poetry and prose manuscripts. Old Saxon prose data and additional poetry were extracted from this corpus and used alongside HeliPaD. The data present structural and linguistic annotations, as well as syntactic sentence information. The smallest unit is the *token*, which corresponds to both the edited and normalized versions of each word in the text. The normalized tokens were extracted with their POS tag and used to train the models.

4 POS Mapping

Since each corpus presents different POS tag sets, all annotations were mapped to the Universal Dependencies UPOS tag set (de Marneffe et al., 2021) to ensure consistency. The mapping for each corpus is reported in Appendix A.

For Old English, the mapping follows the work of Miani et al. (2025). As HeliPaD, the Old Saxon poetry dataset, relied on the same guidelines as the Old English datasets, the same mapping strategy was applied. The scheme conversions can be found in Table 5 and 6, respectively.

The mapping of the Old Norse tag set, presented in Table 7, was straightforward for all tags, except for auxiliary verbs. The annotation did not present any specific tag, but the auxiliary relation of the verbs was expressed in the `relation` attribute with

[1] https://www-users.york.ac.uk/~lang18/pcorpus.html

[2] https://penn-historical-corpora.uni-mannheim.de/ycoe/YCOEHomepage.html

[3] https://zenodo.org/records/4395040

[4] https://clarino.uib.no/iness

[5] Documentation for the POS annotation was kindly provided by Paul Meurer and Odd Einar Haugen.

Language	Train	Development	Test
Old High German	5,327	665	667
Old English	5,039	629	631
Old Norse	3,029	378	380
Old Saxon	444	55	57

Table 1: Number of sentences in the training, development, and test split for each language, indicated in the Language column.

the aux value. Verbs presenting this value were therefore mapped to the UPOS AUX tag.

For the Old High German data, the documentation included detailed descriptions of the POS annotation (Zeige et al., 2025). The original tag set is highly granular, but it was mapped to the UPOS tag set without the need for additional language-specific rules. However, special attention was required for particles. In the original annotation, several tags referred to particles, such as PTKVZ for separated verb particles, PTKINT for interrogative particles, or PTK for general ones. To ensure consistency with the Universal Dependencies guidelines, only the following tags have been tagged with the PART tag: PTKNEG (negation particles), PTKZU, and PTK. Table 8 lists all the conversions for this corpus.

All mapped annotations were converted into CONLL format; data format required by MaChAmp (van der Goot et al., 2021), the toolkit used to train the models.

5 Experimental Setup

Two monolingual baselines with two different dataset sizes were established, followed by an experiment focusing on training three multilingual models in the same two dataset sizes.

For the first baseline, three monolingual models per language were trained with equal amounts of poetry and prose data. Specifically, one model was trained exclusively with poetry data, one with exclusively prose data, and the third with a combination of both genres. The data were divided into 80% for training, 10% for development, and 10% for testing. The resulting dataset sizes are listed in Table 1.

A second baseline was trained using an equal amount of data for each language. The dataset size was determined by the smallest dataset used in the first baseline, Old Saxon prose, which consists of 444 training sentences and 55 development sentences. For each language, the corresponding data were sampled from the training and development splits of the first baseline. The original test sets were retained.

To investigate cross-lingual and cross-domain transfer learning, firstly, we trained two multilingual models: one model was trained exclusively on poetry data, and the other only on prose data. The experiment was repeated for both dataset sizes described above. Each model was evaluated on the test sets of each language, both within the same genre used for training and across the opposite genre. Then, a single multilingual model was trained using all available data from all languages and both genres. The training was repeated for both dataset sizes, and evaluation was performed on the same test sets.

With respect to the modeling approach adopted, recent work has shown that large language models exhibit notable limitations for POS tagging in historical low-resource settings (Schöffel et al., 2025b). In contrast, traditional approaches, such as MaChAmp, a toolkit for multi-task learning, have demonstrated stronger performance (Miani et al., 2025). For this reason, all models in this study are trained using MaChAmp (van der Goot et al., 2021), which offered a straightforward configuration in a multi-dataset scenario that helped the training process. The MaChAmp toolkit supports multiple NLP tasks and is built around a pre-trained encoder that is fine-tuned for specific tasks during training. For POS tagging, we employed the seq task type, which applies a greedy softmax classification layer over the contextualized token embeddings provided by the encoder. Multilingual BERT was the language model used with default hyperparameters. The models were trained for 20 epochs using three random seeds. Evaluation was performed on the same test sets derived from the original dataset splits, per language and per genre. Accuracy and macro F1 score were computed for each seed; the averaged results will be reported in Section 6.

6 Results

Table 2 reports the results of the monolingual baseline models trained on the original and on the reduced dataset sizes. Tables 3 and 4 present the evaluation results of the multilingual models on the poetry and the prose test sets, respectively. Appendix A includes Table 9 and 10, which list the five most frequent POS bigrams and trigrams per language for the reduced poetry and prose datasets. In addition, Figures 1 and 2 compare the POS tag

