# Cross-Genre Learning for Old English Poetry POS Tagging

**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

Poetry has always distinguished itself from other literary genres in many ways, including grammatically and syntactically. These differences are evident not only in modern literature but also in earlier stages. Linguistic analysis tools struggle to address these differences. This paper focuses on the dichotomy between Old English poetry and prose, specifically in the context of the POS tagging task. Two annotated corpora representing each genre were analyzed to show that there are several types of structural differences between Old English poetry and prose. For POS tagging, we conduct experiments on both a detailed tag set with over 200 tags and a mapping to the UPOS tag set with 17 tags. We establish a baseline and conduct two cross-genre experiments to investigate the effect of different proportions of prose and poetry data. Across both tag sets, our results indicate that if the divergence between two genres is substantial, simply increasing the quantity of training data from the support genre does not necessarily improve prediction accuracy. However, incorporating even a small amount of target data can lead to better performance compared to excluding it entirely. This study not only highlights the linguistic differences between Old English poetry and prose but also emphasizes the importance of developing effective NLP tools for underrepresented historical languages across all genres.

## 1 Introduction

Poetry has always stood apart from other genres, and poetic language differs from other genres on several levels, including those of syntax and grammar. There is a tendency to use incomplete sentences, omit finite verbs, or deviate from standard word order. These choices appear to be motivated by the desire to emphasize specific connections of words or trigger specific emotions in the reader (Nofal, 2011). The adoption of different constructions across genres is a phenomenon that shapes not only modern literary traditions but also those of the past. This is the case of Old English poetry, which has been the focus of studies highlighting its structural, syntactical, and grammatical differences from Old English prose. The dichotomy between the two genres lies in several aspects; for instance, significant emphasis is placed on the types of clauses—whether principal or subordinate—employed in the poems (Mitchell, 1985). Being able to recognize the characteristics of each genre is essential to properly analyze a text.

Linguistic analysis is fundamental for examining and identifying the characteristics of different genres. Several tools have been developed to ease this process, such as Part-of-Speech (POS) tagging tools, which have benefited from significant technological advancements and improvements over time. The development of these tools has also a few shortcomings. It has been shown that modern POS taggers struggle to shift between different genres and offer accurate predictions (Arai, 2021). One possible reason for this limitation is the uneven distribution of data across genres within the corpora. The solutions proposed often involve the addition of new or synthetic data to help refine the performance of these tools (Arai, 2021). These practices are more easily implemented in a high-resource language setting. However, this is not always a suitable approach for older languages that typically have less data. In addition to limited data resources, some languages, such as Old English, have been comparatively underrepresented in POS-tagging research. Old English poetry, in particular, is even less represented in this body of research. Addressing the issue of domain shift between genres in support tools for modern languages is essential for reliable tools with all texts; equally important is the focus on older languages, which form the bedrock of human history, offering insights into interactions between past civilizations and helping to preserve our cultural heritage (van



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 708–724

Gelderen, 2014). In addition, old languages are a topic of interest for many scholars and students who need to have tools with accurate performance as a support for their studies.

This paper explores POS-tagging for Old English poetry and investigates cross-genre learning to address the challenge of domain shift. To do that, two corpora with Old English poetry and prose have been used to establish a baseline for this task. Two experiments were then conducted to investigate the impact of mixing poetry and prose training data in different proportions. Because of the high number of labels in the original tag sets and the slight differences between the tag sets of the two corpora, we have also converted the labels used by both corpora to the Universal Dependencies UPOS tag set (de Marneffe et al., 2021). The paper will present the results for both the original tag sets and the UPOS tag set. Section 2 will present an overview of the related work. Section 3 will present the datasets, the POS mapping, and a series of structural analyses to investigate further the differences between the two genres. Experimental setups will be presented in Section 4. Section 5 will present and analyze the results. Conclusions will be discussed together with future work suggestions in Section 6.

## 2 Related Work

Specific studies on POS tagging tools for Old English poetry appear to be lacking, with only one known POS tagger currently available for Old English. The tagger is part of the CLTK library (Johnson et al., 2021), and has been trained on the available texts from the ISWOC Treebank (Bech and Eide, 2014). While the tool provides several model options, their accuracy remains uncertain.

While there is a lack of studies in this particular area, as noted, there are several studies that explore domain shift issues in POS taggers for historical English. Rayson et al. (2007) highlighted the low performance of existing Modern English POS taggers on Early Modern English datasets. Their study showed that handling orthographic variations increases accuracy. In the same year, Moon and Baldridge (2007) investigated ways to implement a POS Tagger for historical languages based on existing resources from their modern varieties. They used Modern English resources to tag Middle English data using alignments on parallel Biblical texts. The results were promising, but the accuracy of the manually annotated training set was not outperformed. Domain adaptation techniques were the focus of Yang and Eisenstein (2016) who evaluated several methods for the task of POS tagging for Early Modern and Modern British English texts. The combination of FEMA, domain adaptation algorithm designed for sequence labeling problems, and normalization techniques, improved the performances. A few years later, Karimov (2018) focused his attention on Middle English corpora and historical texts. To handle the irregular word order in older English, he applied a moving-average method to generate multidimensional vectors, capturing both character composition and weighted positions. 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 face the problem.

## 3 Data and Tag Sets

The paper aims to establish a baseline for Old English poetry POS taggers and investigate cross-genre learning scenarios. Two corpora were used to train the models:

- the York-Helsinki Parsed Corpus of Old English Poetry (YCOEP) (University of Oxford, 2001): selection of poetic texts from the Old English section of the Helsinki Corpus of English Texts.
- the York Toronto Helsinki Parsed Corpus of Old English (YCOE) (University of Oxford, 2003): syntactically annotated corpus with all the major Old English prose works.

Since the official documentation for the YCOEP dataset is unavailable, the YCOE documentation (University of Oxford, 2003) was adopted as the primary reference for both corpora.

The texts of the corpora are segmented into units called "tokens", which consist of one main verb (or verb sequence) along with all associated arguments and adjuncts. The "tokens" can represent matrix inflectional phrases, complementizer phrases, or independent non-clausal utterances. Each "token" is enclosed in a "wrapper": a pair of unlabeled parenthesis including the parsed text and the identifying metadata (University of Oxford, 2003). From the corpora, the original textual form of each "token", along with words and POS tags, was extracted and converted into CoNLL format, data format supported by MaChAmp, the toolkit for multi-task learning used to train all the models.

### 3.1 POS Mapping

Both YCOE and YCOEP datasets contain a substantial number of POS tags: 201 in the poetry dataset and 289 in the prose dataset. This extensive number of labels offers highly detailed linguistic information (i.e. grammatical features, inflectional features, morphological features); at the same time, it can pose significant challenges for both manual annotation and automated processing. A further complication arises from the inconsistencies between the two tag sets: despite originating from the same project (University of Oxford, 2003) and describing the same language variety, only 173 labels are common to both datasets. Our analysis revealed that the differences can be related to:

- potential spelling errors in the tags;
- discrepancies in linguistic categorization, such as the distinction between comparative and superlative use, which is present in the prose but missing in the poetry; this affects adjectives, adverbs, and quantifiers;
- missing tags, such as *MAN*, present in the YCOE dataset, but not in the YCOEP, is frequently used as a pronoun;
- inconsistencies in tag naming conventions, such as proper nouns labeled as *NPR* in the poetry dataset and as *NR* in the prose one.

