What data should I include in my POS tagging training set?

Zoey Liu
University of Florida

Masoud Jasbi
University of California, Davis

Christan Grant
University of Florida

Kenji Sagae
University of California, Davis

Emily Prud'hommeaux
Boston College

Abstract

Building an NLP training set for understudied languages, including Indigenous and endangered languages, often faces challenges due to varying degrees of resource limitations in the speaker communities. What are some reasonable approaches for training set construction in these cases? We address this question with POS tagging as the test case. Although many might consider POS tagging "a solved problem", it remains a crucial task for descriptive linguistics and language documentation and requires laborious manual annotation. Drawing data from 12 language families, we compare in-context learning, active learning (AL), and random sampling. Our results suggest: (1) for communities whose language data can be ethically shared with an API, using only 1,000 randomly sampled tokens as prompt examples, the proprietary `GPT-4.1-mini` can deliver desirable performance (F1 > 0.83) on par with that from a training set of thousands of tokens in AL iterations; (2) in cases where communities prefer not to share data, 4,500-5,500 tokens selected from AL can yield reasonable results at a pace *statistically* faster than random sampling, evidenced by growth curve modeling.

1 Introduction

When exploring a new language and domain for a variety of computational linguistics and natural language processing (NLP) tasks, we typically need to create a new training set for the task of interest. How should one create a new training set in order to derive more generalizable model performance on new unseen data in the wild? In particular, how much training data is necessary (van der Goot et al., 2024), and perhaps more importantly, what kind of data should be included in the training set?

An important consideration when addressing these questions is *resource availability*. The notion of resource availability, from our perspective, entails: (1) the amount of financial support for developing computational techniques; (2) the number of sources from which data can be collected; (3) the number *and* scope of manual annotations that can be acquired. As such, resource availability exists on a continuum with the abundant resources of the "big languages" that are the focus of most research (Søgaard, 2022) at one end and the severe resource limitation as in Indigenous and endangered languages at the other (Meek, 2012).

"High-resource" languages like English tend to have extremely large speaker populations; there is a commercial incentive to develop robust language technology for these languages. Data for these languages can come from a comparatively wide range of sources: text data can be collected from digitized books, online crowd-sourcing platforms, and web-crawled data (Silveira et al., 2014); and spoken data can be curated from read speech by a large number of participants (Panayotov et al., 2015) or existing media (Gauthier et al., 2016). Lastly, given the number of L1 speakers of these languages, it is straightforward to obtain manual annotations for a large training set. When resources are ample, less effort is required to determine how much and what data to include in a training set.

Now consider the other end of the resource continuum represented by Indigenous and endangered languages, for which the resources described above for languages such as English are, in most cases, unattainable. Because the focus on computational research has thus far been on languages with large speaker populations, there has consistently been much less financial support for and attention to language technology development for Indigenous and endangered languages (Blasi et al., 2022). Given their extremely small speaker populations, the number of sources for data collection for these languages is limited. Text data often comes from restricted domains such as the Bible (Domingues et al., 2024) or grammar books (Zhang et al., 2024), the content of which can be biased towards stilted

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8439–8455
November 4-9, 2025

or formal versions of the language that are detached from the reality of the language as it is spoken. Speech data is often derived from linguistic fieldwork (Shi et al., 2021a) carried out over decades using variable equipment and elicitation techniques. The limited availability of L1 speakers presents challenges for digitization from written texts and manual transcription of audio data. It would be impossible for the training set for an Indigenous or endangered language to ever equal that of a language like English. In addition, the annotations of the training data might vary depending on the linguistic expertise of the L1 speakers. These limitations complicate the question of training set creation, where great care is needed to perform resource allocation in a more thoughtful and efficient way.