Genre	Test	Original Sizes		Reduced Sizes	
		Acc.	F1	Acc.	F1
Old English					
poe	poe	0.959	0.96	0.853	0.726
pro	poe	0.873	0.817	0.757	0.656
poe, pro	*poe*	*0.963*	*0.964*	*0.879*	*0.799*
poe	prose	0.921	0.833	0.828	0.736
pro	prose	0.97	0.968	0.89	0.832
poe, pro	*prose*	*0.977*	*0.968*	*0.912*	*0.867*
Old Norse					
poe	*poe*	*0.925*	*0.928*	*0.853*	0.778
pro	poe	0.754	0.644	0.709	0.610
poe, pro	poe	0.928	0.894	0.847	*0.82*
poe	prose	0.768	0.58	0.659	0.476
pro	*prose*	*0.93*	*0.857*	0.865	*0.773*
poe, pro	prose	0.929	0.853	*0.869*	0.746
Old High German					
poe	poe	0.938	0.842	0.837	0.661
pro	poe	0.828	0.678	0.714	0.58
poe, pro	*poe*	*0.947*	*0.852*	*0.864*	*0.722*
poe	prose	0.834	0.746	0.71	0.584
pro	prose	0.942	0.912	0.84	0.746
poe, pro	*prose*	*0.945*	*0.912*	*0.856*	*0.778*
Old Saxon					
poe	*poe*	*0.894*	*0.816*	*0.894*	*0.816*
pro	poe	0.694	0.638	0.694	0.638
poe, pro	poe	0.89	0.816	0.89	0.816
poe	prose	0.62	0.532	0.62	0.532
pro	prose	0.891	0.8	0.891	0.8
poe, pro	*prose*	*0.897*	*0.82*	*0.897*	*0.82*

Table 2: POS tagging results for the monolingual baseline models. The Genre column indicates on which genre the model was trained: poetry (poe), prose (pro), and a combination of poetry and prose (poe,pro). The Test column specifies the genre of the test set. Acc. presents the scores for the accuracy and F1 for the macro F1 score. Each language presents the results for dataset configurations. *Italics* highlight the highest results.

distributions across the four languages, with Figure 1 illustrating the poetry data and Figure 2 the prose data. The distributions of POS-tags and N-grams are used to support the findings discussion.

6.1 Baselines

In Table 2, the models named poe were trained only with poetry data, the ones named pro only with prose data. The models named poe,pro were trained on a combination of both genres. The models were tested on two test sets per language: one for the poetry data and one for the prose data.

Across all languages, models trained on a single genre achieve higher performance when evaluated on in-domain data than on out-of-domain data. The differences between the models' results are consistent, with Old Norse and Old Saxon presenting the largest performance gaps between in-domain and out-of-domain evaluation across both accuracy and macro F1 score. When considering the POS distributions shown in Figures 1 and 2, it is clear that there are some fundamental differences between the genres for all languages, with more substantial divergences observed in Old Norse. The different structural patterns between poetry and prose are influencing the learning process and making it more difficult for the model to generalize to out-of-domain data.

Of the models trained on both genres, poe,pro, in three languages — Old English, Old High German, and Old Saxon — the results are slightly higher than the single-domain models. In contrast, Old Norse faces a slight decrease in the macro F1 score when the model is evaluated on the poetry test set. In any case, the performance gains obtained by combining genres are not outstanding; thus, we could expect slightly refined predictions for rare tags, but not a substantial advance in the overall learning process.

Considering performance across languages, the Old English models achieve the highest scores, followed by the Old High German models. These models have competitive results on both the in-domain and out-of-domain data, as well as the lowest differences between accuracy and macro F1 score. This indicates that the models can correctly predict all tags, even the rare ones. In contrast, Old Norse and Old Saxon exhibit larger gaps between accuracy and F1 score, ranging from 11% to 27%. The weakest performance is observed for the Old Saxon model trained on poetry and evaluated on prose: it yields the lowest scores overall, and the largest gap compared to the in-domain counterpart. Once again, these results suggest that the structural differences between the genres are influencing the learning process: they are so broad that the models struggle to generalize across domains, particularly for the Old Norse and Old Saxon data.

Table 2 also presents the results of the monolingual baseline models trained with the reduced dataset sizes. As expected, performance is lower than in the original-size baseline, but the behaviors remain consistent. Models trained on both genres achieve better results, with slightly larger gains relative to single-genre models; however, there are still no outstanding differences. These increases could suggest that the reduced amount of training data also narrows the differences between the genres, thereby facilitating a small degree of cross-domain generalization.

In this reduced-data setting, the best-performing

model is the Old English model trained on both genres and evaluated on prose data, achieving 91% for accuracy and 87% for F1 score. It is followed by the Old Saxon model trained and evaluated on poetry data, with 89% for accuracy and 82% for F1 score. The Old Norse model trained on poetry and evaluated on prose data presents the worst results, reaching only 66% for accuracy and 48% for F1 score.

Overall, the reduced-size baseline corroborates the findings from the original-size baseline: genre-specific structural differences are too broad to allow the models to generalize across domains. Combining the genres helps the models to refine the predictions, with slightly higher scores, but not outstanding improvements. These findings highlight the importance of incorporating both in-domain and out-of-domain data to achieve more robust POS tagging performance across genres, and they are consistent with previous work (Miani et al., 2025).

6.2 Cross-Lingual and Cross-Domain Transfer Learning Experiment

Tables 3 and 4 present the results for the evaluation of the multilingual models on the poetry and the prose test sets, respectively. Both tables repeat the results from the monolingual baselines, indicated with the `mono-` tag in the Model column. `multi-` refers to the multilingual models. As for Table 2, in the Model column, `poe` refers to the models trained with only poetry data, `pro` to the ones trained with only prose data, and `poe,pro` to the ones trained with a combination of both genres. The models were trained for both dataset configurations, which are indicated in the columns as *Original Sizes* and as *Reduced Sizes*. All reported accuracy and macro F1 scores correspond to averages over three random seeds. The results on the poetry test sets will be discussed in Section 6.2.1, and the ones for the prose test sets in Section 6.2.2.