The large number of tags and the discrepancies between the two tag sets may negatively impact the performance of the models. For this reason, and to facilitate the structural analysis, both YCOE and YCOEP tag sets were mapped to the Universal Dependencies UPOS tag set (de Marneffe et al., 2021), a widely adopted and standardized POS framework. We will report results for both the original and the UPOS tag set. Table 5, in Appendix A, presents the complete mapping from the original tag sets to the UD categories. For the majority of the tags, the conversion to UPOS was straightforward, but a subset of Old English labels required specific rules for the conversion.

Prepositions, a closed class in both Old and Modern English, exhibit diverse syntactic behaviors in the original annotation scheme, leading to multiple tags. When prepositions are used with a complement, they are tagged as such and mapped to the UD category *ADP* (adposition). When no complement is present, they are annotated as adverbs or adverbial particles, and accordingly mapped to the UD category *ADV* (adverb). Furthermore, certain prepositions appear to be able to function also as subordinate conjunctions, which can complicate the effort to extract a clean closed class. For this reason, only complementizers and the word *'whether'* were mapped to the UD category *SCONJ* (subordinating conjunction).

Participles also pose a conversion challenge. Although they often function adjectivally, neither the YCOE nor the YCOEP tags them as *ADJ*. However, the case is a fully productive category in Old English that can be applied to nouns, adjectives, quantifiers, determiners, numbers, and participles (University of Oxford, 2001). For this reason, when participles display a case, instead of the corresponding participle tag, they will be tagged as *ADJ*.

The original tag set has specific labels for auxiliaries; however, *be* and *have* are always tagged as verbs, even when they function as auxiliaries. To more accurately reflect their syntactic role, we introduced a rule-based refinement: *be* and *have* will be labeled as *AUX* (auxiliary) when (i) followed by another verb, or (ii) followed by a subject (noun, proper noun, or pronoun) and another verb. Future work will aim to identify additional syntactic environments in which *be* and *have* fulfill auxiliary functions but are not annotated as such.

Some POS tags, particularly for verbs, adverbs, and quantifiers, include additional markers such as *RP+* or *NEG+*, respectively indicating the presence of adverbial particles or contracted negative forms. In such cases, the suffix tags are removed, and the token is assigned its core POS tag.

The UPOS mapping led to a decrease in the number of POS tags from over 200 to 17. By adopting this conversion, datasets and POS tags are more easily comparable and can be used to train the models. However, the conversion loses the linguistic granularity that was part of the original tag set such as grammatical features (i.e. case, gender, number, etc.). Other tag set variants could have retained more linguistic information; the exploration of different approaches is left for future work.

### 3.2 Structural Analysis of the Genres

To assess the structural differences between Old English poetry and prose, we conducted a series

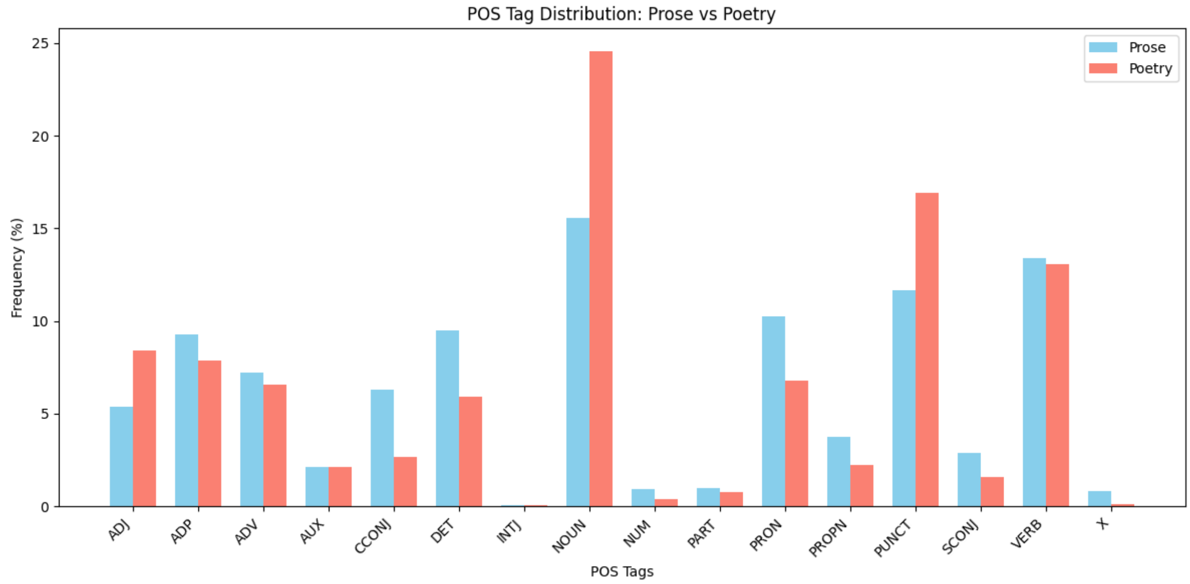


Figure 1: Distribution of POS tag frequencies in samples from the UPOS mapped versions of the YCOEP (poetry) and YCOE (prose) datasets.

of analyses on POS tag distributions using data from the UPOS-mapped versions of the YCOE (prose) and YCOEP (poetry) corpora. The size of the samples used, 5668 sentences, corresponds to the training and development sets employed in the training of the baseline models.

We began by analyzing the frequency of each POS tag. Figure 1 shows a comparison between the frequencies of each tag for both poetry and prose. For both genres, nouns, punctuation, and verbs are the most common tags. Nouns are much more frequent in the poetry compared to the prose, with a difference of approximately 10%. Punctuation is similarly more frequent in poetry, while verbs have similar frequencies. The distribution of the POS tags suggests that, in the poetry, the frequency of content words is higher than that of function words. Prose also shows this behavior, but the gap appears to be smaller. Overall, the prose distribution appears more balanced than that of poetry, suggesting that poetry contains more complex structures. We also extracted the sentence-level POS tag sequences across the two corpora: there are 4814 unique sequences in the poetry and 5195 in the prose. Notably, only 90 are shared by both genres. This low number of overlaps between the two datasets highlights the substantial structural difference between Old English poetry and prose, and the importance of considering it when training models.

A closer look at these differences is given in Table 6, in Appendix A, which displays, for both genres, the ten most common POS bigrams and trigrams, along with their probabilities. Among the bigrams, only five are common to both corpora. These shared bigrams have higher probabilities in the poetry data except for ('DET', 'NOUN'), which rank as the second most frequent pair in the prose data. The ('NOUN', 'NOUN') bigram is particularly interesting, as it does not represent compounds—written as single words in Old English texts (University of Oxford, 2003)—yet it has one of the highest probabilities in the poetry sample (4.59%). It is also present in the prose data but with a lower probability (1.21%). Nouns and punctuation are the most frequent elements in the poetry bigrams: appearing respectively in eight and four pairs. In prose, the most common are verbs and nouns present in five and four bigrams. The high frequencies of these tags are not a surprise if we consider the POS tags distribution presented in Figure 1. This also supports the supposition about the higher frequency of content words in the poetry.