Here we ask: what are some reasonable approaches to use when building a training set with limited annotation sources? Our goal is to inform *new* training set construction, particularly for underrepresented languages using *existing* datasets. This means the task to select as the test case should: (1) be valuable for language documentation; (2) have data available in a wide range of languages with different typological characteristics. With these considerations, we focus on part-of-speech (POS) tagging, taking advantage of the Universal Dependencies (UD) project (de Marneffe et al., 2021). We compare in-context learning with large language model (LLM), uncertainty sampling (Lewis, 1995) from the active learning framework (Settles, 2009), and random sampling (Mirbostani et al., 2023). For in-context learning, we leverage data from 60 languages (one treebank per language) across 12 language families from UD v2.14 (Zeman et al., 2024).[1]; for active learning and random sampling, we expand to 112 treebanks spanning 60 languages of the same language families.

2 Why POS Tagging?

The goal of POS tagging is to automatically assign each token in a given sentence a tag identifying its part-of-speech category.

```
All Boys are not Blue
DET NOUN AUX PART ADJ
```

While the rapid progress in NLP more broadly might have rendered POS tagging less popular, or even "a solved problem", we choose it here because **POS tagging is still widely used in various aspects of linguistics research**. For instance, theoretical linguists rely on the lexical categories of different languages for characterizing their typological profile (Berg, 2014). Cognitive scientists use POS tag distributions as interpretable features to capture structural transfer in second language learning (Liu et al., 2022a), characterize distributional patterns in code-switching (Chi and Bell, 2024), or they employ specific tags for identification of syntactic constructions in child language development (Sagae et al., 2005). Documentary linguists and independent researchers from Indigenous communities rely on POS tags for descriptive purposes and for creating pedagogical materials for new language learners.

In addition, since we aim to explore our question at a cross-linguistic scale, we employ datasets with existing POS annotations that are also relatively consistent across languages; this in turn motivates more comparability in our analysis. The availability of UD makes this possible. This also means that our research questions should be of interest for a variety of tasks and the methods described should be applicable to other scenarios, should cross-linguistic datasets be available.

3 Related Work

In-context learning There is a growing number of studies, mostly using in-context learning, to probe LLMs for their cross-linguistic capabilities in a variety of tasks, including but not limited to native language identification (Zhang and Salle, 2023), machine translation (Robinson et al., 2023), and word sense disambiguation (Cahyawijaya et al., 2025). Among these studies, some attend to low-resource or underrepresented languages specifically (Cahyawijaya et al., 2024), with work ranging from interlinear glossing (Ginn et al., 2024a,b; Shandilya and Palmer, 2025) and grammar creation (Spencer and Kongborrirak, 2025) for endangered languages, to natural language inference for indigenous languages of the Americas (Ebrahimi et al., 2022).

Only a few experiments have looked at POS tagging with in-context learning in particular. Machado and Ruiz (2024) compared three LLMs for POS tagging of Portuguese: `GPT-3`, `LLaMA-7b`, and `Maritaca` (pretrained on Portuguese specifically). Their prompt included 10 sentence examples and each LLM was asked to generate inference for 1,000 sentences. The results showed while `GPT-3` achieved F1 scores above 0.8, the

[1]We use UD data for research purposes abiding by their guidelines licensed by CC-BY-SA-4.0 license.

performance of LLaMA-7b (∼0.56) and Maritaca (∼0.42) was much worse. Adelani et al. (2024) used zero-shot prompting with GPT-4 for 12 low-resource languages from Brazil and two from Africa, and found that this approach outperformed zero-shot cross-linguistic transfer learning with alternative pretrained LMs (Conneau et al., 2020).

Our study goes beyond the scope of prior work using LLMs via experimenting with few-shot in-context learning for 60 languages from 12 language families (60 treebanks). One might be inclined to believe that an LLM would undoubtedly perform well for POS tagging given that the task is comparatively simple; on the other hand, it is not unreasonable to suspect the opposite, given that LLM performance for low-resource languages is quite variable (e.g., Robinson et al. (2023); Ginn et al. (2024b). These assumptions necessitate quantifying the performance of LLMs for POS tagging to empirically inform training set construction.