6.2.1 Evaluation on Poetry Test Sets

The evaluations of the poetry-only multilingual model on the poetry test sets, `multi-poe`, show slightly higher scores compared to the single-genre monolingual baseline models, `mono-poe`, especially for Old Norse, Old High German, and Old Saxon. For the original dataset setting, Old English and Old High German achieve modest improvements; Old Saxon increases the F1 score by 4% points, while Old Norse shows a decrease of 2% points. In the reduced setting, all languages exhibit improvements, particularly in the F1 score, with Old English and Old Norse improving by 6% points and Old High German by 4% points. The improvements could signal slightly refined predictions, including less frequent POS tags. The results suggest that cross-lingual transfer is effective in the poetry domain. Supporting evidence can be found in the POS N-gram statistics reported in Appendix A, specifically in Table 9. The majority of the bigrams and trigrams are largely shared between the languages: for instance, 'NOUN'-'PUNCT', 'NOUN'-'VERB', and 'NOUN'-'VERB'-'PUNCT' are common to all languages. This similarity in POS sequence distributions indicates comparable structural patterns across the four languages, which facilitate the learning process and enable effective cross-lingual transfer.

Model	Original Sizes		Reduced Sizes	
	Acc.	F1	Acc.	F1
Old English				
mono-poe	0.959	0.96	0.853	0.726
mono-pro	0.873	0.817	0.757	0.656
mono-poe, pro	*0.963*	*0.964*	*0.879*	0.799
multi-poe	0.958	0.956	0.862	0.789
multi-pro	0.867	0.798	0.764	0.670
multi-poe, pro	0.960	0.954	0.872	*0.811*
Old Norse				
mono-poe	0.925	*0.928*	*0.853*	0.778
mono-pro	0.754	0.644	0.709	0.610
mono-poe, pro	0.928	0.894	0.847	0.82
multi-poe	0.927	0.907	0.850	*0.836*
multi-pro	0.767	0.641	0.726	0.617
multi-poe, pro	*0.929*	0.909	0.847	0.822
Old High German				
mono-poe	0.938	0.842	0.837	0.661
mono-pro	0.828	0.678	0.714	0.58
mono-poe, pro	*0.947*	*0.852*	0.864	0.722
multi-poe	0.940	0.848	0.847	0.706
multi-pro	0.817	0.668	0.717	0.579
multi-poe, pro	0.945	0.844	*0.866*	*0.728*
Old Saxon				
mono-poe	0.894	0.816	0.894	0.816
mono-pro	0.694	0.638	0.694	0.638
mono-poe, pro	0.89	0.816	0.89	0.816
multi-poe	0.902	*0.852*	*0.903*	*0.826*
multi-pro	0.768	0.738	0.744	0.672
multi-poe, pro	*0.906*	0.844	0.899	0.834

Table 3: Evaluation results of the multilingual models in the **poetry** domain. `mono-` indicates the monolingual baseline models, and `multi-` the multilingual ones. `poe` signals the training only on poetry data, `pro` the one only on prose data, and `poe, pro` on both genres. Each language presents the results for both dataset configurations. Acc. and F1 list respectively the scores for accuracy and macro F1 score. *Italics* highlight the highest results.

In contrast, the multilingual models trained with prose-only data, `multi-pro`, when tested on the poetry domain, showed decreased performance in the original dataset setting but slight improvements in the reduced one. The performance gains of the reduced setting could be related to the comparatively simpler structural patterns of poetry, rather than effective transfer from prose data. However, Old Saxon exhibits the best improvements in both configurations: in the original one, the accuracy increases by 7% points and the F1 score by 10%; in the reduced setting, the metrics increase by 5% and 4% points, respectively. The cross-lingual transfer appears to be more effective for this language compared to the others.

The performance of the multilingual models trained with all available data, `multi-poe,pro`, for all languages except Old Saxon, either remains unchanged or decreases slightly compared to the baseline. Old Saxon stands out with higher scores in both settings: in the original setting, the macro F1 score increases by approximately 3% points, while in the reduced setting it improves by about 2% points. Compared to the one-genre multilingual models, `multi-poe`, results on the poetry test are largely consistent, with no marked improvements or degradations. This could support the hypothesis that the cross-lingual approach does not compensate for the absence of in-domain data, and that it is fundamental to take this into consideration when adopting these techniques.

6.2.2 Evaluation on Prose Test Sets

When evaluated on the prose test sets, both dataset configurations show a lower or equal performance compared to the baseline models. The worst scores, in the original configuration, are reported by the Old High German model `multi-pro`, with a decrease of 6% points for the same metric. In the reduced configuration, the Old Norse `multi-pro` model presents a decreased F1 score of 4% points, and the same model in Old Saxon shows a difference of 6% points from the `mono-pro` model. The results are consistent with the findings from the POS N-gram analysis: the prose datasets are structurally more heterogeneous across languages, and these structural divergences seem to limit the effectiveness of cross-lingual transfer. Notably, even the reduced dataset setting—which led to improved performance in the poetry-based experiment reported in Table 3—does not yield comparable benefits for prose.