Regarding trigrams, only two are shared between the datasets, and as for the bigrams, these common combinations have higher probabilities in the poetry data. Also, in this case, nouns, punctuation, and verbs have higher frequencies. In the poetry results, nouns are present in each trigram. Punctuation tags increase, appearing seven times. The distribution of the POS tags in the poetic trigrams seems to indicate the presence of more frag-

| UniGram Model | | |
|---|---|---|
| **Genre** | **Poetry Test Set** | **Prose Test Set** |
| Poetry | 9.603 | 12.353 |
| Prose | 10.352 | 11.377 |
| **BiGram Model** | | |
| **Genre** | **Poetry Test Set** | **Prose Test Set** |
| Poetry | 9.093 | 11.349 |
| Prose | 10.464 | 9.866 |
| **TriGram Model** | | |
| **Genre** | **Poetry Test Set** | **Prose Test Set** |
| Poetry | 7.428 | 9.5 |
| Prose | 8.862 | 7.994 |

Table 1: Perplexity Scores for N-Gram Models on samples from the UPOS mapped versions of the YCOEP (poetry) and YCOE (prose) datasets. The Genre column indicates the genre of the training data. Poetry and Prose Test Set display the perplexity score of the model computed on the corresponding test test.

mented constructions. The prose sample, on the other hand, shows an increasing number of trigrams with nouns and a consistent amount of verbs. Even with the poetic trigrams, we can observe a stronger presence of content words. Functional words are more present in the trigrams, but not at the same level as in the prose ones.

To further investigate the genre differences, we calculated the perplexity scores for unigrams, bigrams and trigrams of two poetry and prose test sets with two models: one trained only with poetry data and one only with prose data. For both datasets, 6299 sentences (i.e. the amount of poetry data) were extracted and divided into training set (80%), development set (10%) and test set (10%). All the models were implemented using the Language Model module from the Natural Language Toolkit (NLTK) (Bird et al., 2009). Laplace smoothing was used to ensure non-zero probabilities for unseen sequences. Perplexity was computed using the NLTK's built-in function. The results are presented in Table 1.

As expected, both models exhibited lower perplexity scores when evaluated on their own test set, and higher scores when evaluated on the other genre. A general trend across all models is the decrease in perplexity with increasing n-gram size: as more context is incorporated, the predicting abilities of the models improve. In addition, the difference between the in-genre and out-of-genre training increases with n-gram size. This indicates that the differences between the two genres are more pronounced with higher-order n-grams, being consistent with observations about structural differences between poetry and prose (Nofal, 2011; Mitchell, 1985). These findings highlight the importance of taking into account these variations when developing NLP tools.

## 4 Experimental Setup

The poetry dataset, YCOEP, contains 6299 sentences. For the baseline model, the poetry data were divided into training set (80%), development set (10%), and test set (10%). This resulted in 5039 sentences for training, 629 for development, and 631 for test sets. To create a comparable dataset for the prose genre, a subset of the YCOE corpus was selected: a sample of 5668 sentences—matching the combined size of the poetry training and development sets.

In our first experiment, we investigated the models' performances in a scenario of limited target genre data combined with a greater amount of support data. In this experiment, the same data used to train the poetry-only baseline model were used as target genre data. The support data consisted of progressively larger subsets of prose data, up to the full prose dataset consisting of 109,703 lines. The prose data was always only divided into a training set (90%) and a development set (10%).

Our second experiment was designed with two primary objectives: (i) to determine the minimum amount of target genre data required to maintain acceptable model performance, and (ii) to examine the impact of progressively reducing the amount of target genre data while keeping the quantity of support genre data constant. In this experiment, the prose data used to train the baseline models was used as support genre data. The amount of poetry data was progressively decreased until it reached 57 sentences; the data were divided into a training set (90%) and a development set (10%).

All models were trained with MaChAmp, a toolkit for multi-task learning and fine-tuning offering a wide variety of tasks. It offers an easy configuration, especially for dealing with multiple datasets, together with a wide range of NLP tasks (i.e., POS tagging, text classification, etc.). It utilizes a shared pre-trained encoder, which is fine-tuned during training. Each task is equipped with its decoder (van der Goot et al., 2021). For our experiments, we employed the *seq* task type, for which MaChAmp applies a greedy softmax

| | OG Tag Set | | UPOS Tag Set | |
|---|---|---|---|---|
| **Genre** | **Acc.** | **F1** | **Acc.** | **F1** |
| poe (5668) | 0.909 | 0.708 | 0.961 | 0.944 |
| pro (5668) | 0.762 | 0.464 | 0.879 | 0.840 |
| poe, pro (11,336) | 0.917 | 0.707 | 0.966 | 0.948 |

Table 2: Results for baseline POS taggers trained with the Original (OG) tag set and the Universal Dependencies (UD) tag set. The models were tested on a poetry test set. The data belong to either poetry, prose genres, or a combination of both.

classification layer over the contextualized token embeddings provided by the encoder. All the models were based on multilingual BERT, the default language model in MaChAmp, and trained with default hyperparameters. Each model was trained for 20 epochs with three different random seeds. The evaluation was performed primarily on the poetry test set from the original dataset split; in addition, a prose test set was used to evaluate the baseline models' performance on the opposite genre. For each seed, we computed accuracy and macro F1 score across all tags; the results will report the average performance over the three seeds.

## 5 Results

Tables 2, 3, and 4 present the results for the baseline models, the first experiment, and the second experiment, respectively, for both tag sets. Appendix A additionally includes the evaluation of the baseline models on the prose test set (Table 7).

### 5.1 Baseline

Table 2 reports the results obtained from the baseline models. The first model (*poe*) was trained only on poetry data, and despite relying on the smallest dataset, it showed strong performance with both tag sets. By reducing the number of tags from 200 to 17, both accuracy and F1 score values increase. This approach helps reduce the number of rare classes leading to more informative results, but at the same time, a deeper level of linguistic information is lost.

The second model (*pro*) was trained solely on prose data and evaluated on poetry data. Compared to the first model, the performances across both tag sets, drop significantly. With the original tag set, model accuracy declines from 90% to 76%, accompanied by a decrease in F1 score from 70% to 46%. The same trend is observed with the UPOS tag set, although the decline is less pronounced. This behavior can be explained by the different syntactical structures of the two genres. As it has been shown in section 3.2, the distribution of the POS tags in the prose differs significantly from the poetry one; these differences are so broad that the model is not able to learn to correctly predict the poetry POS tags.

The third model (*poe, pro*) is trained with data from both genres, which results in the largest dataset (11,336 sentences) among the three. This model has better performances than the second, but not compared to the first: the second model is outperformed because of the presence of the target genre which is missing from the second model. Compared to the first model, there is only a marginal improvement in accuracy and almost no change in the F1 score. One might expect to have higher results with a larger dataset, but this is not the case. Even with the same amount of target and support data, the differences between the two genres are too broad for the model to learn information suitable to tag data from the target genre.

Table 7 reports the evaluation of the baseline models on the prose test set. The model trained solely on the target genre (i.e. the *pro* model here) achieves better results than the one trained only with support data (i.e. the *poe* model in this case). This is consistent with the results and findings from the poetry test set evaluation. Despite the larger dataset size, the combined *poe, pro* model does not outperform the *pro* model, suggesting that the differences between the genres are too broad to provide useful additional information. Notably, results on prose are slightly higher than on poetry, indicating possible asymmetry between genres as also suggested by their POS tag distributions (Figure 1). Poetry, less balanced and structurally more complex, requires more robust training and is harder to predict, while prose's simpler, more balanced patterns lead to higher performance.