Active learning The active learning (AL) framework has long been proposed as a method for informative data selection when resources for acquiring manual annotations are limited (Palmer, 2009; Hwa, 2004; Hwa et al., 2003; Osborne and Baldridge, 2004; Baldridge and Osborne, 2004; Steedman et al., 2003). The goal is to derive models that can reach a certain level of performance from comparatively less training data.

The general process of AL is as follows. We have a designated test set and a *training data pool* to draw training samples. We first build an *initial training set* of a certain size from the training data pool, and train a model on this initial training set. We then apply the trained model to the *residual data* from the training data pool; the model generates a score (e.g., a confidence score) for each data point in the residual data that reflects how *confident* the model is in its prediction for that particular data point. Then a number of X data points for which the model is the least certain about will be added to the initial training set with its correct output label. We build a model using this expanded training set to start the next iteration of AL; this process continues iteratively with the value of X fixed in every cycle. In real-world scenarios, the initial training set and the additional data selected in each cycle of AL are mostly annotated by human annotators then passed on to the next iteration. Here we perform computational simulation of this process in controlled experimental settings, using existing annotations provided by UD.

While AL is promising, current research on this topic, including work focusing on POS tagging, faces shortcomings. First, the majority of AL-related studies still attend to individual languages, most of which are English. Large-scale cross-linguistic investigations are still rare, with a few exceptions that have focused on the task of morphological inflection (Muradoglu et al., 2024; Muradoglu and Hulden, 2022). While some studies have looked into AL for POS tagging, they are again constrained by the language samples in their individual experiments (e.g., only English in Stratos and Collins (2015), eight Indo-European languages in Duong et al. (2014), and six low-resource languages in Chaudhary et al. (2021)). Second, the analyses of results from AL in the literature tend to take an "eyeballing" approach. These analyses rely on raw numbers and visualizations of the learning curves to draw conclusions about whether and to what extent AL is helpful.

Our work addresses these gaps. We investigate AL for POS tagging across 60 languages; the different treebanks for the same language are deliberately studied individually to conserve potential impact from different domains on the observations. In addition, unlike previous research that tends to lack statistical validation, we show how results of different sampling methods (AL vs. random sampling) can be compared systematically using *growth curve modeling* (Panik, 2014), originated from literature on biological analysis (Richards, 1959).

4 Experiments

4.1 Training set construction

We take treebanks of contemporary languages from UD v2.14, for which the training set has at least 3,000 tokens. This results in 112 treebanks across 60 languages spanning 12 language families (see Appendix A; language family information is taken from The World Atlas of Language Structures (Dryer and Haspelmath, 2013)); all treebanks have a pre-defined train/test split. For each treebank, we treat the training set as the training data pool for AL, and the full test set as the new unseen test data to evaluate model performance from each iteration.

We set the initial training set size to be 1,000 tokens to reflect a typical goal for a session of manual annotation.[2] This decision is informed by

[2] In preliminary analysis, we also experiment with an initial training set size of {50, 100, 500} tokens; the results are qualitatively comparable.

Garrette and Baldridge (2013), which showed that for each of Kinyarwanda, Malagasy, and English, non-native speakers were able to manually annotate POS tags of full sentences totaling 1,500 (Kinyarwanda) to 2,600 (English) tokens in two hours. Additionally, we set a maximum training set size to be 100K tokens, since our primary concern is a low-resource scenario; that is, the AL or the random sampling process for this treebank will stop when the training set reaches 100K tokens.