Model	Original Sizes		Reduced Sizes	
	Acc.	F1	Acc.	F1
Old English				
mono-poe	0.921	0.833	0.828	0.736
mono-pro	0.97	0.968	0.89	0.832
mono-poe, pro	*0.977*	*0.968*	*0.912*	*0.867*
multi-poe	0.921	0.819	0.838	0.745
multi-pro	0.965	0.947	0.884	0.814
multi-poe, pro	0.971	0.937	0.904	0.826
Old Norse				
mono-poe	0.768	0.58	0.659	0.476
mono-pro	*0.93*	*0.857*	0.865	0.773
mono-poe, pro	0.929	0.853	*0.869*	*0.746*
multi-poe	0.765	0.601	0.686	0.517
multi-pro	0.929	0.856	0.865	0.735
multi-poe, pro	0.929	0.822	0.861	0.693
Old High German				
mono-poe	0.834	0.746	0.71	0.584
mono-pro	0.942	0.912	0.84	0.746
mono-poe, pro	*0.945*	*0.912*	0.856	*0.778*
multi-poe	0.833	0.698	0.744	0.617
multi-pro	0.942	0.856	0.854	0.723
multi-poe, pro	0.944	0.86	*0.867*	0.737
Old Saxon				
mono-poe	0.62	0.532	0.62	0.532
mono-pro	0.891	0.8	0.891	0.8
mono-poe, pro	0.897	0.82	0.897	*0.82*
multi-poe	0.726	0.557	0.699	0.552
multi-pro	0.909	0.795	0.894	0.745
multi-poe, pro	*0.918*	*0.824*	*0.903*	0.769

Table 4: Evaluation results of the multilingual models in the **prose** domain. `mono-` indicates the monolingual baseline models, and `multi-` the multilingual ones. `poe` signals the training only on poetry data, `pro` the one only on prose data, and `poe, pro` on both genres. Each language presents the results for both dataset configurations. Acc. presents the scores for the accuracy and F1 for the macro F1 score. *Italics* highlight the highest results.

In contrast, the evaluation of the multilingual model trained on the opposite genre shows a different behavior compared to previous results. In the original setting, the poetry-only multilingual models, `multi-poe`, tested on the prose test sets have similar results compared to the baseline, for Old English and Old High German. Old Norse increases its F1 score by 2% points, and Old Saxon by 10% points for the accuracy, and by 2% points for the F1 score. In the reduced configuration, all the languages show slight improvements compared to the baseline, except Old English. Old Norse presents an increase of between 3% and 4% points for both metrics; Old High German increases both metrics by 3 points, and Old Saxon shows a substantial increase in the accuracy. An analysis of the POS N-gram distributions for prose reveals fewer shared bigrams and trigrams across languages, with

only a few common ones, such as 'DET'-'NOUN' or 'DET'-'NOUN'-'PUNCT'. These structural differences seem to reduce the effectiveness of the cross-lingual transfer, as the training data exhibit different distributions. The marginally better results observed in the reduced setting could more likely be attributed to a narrowed range of structural variation, which can hide the struggles in domain switch, rather than an actual gain from the learning process.

Compared to the baseline, the evaluation of the `multi-poe,pro` models on the prose test set reveals a general decrease — between 3% and 6% points for both metrics — in both configurations, except for Old Saxon, which shows again a very slight improvement. Compared to the one-genre multilingual models, `multi-pro`, the scores do not show any interesting improvements, as in Table 3. The behaviors of the models evaluated on the prose test sets are consistent and, in most cases, worse than the ones from the poetry test sets, supporting once again the idea that the cross-lingual approach does not provide better domain-generalization abilities; on the contrary, it could lead to worse performance because the substantial structural differences between the genres are negatively influencing the learning process.

6.3 Discussion

The evaluation results for the monolingual baseline models were consistent with previous findings, indicating that different structural distributions can influence the tagging process, resulting in lower scores than when the models are evaluated on the opposite genre. This behavior is consistent across all languages and in both dataset configurations.

The evaluation of the multilingual models trained on a single genre revealed performance improvements, particularly in the poetry domain. The shared POS distributions enabled the models to generalize successfully, even in the reduced dataset setting. In contrast, the prose domain presented a substantially harder scenario for the learning process. All languages presented complex structural distribution for the prose with few shared features that influenced the transfer learning.

On the contrary, combining all available data in a multilingual model proved to be a less effective approach. Old Norse and Old Saxon showed slight improvements, especially in the reduced setting, suggesting that the gains were related to narrower structural variation rather than robust cross-domain learning. The genre differences were too broad and impacted the POS tagging abilities of the model. Overall, combining data from different languages and genres does not mitigate the structural differences between domains.

7 Conclusions

The study investigates cross-lingual and cross-domain transfer learning on POS tagging for four historical Germanic low-resource languages and two literary genres. POS annotated poetry and prose data were extracted and mapped to the UPOS tag set. The data were used to establish a monolingual baseline and train three multilingual models in two different dataset configurations.

The results were overall consistent with previous findings, indicating that the differences in the structures of the corpora can influence the transfer learning, worsening the performance of the models. Multilingual models trained on only one genre showed improvements compared to the monolingual ones, while the multilingual models trained with all available data showed lower performance influenced by the substantial differences in the corpora.

Future work should explore genre-aware strategies or targeted domain adaptation techniques; comparing these findings with LLM-based tagging could also lead to interesting analysis. Additionally, exploring language-family-specific pretraining may further improve transfer learning for historical low-resource languages.