### 5.2 Limited Target Data and Increasing Support Data Scenario

The first experiment involves a constant amount of poetry (5668 lines) combined with progressively larger subsets of prose data, up to the full prose dataset consisting of 109,703 lines. The results of the experiment are presented in Table 3.

Consistent with the findings from Table 2, the UPOS tag set has higher scores than the original

| | OG Tag Set | | UPOS Tag Set | |
|---|---|---|---|---|
| Size | Acc. | F1 | Acc. | F1 |
| *0* | *0.909* | *0.708* | *0.961* | *0.944* |
| 1417 | 0.916 | 0.705 | 0.962 | 0.944 |
| 2834 | 0.916 | 0.698 | 0.964 | 0.946 |
| *5668* | *0.917* | *0.707* | *0.966* | *0.948* |
| 11,336 | 0.919 | 0.692 | 0.966 | 0.948 |
| 22,672 | 0.917 | 0.680 | 0.969 | 0.953 |
| 34,008 | 0.921 | 0.676 | 0.970 | 0.951 |
| 45,344 | 0.920 | 0.676 | 0.969 | 0.949 |
| 56,680 | 0.920 | 0.697 | 0.969 | 0.945 |
| 68,016 | 0.923 | 0.684 | 0.969 | 0.945 |
| 79,352 | 0.921 | 0.684 | 0.970 | 0.942 |
| 90,688 | 0.921 | 0.672 | 0.969 | 0.944 |
| 109,703 | 0.921 | 0.688 | 0.970 | 0.942 |

Table 3: Results for the first experiment. In this experiment, the amount of poetry is consistent (5668 lines) while the amount of prose increases systematically. The Size column indicates the amount of prose added to the dataset. *Italic* is used to indicate the baseline results.

| | OG Tag Set | | UPOS Tag Set | |
|---|---|---|---|---|
| Size | Acc. | F1 | Acc. | F1 |
| *5668* | *0.917* | *0.707* | *0.966* | *0.948* |
| 4534 | 0.913 | 0.676 | 0.964 | 0.947 |
| 3779 | 0.908 | 0.661 | 0.961 | 0.939 |
| 2834 | 0.897 | 0.626 | 0.957 | 0.934 |
| 1889 | 0.883 | 0.598 | 0.949 | 0.930 |
| 945 | 0.857 | 0.554 | 0.936 | 0.923 |
| 472 | 0.824 | 0.519 | 0.919 | 0.907 |
| 227 | 0.802 | 0.490 | 0.904 | 0.891 |
| 113 | 0.784 | 0.482 | 0.893 | 0.878 |
| 57 | 0.775 | 0.475 | 0.889 | 0.867 |
| *0* | *0.762* | *0.464* | *0.879* | *0.840* |

Table 4: Results for the second experiment. The amount of prose data is set to 5668 lines, while the amount of poetry decreases. The Size column indicates the amount of poetry for each model. *Italic* is used to indicate the baseline results.

one, especially for what concerns the F1 score. Accuracy also improves, but the difference is notably smaller than the one observed for the other measure.

With both tag sets, independently of the amount of prose data, the accuracy increases slightly compared to the poetry-only model (i.e. size 0 model). The F1 score is more or less consistent with the UPOS tag set, but it declines more with the original tag set. As for the baseline models, we might expect outperforming results as the dataset size increases, but this is not happening. Even the last model, trained with the largest dataset (109,703 lines) has either lower results than the baseline (OG tag set) or almost the same values (UPOS). These results suggest that indiscriminately increasing training data is not a universally effective strategy: the intrinsic differences between the two genres could be too diverse for the model to learn properly the patterns.

Interestingly, the models trained with smaller subsets of prose data—comprising 1417, 2834, and 5668 lines—have slightly higher results than those trained with larger amounts of prose. This finding could signal that a limited quantity of support data could contribute to the training of the model. It could be the case that selecting a smaller quantity of data with similar patterns to the target genre, could refine the predictions without overwhelming the target genre's patterns. Future studies will focus on this finding.

### 5.3 Decreasing Target Data and Consistent Support Data Scenario

Table 4 presents the results for the second experiment: the amount of prose data remains constant (5668 lines), while the amount of poetry data is progressively reduced across models.

For both tag sets, accuracy, and F1 score values decline as the size of the poetry data decreases. The decline is more pronounced with the OG tag set, especially for the F1 score, which drops by 23% points compared to the 70% of the *poe, pro* baseline model. This progressive decline is again an indication of the differences between the two genres. When the proportion of target data decreases, the model has fewer genre-specific patterns to learn from; thus, the model struggles to predict unseen patterns. However, it is noteworthy that even the model trained with the smallest amount of poetry data—only 57 lines—achieves slightly better performances than the baseline model trained with only prose data (i.e. size 0 model). This finding emphasizes the importance of the target genre in the training data. Even in a minimal amount, the target genre can improve the performance of the model, suggesting that the specific features of a genre cannot be learned even from large quantities of out-of-genre data.

### 5.4 Tag-Level Error Analysis

Appendix A presents the normalized confusion matrices averaged over the three seeds for the baseline models evaluated on the poetry test set, as well as those evaluated on the prose test set. It includes also two key models from both experiments.

Figure 2 presents the results for the *poe* baseline model. ADJ, ADV, and X are the tags with the lowest scores: ADJ is primarily confused with NOUN and VERB, while ADV is misclassified across eight other tags. This suggests model uncertainty, probably related to its medium-to-low frequency in the dataset. X is confused with ADJ, NOUN, and VERB; but it has a very low frequency, resulting in a lack of training data. The *pro* baseline model (Figure 3) shows similar misclassification patterns. ADJ, ADV, and X remain among the most confused tags; in addition, the model wrongly assigns AUX, NOUN, PROPN, and VERB. AUX is misclassified mostly with VERB, which may be related to the mapping choices described in Section 3.1. Unlike in the *poe* model, NOUN is frequently misclassified, possibly due to its lower frequency in the prose compared to the poetry. This reduces its available training data, worsening the model's performance. VERB is misclassified mainly with ADJ and NOUN, with smaller errors with other six tags. Figure 4 shows the results for the combined *poe, pro* baseline model. ADJ, ADV, and X still have lower scores, but overall results are slightly higher compared to the poetry-only model. The plot supports earlier findings: combining target and support genres slightly helps the model to generalize because of the increased diversity in the training data. However, the improvements remain very modest relative to the much larger dataset size (11,336 sentences).

Figure 5 presents the results for the *poe* baseline model tested on the prose test set. ADJ, ADV, and X remain among the main misclassified tags, along with AUX, INTJ, NOUN, NUM, PART, and SCONJ. According to the POS tag distributions (Figure 1), many of these tags present significant frequency differences between prose and poetry: the lack of data per tag in the training data may be the cause of the model's uncertainty. Overall, the *poe* model performs better on prose than the *pro* model does on poetry, supporting the presence of an asymmetry between genres. Poetry's complex structures require more robust learning, while prose patterns are more balanced and predictable, increasing the model's performance. This appears to be also supported by the scores in Figure 6: the *pro* baseline model tested on the prose test set has higher values than the *poe* model tested on the same genre test set. This is most likely related to the simpler and more predictable patterns present in the prose. ADJ is still a frequently misclassified class, together with INTJ and NUM. Figure 7 presents the results for the combined *poe, pro* model tested on prose. Consistently with the previous results, the performances are slightly better than Figure 4, supporting the idea of an asymmetry between the genres. ADJ, INTJ, and X are still challenging tags.