For AL, how an initial training set should be constructed is not always clear. Previous work has adopted two main approaches for constructing an initial training set: *cold-start* AL with randomly sampled initial sets (Yu et al., 2023; Jin et al., 2022; Houlsby et al., 2014), and *warm-start* AL with external already-annotated dataset or multilingual pretrained language models (Varadarajan et al., 2023; Zhu et al., 2019). To ensure comparability of experimental settings across languages, we adopt the cold-start approach; for each treebank, we randomly sample an initial training set that is constant for LLM prompting as well as kicking off the AL and random sampling process, respectively.[3]

For data selection in each iteration of AL, while there can be different metrics based on the model architecture and output (Mirbostani et al., 2023), here we rely on uncertainty sampling (Lewis, 1995), using the confidence score (cf. Yuan et al. (2020); Dasgupta (2011)) from conditional random fields (CRF) (Lafferty et al., 2001) (Section 4.3) which is measured as the average marginal probabilities across all tokens of a sentence; the lower the marginal probability, the less certain the model is about its prediction for a particular sentence. We then select a number of sentences with the lowest confidence scores totaling approximately 500 tokens, a value kept constant throughout AL for each treebank. For random sampling, on the other hand, in each iteration we randomly select sentences also totaling around 500 tokens to add to the previous training set; this process continues in an iterative fashion as well, ensuring that the resulting training set size from each random sampling iteration is comparable to that from the AL iteration.

4.2 In-context learning

For speech communities that consider LLMs safe or ethical to use for computational tasks for their own languages, LLMs might be a viable go-to given recent advancements, with either in-context learning or fine-tuning. This section describes our experiments and results for POS tagging using in-context learning.

While having thousands of sentence examples for POS tagging in the prompt might lead to better or higher performance, here for each treebank we focus on including just the initial training set of ∼1,000 tokens (see Section 4.1); this only refers to the sentence examples of POS tagging (an average of 61 sentences across treebanks), thereby excluding other instructions given to the LLM in the prompt (Table 3 in Appendix B). With each treebank, we use the same prompt format and only switch out the language name and the sentence examples. Our goal is to see how well an LLM can perform POS tagging with just these 1,000 tokens, then use the results as comparisons to AL, in order to assess their respective strengths and weaknesses. By doing this we hope to offer insights for what data selection methods (and models) to use for dataset creation.

In reality, any research that potentially involves data from an Indigenous or endangered language can only be conducted after researchers obtain consent from the language community. Taking that into account, we initially explore several LLM variants from the `Llama` family run locally through Ollama, which is completely free, for a smaller number of treebanks. Despite giving the UD tag set in the prompt as well as clear instructions about the language of interest and the desired output format, these smaller LLMs are not successful, somewhat consistently generating tags not in the UD tag set or outputting tag sequences where the number of tags does not match the input sentence length.

We therefore turn to the state-of-the-art proprietary model `GPT-4.1-mini` via the Python API of OpenAI. We select a subset of 60 treebanks (one per language; Table 1), balancing language family and speaker population, along with test set size and the corresponding estimated cost. For evaluation, we again adopt the original test set in full from every treebank. Model performance is measured as weighted F1 score. Specifically, we compute the F1 score of each POS tag, weight it by its frequency in the test set, then take the average of the F1 scores across all POS tags.

Results As shown in Table 1, with the exceptions of the treebanks for two Uralic languages, Erzya and North Sami, and simplified Mandarin Chinese, F1

[3] In practice, we randomly sample 3 initial training sets to control for potential variation; there is no observable difference in the results from the different samples.