Limitations

This study investigated cross-lingual and domain transfer learning on POS tagging for four historical Germanic low-resource languages and two literary genres.

All four languages are morphologically rich and originally annotated with highly fine-grained POS tag sets. Mapping the tag sets to the UPOS tag set was the most suitable approach for enabling cross-lingual learning and comparison of model performance; however, a part of the linguistic information is lost. Moreover, annotation errors in the original datasets, as well as mapping inaccuracies, may affect the quality of the resulting UPOS labels.

To ensure comparable experiments, the dataset sizes needed to be substantially reduced. This may have negatively impacted model performance and potentially obscured better performance that could

be achieved with larger training datasets. In addition, all models were trained using default hyperparameters, which ensures consistency, but more careful hyperparameter optimization could improve performance and should be explored in future work.

Acknowledgments

This work has been supported by the Swedish Graduate School of Digital Philology, funded by the Swedish Research Council (grant 2022-06343). Computations were enabled by resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We would like to thank Paul Meurer and Odd Einar Haugen for providing documentation on POS tagging for the Old Norse corpus, Menotec.

References

Momen Abo Mokh, Houda Bouamor, and Nizar Habash. 2022. Improving POS tagging for Arabic dialects on out-of-domain texts. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 205–215. Association for Computational Linguistics.

Hirona Jacqueline Arai. 2021. *Optimizing an automatic part of speech tagger for poetry text using data augmentation*. Undergraduate thesis, Middlebury College, Computer Science Department.

Nomi Baruah and Pritom Jyoti Goutom. 2025. A comparative analysis of deep learning and machine learning for pos tagging. *Expert Systems with Applications*, 288:128026.

Kristin Bech and Kristine Eide. 2014. The iswoc corpus. Department of Literature, Area Studies and European Languages, University of Oslo.

Shabnam Behzad and Amir Zeldes. 2020. A cross-genre ensemble approach to robust reddit part of speech tagging. *arXiv e-prints*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Brendan Ferreri Hanberry. 2015. Application of a pos tagger to a novel chronological division of early modern german text. Master's thesis, University of North Carolina at Chapel Hill.

Kia Kirstein Hansen and Rob van der Goot. 2023. Cross-domain evaluation of pos taggers: From wall street journal to fandom wiki. *arXiv e-prints*.

Odd Einar Haugen and Fartein Th. Øverland. 2014. Guidelines for morphological and syntactic annotation of old norwegian texts. *Bergen Language and Linguistics Studies (BeLLS)*, 4(2).

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Mariya Koleva, Melissa Farasyn, Bart Desmet, Anne Breitbarth, and Veronique Hoste. 2017. An automatic part-of-speech tagger for middle low german. *International Journal of Corpus Linguistics*, 22(1):108–141.

Irene Miani, Sara Stymne, and Gregory R. Darwin. 2025. Cross-genre learning for Old English poetry POS tagging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 708–724, Vienna, Austria. Association for Computational Linguistics.

Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2023. Cross-lingual constituency parsing for middle high german: A delexicalized approach. *arXiv preprint arXiv:2308.04645*.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data and model architecture on ranking transfer languages for cross-lingual pos tagging. *arXiv e-prints*.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29, Istanbul, Turkey.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' pos-tagger on early modern german text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–23, Portland, OR, USA. Association for Computational Linguistics.

Matthias Schöffel, Esteban Garces Arias, Marinus Wiedner, Paula Ruppert, Meimingwei Li, Christian Heumann, and Matthias Aßenmacher. 2025a. Unveiling factors for enhanced pos tagging: A study of low-resource medieval romance languages. *Preprint*, arXiv:2506.17715.

Matthias Schöffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher. 2025b. Modern models, medieval texts: A pos tagging study of old occitan. *arXiv preprint arXiv:2503.07827*.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Lars Erik Zeige, Gohar Schnelle, Martin Klotz, Karin Donhauser, and and Lühr Rosemarie Gippert, Jost. 2025. Deutsch Diachron Digital - Referenzkorpus Altdeutsch (Version 1.2). Humboldt-Universität zu Berlin.

A Appendix

POS Mapping and Statistical Analysis

The POS tagging conversion schemes for each language, discussed in Section 4, are presented in Table 5, 6, 7, and 8. Here, POS indicates the name of the syntactic category, and UPOS, its corresponding Universal POS tag. The following column lists the original tag sets used in each corpus.

The statistical analysis results on the reduced training configurations, which support the discussion in Section 6, are shown in Table 9 and 10, and Figures 1 and 2. Specifically, the POS N-gram analysis of the five most frequent POS bigrams and trigrams for each language is listed in Tables 9 for the poetry and 10 for the prose. Figures 1 and 2 illustrate the overall POS tag distributions across all languages, respectively, in the poetry reduced configurations and the prose ones.