Figure 8 and 9 present key models from each experiment. Figure 8 shows results for the model trained with a fixed amount of poetry (5668 sentences), and the entirety of the prose data (109,703 sentences) from the first experiment (Section 5.2). ADJ, ADV, and X remain lower-scoring tags, but overall, the performances improve compared to baseline models. Because of the large dataset size, the model is trained on a very diverse training set, which leads to refined predictions. However, as for the *poe, pro* baseline model, the results are disproportionately small compared to the amount of data provided, supporting earlier findings that larger dataset sizes do not ensure the best results. Figure 9 reports results for the model trained with a fixed amount of prose (5668 sentences) and minimal poetry (57 sentences) from the second experiment (Section 5.3). The misclassified tags are the same as for the *pro* baseline model (i.e. ADJ, ADV, X, AUX, NOUN, PROPN, and VERB). Nonetheless, overall scores are slightly higher, suggesting that even small amounts of target data in the training set can strengthen the model's performance, as previously observed.

Overall, the error analysis supports previous findings, reinforcing the notion of an asymmetry between Old English poetry and prose, which can be somewhat mitigated by the combination of target and support data. Selecting an appropriate dataset size also proves to be relevant. Across all plots, ADJ, ADV, and X consistently emerge as the most challenging tags for the models. A deeper, more detailed qualitative analysis could reveal hidden patterns and provide explanations for these and other misclassifications; such analysis is left for future work.

## 6 Conclusions and Future Work

The study explores the differences between Old English poetry and prose, focusing on the POS tagging task. Two datasets, YCOE and YCOEP, were mapped to the UPOS tag set and used to establish a baseline and conduct two cross-genre experiments. Additionally, a series of analyses of the distributions of the POS tags within the sentences of both datasets have been conducted to investigate the differences between the two genres.

Baseline results suggested an asymmetry between target and support genres, causing the model to struggle to predict the correct target POS tags. This limitation was also present when the training data included the same amount of target and support data, suggesting that quantity cannot account for genre-specific patterns in the data.

The first experiment involved a constant amount of target data combined with an increasing amount of support data. Results showed that indiscriminately enlarging the training data is not always an effective solution. If the divergence between the two genres is substantial, selecting the largest amount of support data could simply lead to the same performance as the absence of the support data. Conversely, selecting a smaller and more controlled amount of support data could result in more refined performances.

The second experiment fixed the amount of support data while gradually decreasing the target data. As expected, the performance of the models declined as the target data was reduced: the model had fewer genre-specific instances to learn from, so it was unable to correctly predict unseen target data. However, even a minimal amount of target data can result in better performance compared to the complete absence of the genre itself.

The error analysis revealed that certain tags, ADJ, ADV and X, consistently challenge all models. It also reinforced earlier findings by highlighting the asymmetry between genres and emphasizing the importance of dataset size.

These findings highlight the necessity of developing linguistic analysis tools able to handle a wide range of genres with equal proficiency. Moreover, this study contributes to the development of more robust NLP tools for underrepresented historical languages and supports broader efforts to preserve and analyze linguistic heritage.

Future research will focus on selecting small support datasets that mirror the sentence-level POS tag sequences in the target data. In addition, it will include qualitative analyses of the predictions to uncover hidden patterns and better understand the models' errors. Since this paper explores only data concatenation for combining data from different genres, future work will investigate more advanced methods such as multi-lingual learning or treebank embeddings (Stymne et al., 2018). In future works, we aim to investigate further ways to deal with historical, low-resource languages. Additional underrepresented historical languages and other tasks relevant to the linguistic analyses will also be taken into consideration.

## 7 Limitations

This study offers insight into the linguistic differences between Old English poetry and prose, and how these differences can affect linguistic analysis tools, such as POS taggers. In doing so, it also encounters some limitations.

Firstly, Old English is a morphologically rich language, and the granularity of the original tag sets reflects this complexity. As a result, losing linguistic information when converting these detailed tags to UPOS is inevitable. While we made an effort to map the original tags in a reliable way, there may still be conversion errors influencing the UPOS quality. Additionally, the study relies solely on combining data from different genres as a method of concatenation; future work will investigate alternative approaches. Secondly, the models were trained with MaChAmp default hyperparameter settings. A more focused investigation into hyperparameter optimization could influence the models' performances, especially given the unique characteristics of Old English poetic data.

## References


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

Kristin Bech and Kristine Eide. 2014. The iswoc corpus. Department of Literature, Area Studies and European Languages, University of Oslo. Accessed: 2025-04-22.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc., Sebastopol, CA.

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

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.

Raoul Karimov. 2018. Combined machine-learning approach to pos-tagging of middle english corpora. *Crossroads. A Journal of English Studies*, 21:42–52.

Bruce Mitchell. 1985. *Old English Syntax*, volume 1. Clarendon Press, Oxford.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 390–399. Association for Computational Linguistics.

Khalil Hassan Nofal. 2011. Syntactic aspects of poetry: A pragmatic perspective. *The Buckingham Journal of Language and Linguistics*, 4.

Paul Rayson, Dawn Archer, Alistar Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of the Corpus Linguistics Conference: CL2007*. UCREL, University of Birmingham.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

University of Oxford. 2001. The york-helsinki parsed corpus of old english poetry (YCOEP). Oxford Text Archive.

University of Oxford. 2003. The york-toronto-helsinki parsed corpus of old english prose (YCOE). Oxford Text Archive.

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.

Elly van Gelderen. 2014. *A History of the English Language*. John Benjamins Publishing Company, Amsterdam. Casalini ID: 5001619.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical english. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*.

# A Appendix

### POS Mapping and Tag Distributions

Table 5 presents the tag set conversion scheme. The POS and UPOS columns denote the name of the label and its corresponding Universal POS tag, while the YCOEP and YCOE columns list the corresponding tags according to our conversion. Table 6 reports the ten most common bigrams and trigrams for each genre, along with their probabilities, based on a representative sample of the datasets. Table 7 presents the baseline models from Section 5.1 evaluated on the prose test set.

Figures 2, 3, and 4 show the heatmaps of the normalized confusion matrices for the baseline models tested on the poetry test set. Figures 5, 6, and 7 show the corresponding heatmaps for the baseline models tested on the prose test set. Figures 8 and 9 present the heatmaps for two key models from the experiments.