Treebank	Language family	GPT-4.1-mini F1	N of tokens from AL
UD_Arabic-PADT	Afro-Asiatic	0.89	6,706
UD_Hebrew-HTB		0.94	48,905
UD_Maghrebi_Arabic_French-Arabizi		0.83	12,724
UD_Maltese-MUDT		0.93	max=0.92
UD_Indonesian-GSD	Austronesian	0.92	11,185
UD_Tamil-TTB	Dravidian	0.84	4,084
UD_Telugu-MTG		0.93	max=0.92
UD_Naija-NSC	English-based creole	0.94	5,535
UD_Afrikaans-AfriBooms	IE	0.95	12,407
UD_Armenian-ArmTDP		0.92	max=0.91
UD_Belarusian-HSE		0.94	18,693
UD_Bulgarian-BTB		0.97	27,918
UD_Catalan-AnCora		0.96	28,006
UD_Croatian-SET		0.95	25,606
UD_Czech-CAC		0.97	19,815
UD_Danish-DDT		0.93	26,103
UD_Dutch-LassySmall		0.92	24,003
UD_English-ParTUT		0.93	7,672
UD_Faroese-FarPaHC		0.94	12,813
UD_French-ParTUT		0.97	max=0.96
UD_Galician-TreeGal		0.93	max=0.92
UD_German-GSD		0.94	97,258
UD_Greek-GUD		0.95	max=0.94
UD_Hindi-HDTB		0.90	7,086
UD_Icelandic-GC		0.90	23,950
UD_Irish-IDT		0.90	12,886
UD_Italian-ParTUT		0.96	33,509
UD_Lithuanian-HSE		0.92	max=0.72
UD_Latvian-LVTB		0.93	43,156
UD_Low$_{S}axon - LSDC$		0.85	max=0.84
UD_Norwegian-Nynorsk		0.92	9,596
UD_Persian-Seraji		0.94	19,590
UD_Polish-LFG		0.95	19,107
UD_Portuguese-GSD		0.94	17,995
UD_Pomak-Philotis		0.86	2,532
UD_Romanian-RRT		0.94	28,993
UD_Russian-GSD		0.95	29,490
UD_Serbian-SET		0.96	15,397
UD_Slovenian-SST		0.91	11,724
UD_Spanish-GSD		0.94	43,137
UD_Scottish_Gaelic-ARCOSG		0.85	3,103
UD_Slovak-SNK		0.94	max=0.91
UD_Swedish-LinES		0.95	max=0.94
UD_Ukrainian-IU		0.94	31,300
UD_Urdu-UDTB		0.90	12,401
UD_Welsh-CCG		0.85	4,024
UD_Western_Armenian-ArmTDP		0.92	18,767
UD_Basque-BDT	Isolate	0.88	10,647
UD_Japanese-GSD	Japonic	0.91	15,793
UD_Korean-GSD	Koreanic	0.87	16,804
UD_Wolof-WTB	Niger-Congo	0.87	4,060
UD_Chinese-GSDSimp	Sino-Tibetan	0.84	19,082
UD_Vietnamese-VTB		0.87	max=0.85
UD_Turkish-Kenet	Turkic	0.87	12,089
UD_Uyghur-UDT		0.86	7,026
UD_Erzya-JR	Uralic	0.70	1,498
UD_Estonian-EWT		0.91	max=0.90
UD_Finnish-TDT		0.94	max=0.93
UD_Hungarian-Szeged		0.93	max=0.90
UD_North_Sami-Giella		0.78	3,007

Table 1: Weighted F1 scores from GPT-4.1-mini across 60 treebanks; N of tokens from AL refers to the number of tokens in a training set in the AL process needed to reach comparable performance as that of GPT-4.1-mini; e.g., max=0.92 means that the maximum F1 score from AL (regardless of training data size) is 0.92 and lower than the result from GPT-4.1-mini.

scores from GPT-4.1-mini are consistently above 0.83 for all other treebanks, with the majority having performance equal to or above 0.90. A model with such strong performance can most likely suffice *in the wild* to obtain first-pass automatic annotations of new unseen data before possibly going through manual correction. A sneak peak at the results from AL reveals that to perform on par with GPT-4.1-mini, a minimum of thousands of tokens in the training set selected from the AL iter-

ations is required, with some treebanks requiring even more. For cases such as UD_Maltese-MUDT and UD_Hungarian-Szeged, even the maximum F1 score from AL is still lower than that from the LLM. These observations might suggest that in cases where language data can be processed via the GPT-4.1-mini API, 1,000 randomly selected tokens as prompt examples can deliver desirable performance, at least for the languages evaluated. At the same time, we acknowledge the possibility that the UD treebanks might have been included in the training data for GPT-4.1-mini; additionally, none of the 60 languages here is truly low-resource (Liu et al., 2022b). It is not unlikely that for Indigenous and critically endangered languages, the performance of LLMs is (much) worse, necessitating the approach of AL, as we will describe below.