POS	UPOS	YCOEP	YCOE
Adjective	ADJ	VBNˆD, WADJˆN, ADJˆG, VAGˆN, WADJˆD, VAGˆD, VBNˆG, VBNˆA, ADJˆA, ADJˆN, ADJˆI, BENˆN, ADJˆD, ADJ, VAGˆA, WADJˆA, ADJP-NOM, VAGˆG, VBNˆN	HVNˆN, ADJ, ADJSˆA, VAGˆI, WADJˆG, ADJS, ADJRˆN, MAGˆG, ADJˆA, HAGˆA, WADJˆD, VAGˆN, ADJRˆI, WADJˆI, HAGˆN, WADJˆN, VAGˆA, ADJR, ADJˆD
Adposition	ADP	P	P21, P22, PP, P+Dˆ I, P
Adverb	ADV	RP, ADVˆD, WADVˆD, RP-1, ADVˆL, ADV, WADVˆDX, ADVˆDX, WADVˆL, WADVˆT, WADV-1, WADV, ADVˆT, ADVP	ADVˆT22, WADVˆD, ADV, ADVRˆD, ADVP, ADVP-LOC, ADVSˆT, ADVˆT21
Auxiliary	AUX	AXDS, AXP, MDI, AXPS, MDPS, AXI, MDPI, AXN, MDDI, MDD, AXDI, AXPI, MD, AXD, MDP, AX, MDDS	MD, AXG, MDDI, AXDS, AXDI, MDPS, MDP, AXI, MDD
Coordinating Conjunction	CCONJ	CONJ	CONJ
Determiner	DET	QˆG, QˆI, DˆG, DˆD, Q, QˆA, DˆI, DˆN, DˆA, QˆN, QˆD	QˆD, QSˆA, D, QRˆN, Q+QˆN, DˆN
Interjection	INTJ	INTJ	INTJ
Noun	NOUN	NP-ACC-SBJ, NˆG, NP-ACC, NP-DAT, NˆD, NˆI, NˆN, NP-NOM, NˆA	NP-ACC, NˆG, NP-NOM, NˆA, NˆN
Numeral	NUM	NUMˆA, NUMˆG, NUMˆI, NUM, NUMˆD, NUMˆN	NUMˆD, NUMˆG, NUMˆA, NUMˆN
Particle	PART	TO, UTP, FP, NEG	TO, FP, NEG, UTP
Pronoun	PRON	PRO$, PROˆA, PROˆD, WPRO, PROˆN, MANˆN	PRO, WPRO, PROˆG, MANˆN
Proper Noun	PROPN	NPR, NPRˆN, NPRˆG	NR, NRˆN, NRˆG
Punctuation	PUNCT	, .	, .
Subordinating Conjunction	SCONJ	WNP-ACC, WNP-NOM, WQ, C	C, WQ, CP-REL
Verb	VERB	BE, VBD, VBN, VBPI, VAG, VBDI, BED, VB, VBP	BE, VBD, VBN, VAG, VB, VBP
Other	X	FW, UNKNOWN	FW, UNKNOWN
Symbol	SYM	-	-

Table 5: Mapping of YCOEP and YCOE to UPOS.

POS	UPOS	HeliPaD
Adjective	ADJ	ADJ, ADJR, ADJS, WADJ, VGI, VNI
Adposition	ADP	P
Adverb	ADV	ADV, ADVR, ADVS, WADV, ALSO
Auxiliary	AUX	AX, MD, MDI, MDPI, MDPS, MDDI, MDDS, MG
Coordinating Conjunction	CCONJ	CONJ
Determiner	DET	D, Q, QR, QS
Interjection	INTJ	INTJ
Noun	NOUN	N
Numeral	NUM	NUM
Particle	PART	RP, NEG, TO, UTP
Pronoun	PRON	MAN, PRO, PRO$, WPRO, WPRO$
Proper Noun	PROPN	NPR
Punctuation	PUNCT	. , ’ "
Subordinating Conjunction	SCONJ	C, WQ
Verb	VERB	HV, HVI, HVPI, HVPS, HVDI, HVDS, BE, BEI, BEPI, BEPS, BEDI, BEDS, RD, RDI, RDPI, RDPS, RDDI, RN, VB, VBI, VBPI, VBPS, VBDI, VBDS, VG, VN
Other	X	FW
Symbol	SYM	-

Table 6: Mapping of HeliPaD to UPOS.

POS	UPOS	Menotec
Adjective	ADJ	A, Mo
Adposition	ADP	R
Adverb	ADV	Df, Du, S-
Auxiliary	AUX	AUX
Coordinating Conjunction	CCONJ	C
Determiner	DET	Pd, Py, Ps
Interjection	INTJ	I
Noun	NOUN	Nb
Numeral	NUM	Ma
Particle	PART	N
Pronoun	PRON	Pp, Pk, Pi, Px
Proper Noun	PROPN	Ne
Punctuation	PUNCT	. , ’ "
Subordinating Conjunction	SCONJ	G
Verb	VERB	V
Other	X	X, F
Symbol	SYM	-

Table 7: Mapping of Menotec to UPOS.

POS	UPOS	DDD-ReA 1.2
Adjective	ADJ	ADJ, ADJN, ADJNE, ADJE, ADJD, ADJDE, ADJO, ADJON, ADJOS, ADJS
Adposition	ADP	AP, APPO, APPR, APZR, PTKVZ, PTKINT
Adverb	ADV	ADV, ADVM, ADVNEG, ADVREL, PTKVZ, PWAV, PWAVREL, PWGAV, PWGAVREL
Auxiliary	AUX	VAIMP, VAINF, VAINFS, VAPP, VAPS, VMFIN, VMINF, VMINFS, VMPP, VMPS
Coordinating Conjunction	CCONJ	KON
Determiner	DET	DD, DDA, DDN, DDREL, DDS, DDSREL, DI, DIA, DIN, DINEG, DINEGN, DINEGS, DIS, DPOS, DPOSD, DPOSN, DPOSS, DW, DWG, DWGREL, DWREL, DWS, DWSREL
Interjection	INTJ	ITJ, PTKANT
Noun	NOUN	NA
Numeral	NUM	CARD, CARDN, CARDS
Particle	PART	PTK, PTKNEG, PTKZU
Pronoun	PRON	PI, PINEG, PPER, PRF, PW, PWG, PWGREL, PWREL, PTKREL
Proper Noun	PROPN	NE, NEO
Punctuation	PUNCT	$(, $,, $., $:, $;, $
Subordinating Conjunction	SCONJ	KOUS, KO, KOKOM
Verb	VERB	VV, VVFIN, VVIMP, VVINF, VVINFS, VVPP, VVPPA, VVPPD, VVPPN, VVPS, VVPSA, VVPSN, VVPSS, VVPPS
Other	X	?
Symbol	SYM	-