| 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, BENˆD, WADJˆA, VAGˆD, VBNˆI, VBNˆN, VBNˆD, BENˆA, BENˆG, ADJˆN, ADJSˆG, ADJSˆD, ADJˆI, VBNˆA, ADJRˆG, ADJˆG, ADJSˆN, ADJRˆD, WADJ, VAGˆG, VBNˆG, ADJRˆA, BENˆN |
| Adposition | ADP | P | P21, P22, PP, P+DˆI, 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, RP-1, ADVSˆL, ADV22, ADV+P, RP-4, RPX, ADVˆL, ADVˆD, ADVˆT, ADVS, ADVR, WADVˆT, WADVP-LOC-1, ADVRˆL, P+ADV, WADVˆL, WADV, RP, WADV+P, ADVRˆT, ADV21, ADVP-TMP |
| 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, AXP, AXPS, AXD, AXPI, MDˆD, MDDS, MDI, MDPI, AX |
| 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, QRˆA, QSˆD, DˆG, DˆA, Q, Q+NˆA, QˆG, QRˆG, Q21, DˆD, Q+NˆG, QRˆD, DˆI, QˆN, Q+QˆA, QSˆG, Q+NˆN, QˆA, QˆI, Q22, QR, QP-NOM, QS, QSˆN |
| Interjection | INTJ | INTJ | INTJ |
| Noun | NOUN | NP-ACC-SBJ, NˆG, NP-ACC, NP-DAT, NˆD, NˆI, NˆN, NP-NOM, NˆA, NP-DAT-PRN-1, NP-DAT-ADT | NP-ACC, NˆG, NP-NOM-x, NˆA, NˆN, NP-GEN, NP, N, NP-SBJ, NP-NOM, NˆI, NˆD |
| Numeral | NUM | NUMˆA, NUMˆG, NUMˆI, NUM, NUMˆD, NUMˆN | NUMˆD, NUMˆG, NUMˆA, NUMˆN, NUM, NUMˆI |
| Particle | PART | TO, UTP, FP, NEG, FP-5 | TO, FP, NEG, UTP |
| Pronoun | PRON | PRO$ ˆG, WPROˆA, PROˆA, PRO$, PROˆD, WPRO, PROˆN, PRO$ ˆN, PROˆI, WPROˆD, PRO$ ˆA, PRO$ ˆD, MANˆN, PROˆG, MANˆA, PRO$ ˆI, WPROˆG, WPROˆN, WPROˆI | PROˆD, PRO, WPRO, PROˆG, WPROˆN, PRO$ ˆN, WPROˆD, MANˆN, PRO$ ˆA, PRO$ ˆD, PROˆN, WPROˆG, PROˆA, WPROˆI, PRO$ ˆG, WPROˆA, PRO$, PRO$ ˆI |
| Proper Noun | PROPN | NPR, NPRˆN, NPRˆG, NPRˆD, NPRˆA | NRˆN, NR, NRˆG, NRˆA, NRˆD |
| Punctuation | PUNCT | , . | , . |
| Subordinating Conjunction | SCONJ | WNP-ACC-2, WNP-NOM-2, WNP-ACC-3, WNP-NOM-1, WQ, C, WNP-NOM-6 | C, WNP-NOM-2, WQ-1, WNP-ACC-1, WNP-ACC-3, CP-REL, WQ, WNP-ACC-2, WNP-NOM-1 |
| Verb | VERB | BE, VBD, VBN, VBPI, VAG, VBDI, BEDS, HVPS, HVD, HVPI, BED, VB, VBPS, VBDS, HVP, VBPH, BEDI, HVI, HV, VBˆD, VB-3, BEPS, HVDI, BEI, BEP, BEPI, VBI, VBˆA, BEN, VBP | BAG, HVˆD, VBD, HVDI, VB, HVN, VBP, HVPS, HV, HVD, VBDS, BEDI, HVP, VBPH, BE, BEDS, BEI, BEPS, VBN, VAG, BEPH, BEˆD, BED, BEN, HVI, HVDS, BEPI, HVPI, HAG, VBˆD, VBPS, VBPI, VBDI, VBI, BEP |
| Other | X | FW, UNKNOWN | XX, FW, UNKNOWN |
| Symbol | SYM | - | - |

Table 5: Mapping of YCOEP and YCOE to UPOS.

| Poetry | | | |
|---|---|---|---|
| **BiGram** | **Prob.** | **TriGram** | **Prob.** |
| ('NOUN', 'PUNCT') | 7.44% | ('NOUN', 'VERB', 'PUNCT') | 3.56% |
| ('VERB', 'PUNCT') | 5.67% | ('ADJ', 'NOUN', 'PUNCT') | 2.07% |
| ('NOUN', 'VERB') | 5.59% | ('NOUN', 'NOUN', 'PUNCT') | 2.03% |
| ('ADP', 'NOUN') | 4.66% | ('ADP', 'NOUN', 'PUNCT') | 1.78% |
| ('NOUN', 'NOUN') | 4.59% | ('NOUN', 'PUNCT', 'NOUN') | 1.41% |
| ('ADJ', 'NOUN') | 3.65% | ('NOUN', 'ADJ', 'PUNCT') | 1.40% |
| ('PUNCT', 'NOUN') | 3.36% | ('NOUN', 'ADP', 'NOUN') | 1.23% |
| ('DET', 'NOUN') | 2.44% | ('VERB', 'PUNCT', 'NOUN') | 1.16% |
| ('NOUN', 'ADJ') | 2.42% | ('NOUN', 'NOUN', 'VERB') | 1.15% |
| ('ADJ', 'PUNCT') | 2.39% | ('ADP', 'NOUN', 'NOUN') | 1.08% |
| **Prose** | | | |
| **BiGram** | **Prob.** | **TriGram** | **Prob.** |
| ('NOUN', 'PUNCT') | 4.99% | ('ADP', 'DET', 'NOUN') | 1.72% |
| ('DET', 'NOUN') | 4.96% | ('DET', 'ADJ', 'NOUN') | 1.45% |
| ('VERB', 'PUNCT') | 3.95% | ('NOUN', 'VERB', 'PUNCT') | 1.29% |
| ('PRON', 'VERB') | 3.10% | ('DET', 'NOUN', 'PUNCT') | 1.10% |
| ('NOUN', 'VERB') | 3.00% | ('ADJ', 'NOUN', 'PUNCT') | 1.10% |
| ('ADJ', 'NOUN') | 2.91% | ('DET', 'NOUN', 'VERB') | 1.09% |
| ('ADP', 'DET') | 2.89% | ('VERB', 'DET', 'NOUN') | 0.95% |
| ('ADV', 'VERB') | 2.42% | ('ADP', 'PRON', 'NOUN') | 0.85% |
| ('ADP', 'PRON') | 2.35% | ('VERB', 'ADP', 'DET') | 0.76% |
| ('VERB', 'ADP') | 2.29% | ('PRON', 'VERB', 'PUNCT') | 0.75% |

Table 6: Ten most frequent bigrams and trigrams with probabilities of representative samples from YCOEP and YCOE.

| | **OG Tag Set** | | **UPOS Tag Set** | |
|---|---|---|---|---|
| **Genre** | **Acc.** | **F1** | **Acc.** | **F1** |
| poe (5668) | 0.798 | 0.446 | 0.918 | 0.852 |
| pro (5668) | 0.936 | 0.789 | 0.971 | 0.962 |
| poe, pro (11,336) | 0.937 | 0.770 | 0.976 | 0.968 |

Table 7: Results for baseline POS taggers trained with the Original (OG) tag set and the Universal Dependencies (UD) tag set and tested on a prose test set. The data belong to either poetry, prose genres, or a combination of both.