Table 8: Mapping of the Deutsch Diachron Digital, Referenzkorpus Altdeutsch (Version 1.2) to UPOS.

Old English			
BiGram	**Prob.**	**TriGram**	**Prob.**
('NOUN', 'PUNCT')	6.88%	('NOUN', 'VERB', 'PUNCT')	3.76%
('VERB', 'PUNCT')	5.77%	('ADJ', 'NOUN', 'PUNCT')	2.01%
('NOUN', 'VERB')	5.58%	('ADP', 'NOUN', 'PUNCT')	1.95%
('ADP', 'NOUN')	5.04%	('NOUN', 'NOUN', 'PUNCT')	1.65%
('NOUN', 'NOUN')	5.04%	('NOUN', 'ADP', 'NOUN')	1.49%
Old Norse			
BiGram	**Prob.**	**TriGram**	**Prob.**
('NOUN', 'PUNCT')	5.06%	('NOUN', 'VERB', 'PUNCT')	2.10%
('NOUN', 'VERB')	4.63%	('VERB', 'ADP', 'NOUN')	1.39%
('VERB', 'PUNCT')	4.52%	('ADP', 'NOUN', 'PUNCT')	1.16%
('ADP', 'NOUN')	4.01%	('NOUN', 'NOUN', 'PUNCT')	1.12%
('VERB', 'PRON')	3.50%	('ADJ', 'NOUN', 'PUNCT')	0.91%
Old High German			
BiGram	**Prob.**	**TriGram**	**Prob.**
('VERB', 'PUNCT')	5.79%	('DET', 'NOUN', 'PUNCT')	2.15%
('DET', 'NOUN')	5.53%	('NOUN', 'VERB', 'PUNCT')	1.51%
('NOUN', 'PUNCT')	4.68%	('ADV', 'VERB', 'PUNCT')	1.37%
('PRON', 'ADV')	3.13%	('PRON', 'VERB', 'PUNCT')	1.29%
('PRON', 'VERB')	2.97%	('ADP', 'DET', 'NOUN')	1.19%
Old Saxon			
BiGram	**Prob.**	**TriGram**	**Prob.**
('NOUN', 'PUNCT')	6.05%	('NOUN', 'VERB', 'PUNCT')	2.67%
('DET', 'NOUN')	5.42%	('ADP', 'DET', 'NOUN')	2.60%
('VERB', 'PUNCT')	4.18%	('ADJ', 'NOUN', 'PUNCT')	1.69%
('NOUN', 'VERB')	3.95%	('PUNCT', 'SCONJ', 'PRON')	1.59%
('VERB', 'PRON')	2.90%	('DET', 'NOUN', 'PUNCT')	1.43%

Table 9: Five most frequent POS bigrams and trigrams, along with their frequencies, extracted from the reduced training datasets for each language in the **poetry** corpus.

Old English			
BiGram	**Prob.**	**TriGram**	**Prob.**
(’DET’, ’NOUN’)	5.42%	(’ADP’, ’DET’, ’NOUN’)	2.12%
(’NOUN’, ’PUNCT’)	4.69%	(’DET’, ’ADJ’, ’NOUN’)	1.49%
(’VERB’, ’PUNCT’)	3.88%	(’NOUN’, ’VERB’, ’PUNCT’)	1.46%
(’ADP’, ’DET’)	3.43%	(’DET’, ’NOUN’, ’PUNCT’)	1.46%
(’PRON’, ’VERB’)	3.26%	(’DET’, ’NOUN’, ’VERB’)	1.09%
Old Norse			
BiGram	**Prob.**	**TriGram**	**Prob.**
(’NOUN’, ’PUNCT’)	4.27%	(’SCONJ’, ’PRON’, ’VERB’)	1.65%
(’ADP’, ’NOUN’)	4.01%	(’VERB’, ’ADP’, ’NOUN’)	1.32%
(’VERB’, ’PRON’)	3.06%	(’ADP’, ’NOUN’, ’PUNCT’)	1.31%
(’PRON’, ’VERB’)	2.81%	(’NOUN’, ’ADP’, ’PUNCT’)	1.10%
(’ADV’, ’VERB’)	3.06%	(’PUNCT’, ’CCONJ’, ’VERB’)	0.94%
Old High German			
BiGram	**Prob.**	**TriGram**	**Prob.**
(’NOUN’, ’PUNCT’)	6.34%	(’DET’, ’NOUN’, ’PUNCT’)	2.36%
(’DET’, ’NOUN’)	6.14%	(’ADP’, ’DET’, ’NOUN’)	2.05%
(’VERB’, ’PUNCT’)	4.44%	(’VERB’, ’DET’, ’NOUN’)	1.16%
(’VERB’, ’PRON’)	2.83%	(’VERB’, ’PRON’, ’PUNCT’)	1.13%
(’PRON’, ’VERB’)	2.82%	(’PRON’, ’VERB’, ’PUNCT’)	1.01%
Old Saxon			
BiGram	**Prob.**	**TriGram**	**Prob.**
(’NOUN’, ’DET’)	5.67%	(’NOUN’, ’DET’, ’PUNCT’)	2.28%
(’ADP’, ’NOUN’)	4.86%	(’VERB’, ’NOUN’, ’DET’)	1.19%
(’NOUN’, ’PUNCT’)	4.37%	(’ADP’, ’NOUN’, ’NOUN’)	1.05%
(’VERB’, ’AUX’)	3.08%	(’NOUN’, ’PUNCT’, ’CCONJ’)	1.04%
(’PUNCT’, ’CCONJ’)	2.79%	(’ADP’, ’NOUN’, ’DET’)	1.04%