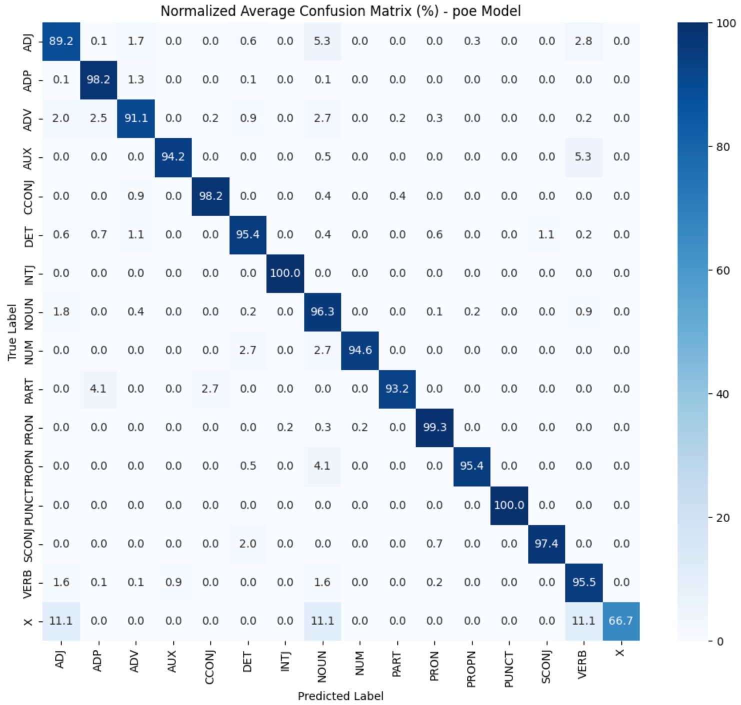


Figure 2: Normalized confusion matrix averaged over all seeds for the *poe* baseline model (Table 2) evaluated on the **poetry test set**.

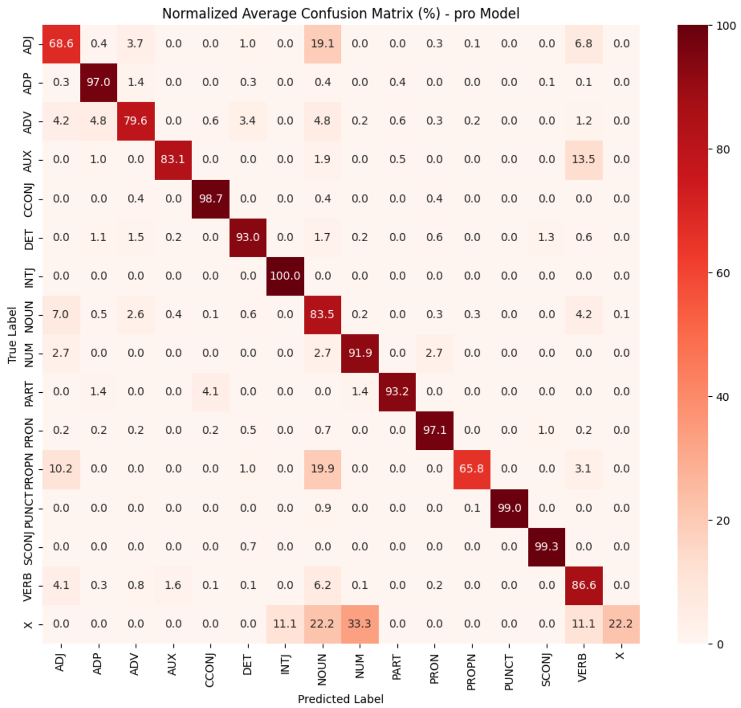


Figure 3: Normalized confusion matrix averaged over all seeds for the *pro* baseline model (Table 2) evaluated on the **poetry test set**.

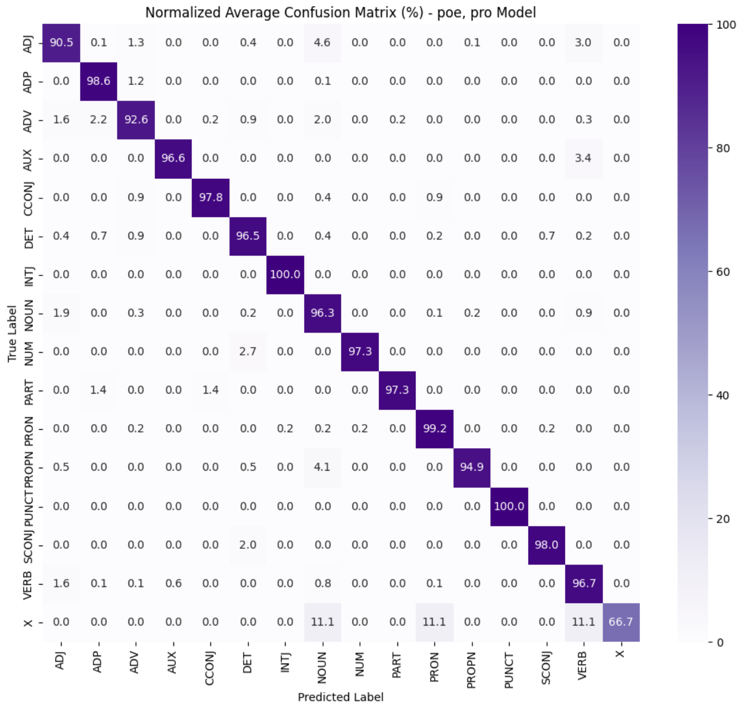


Figure 4: Normalized confusion matrix averaged over all seeds for the *poe, pro* baseline model (Table 2) evaluated on the **poetry test set**.

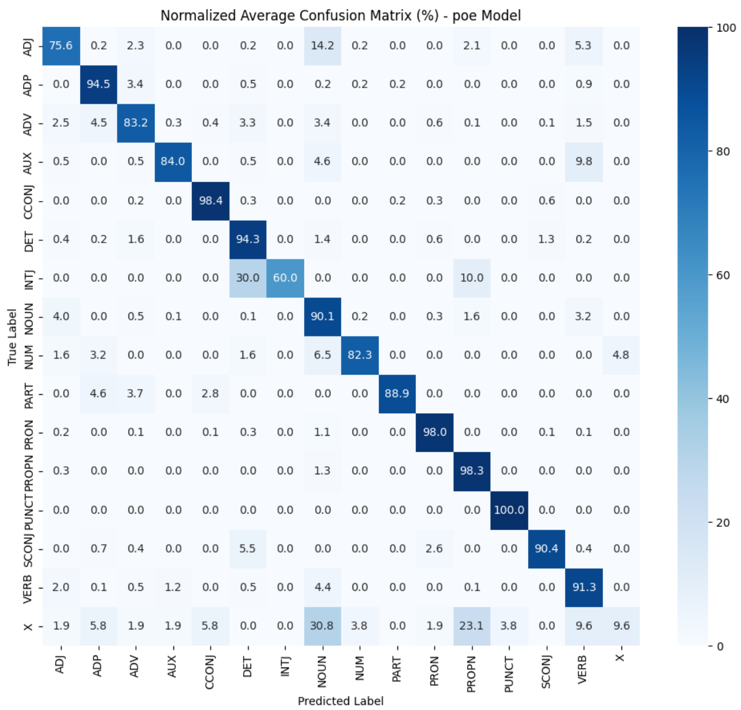


Figure 5: Normalized confusion matrix averaged over all seeds for the *poe* baseline model (Table 7) evaluated on the **prose test set**.