Table 10: Five most frequent POS bigrams and trigrams, along with their frequencies, extracted from the reduced training datasets for each language in the **prose** corpus.

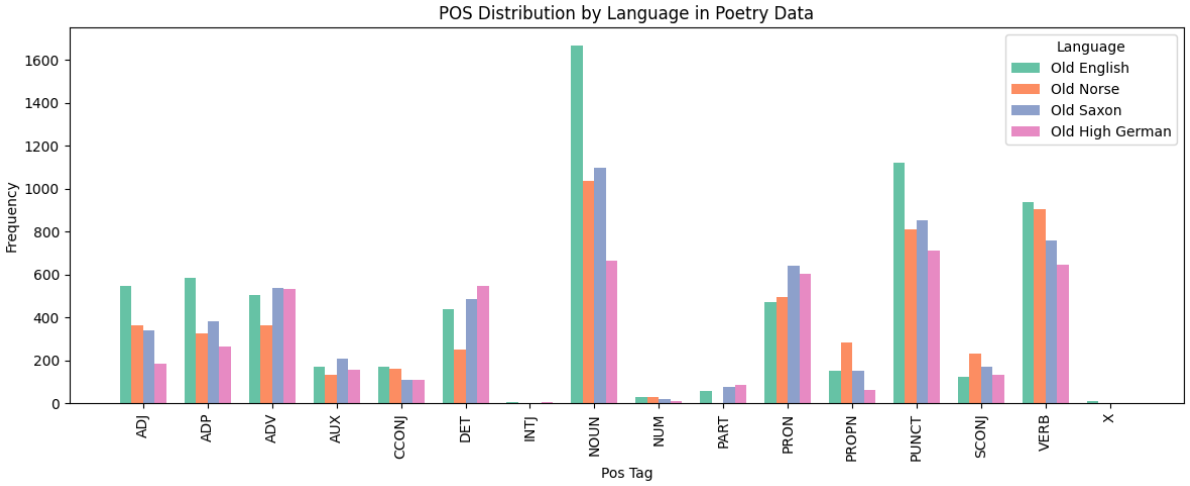

Figure 1: Distribution of part-of-speech (POS) tag frequencies in **poetry** reduced datasets from four historical languages: Old English, Old Norse, Old High German, and Old Saxon.

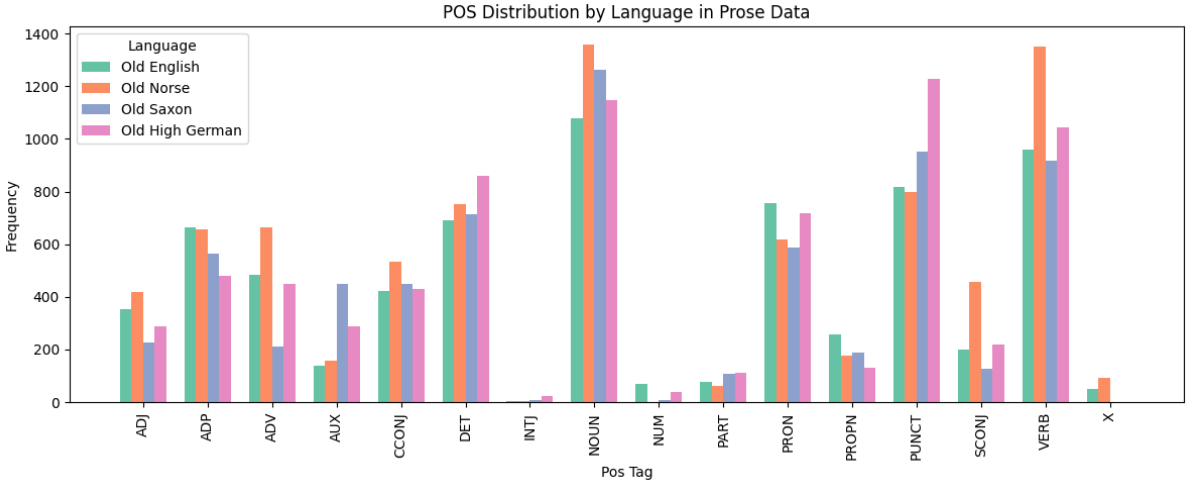

Figure 2: Distribution of part-of-speech (POS) tag frequencies in **prose** reduced datasets from four historical languages: Old English, Old Norse, Old High German, and Old Saxon.