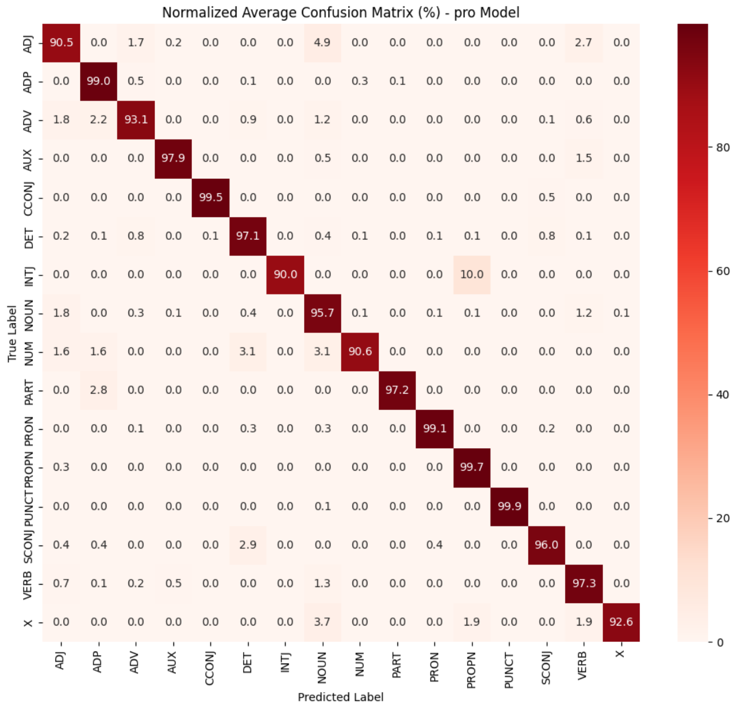


Figure 6: Normalized confusion matrix averaged over all seeds for the *pro* baseline model (Table 7) evaluated on the **prose test set**.

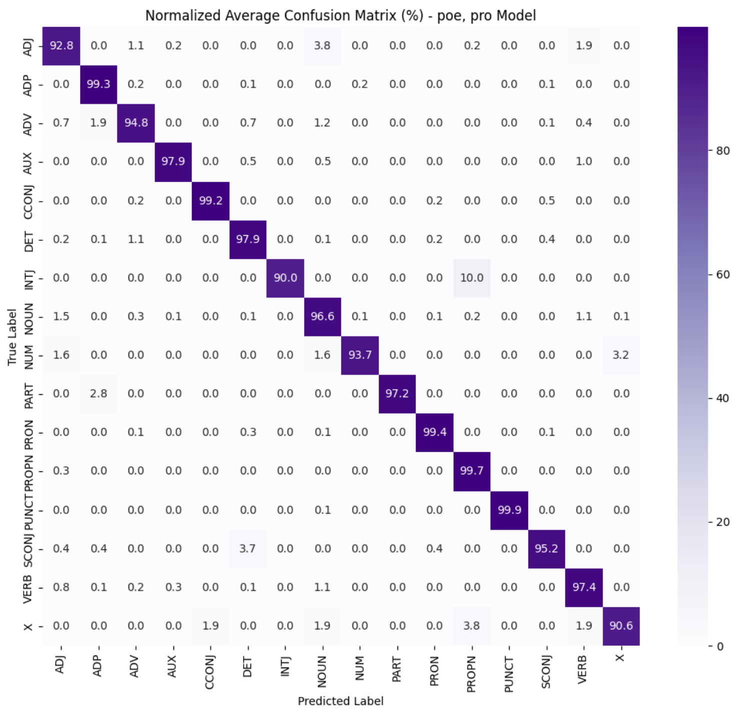


Figure 7: Normalized confusion matrix averaged over all seeds for the *poe, pro* baseline model (Table 7) evaluated on the **prose test set**.

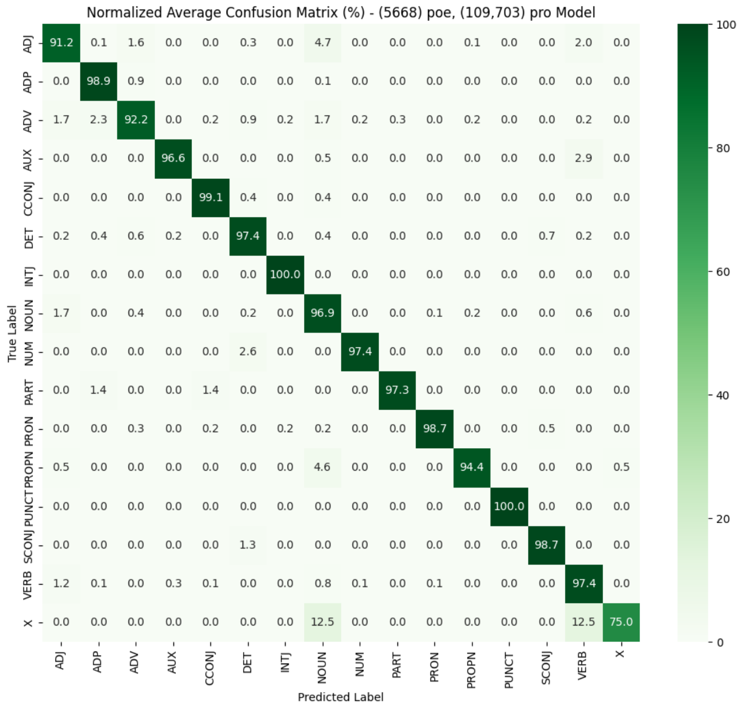

Figure 8: Normalized confusion matrix averaged over all seeds for the (5668 sent.) poe, (109,703 sent.) pro model (Table 3) evaluated on the **poetry test set**.

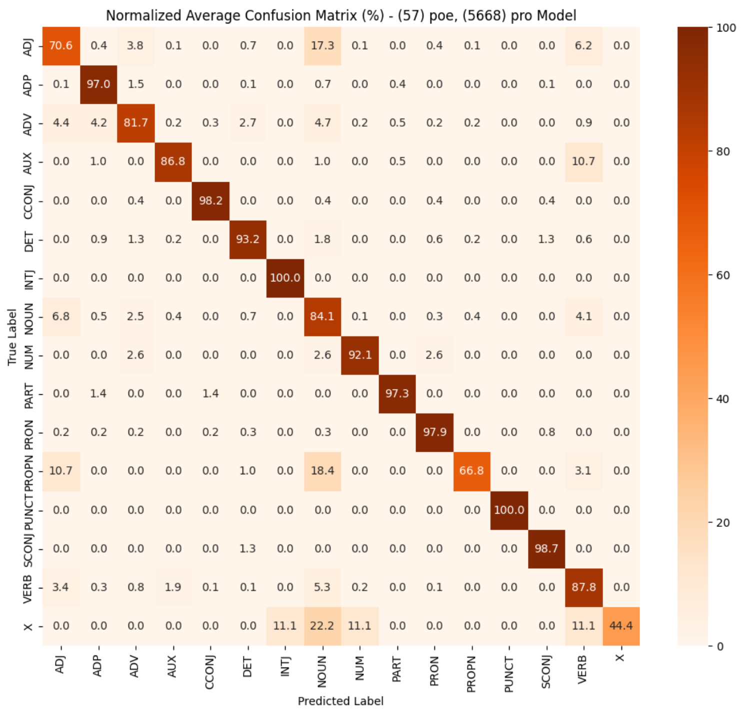

Figure 9: Normalized confusion matrix averaged over all seeds for the (57 sent.) poe, (5668 sent.) pro model (Table 4) evaluated on the **poetry test set**.