4.3 AL and random sampling

While LLMs can be powerful, for many Indigenous speech communities, data is frequently not shared with outsiders, making the use of models such as GPT-4.1-mini unethical and incompatible with the community's values. Given the weak POS tagging performance of LLMs that are small enough to be run locally via Ollama, we turn to (old-fashioned) AL.

Model We use CRF throughout experiments in this section. Despite being a relatively simple statistical model, prior research has shown the effectiveness of CRF over several neural models in sequence labeling tasks (e.g., morphological segmentation (Liu and Dorr, 2024)) in scenarios with limited data.[4] Given a sentence, CRF uses a curated feature set of each individual token to predict its POS tag (see implementation details in Appendix C). After initial experimentation, for each token, the feature set includes information such as the number of characters, character-level n-grams (up to four) in the word itself, as well as the same information about the previous one and two words (if available). Model performance for each treebank is measured as weighted F1 score derived from the full test set.

Statistical analysis As mentioned previously (Section 3), prior research on AL lacks statistical analysis to verify whether it is better than alternative baselines. We address this gap here using growth curve models (Panik, 2014), the variants of which have been applied in linguistics to computationally assess the developmental trajectory of phenomena such as infant development (Neale and McArdle, 2000) and cognitive control (Erb et al., 2023). With growth curve modeling, we can identify a training configuration that results in more rapid performance improvements and hence has the potential to more efficiently allocate annotation efforts in a non-simulated AL scenario when only limited resources are available, which would particularly benefit endangered languages.

To compare the performance of the two sampling methods, we rely on nonlinear growth curve analyses of F1 scores (Panik, 2014) as training sample size increases. We use the four-parameter Weibull model (Ratkowsky, 1983) shown in Eq. 1 below, modeling F1 score as a function of training size (t). F1 scores range between (0, 1) with model performance possibly improving nonlinearly until it reaches an upper limit at which performance does not improve any further (see Figure 1 in Section 4.3). Similarly, a four-parameter nonlinear Weibull growth model assumes an upper asymptote α as ceiling for the F1 scores, a lower asymptote β where model performance starts, a growth or improvement rate γ which shows how quickly the model improves, and finally a shape parameter δ which determines the shape of the learning curve, with lower δ indicating sharper approach to the upper asymptote; e is Euler's Number:

$$F1 = \alpha - (\alpha - \beta)e^{-(\gamma t)^{\delta}} \quad (1)$$

We compare the average growth rate γ between uncertainty sampling from AL and random sampling. An observed higher growth rate would suggest that the performance of one method improves more quickly than another, which as noted above could contribute to decreasing the amount of human annotation effort needed to create robust models. To fit our nonlinear growth curves we used the BRMS package in R (Bürkner, 2017). (For details of parameterization and fitting of the growth curve models, see Appendix F).

A walk-through with Irish We first present a walk-through of the results for the UD_Irish-IDT treebank of Irish, classified as endangered by Ethnologue (Eberhard et al., 2025). Based on the Expanded Graded Intergenerational Disruption Scale (EGIDS), a scale ranging from 0 (International) to 10 (Extinct) (Lewis and Simons, 2010, 2017), Irish

[4] In preliminary work, we compare CRF with a neural alternative, the transformer architecture TRANSFORMER_TINY from fairseq (Ott et al., 2019); the neural model, while being computationally much more expensive, consistently underperforms compared to CRF.

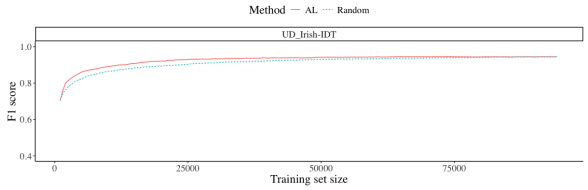

Figure 1: Learning curves for UD_Irish-IDT from AL and random sampling; x-axis represents the training set sizes and y-axis corresponds to the weighted F1 scores.

is level 6b (*Threatened*).[5]

As illustrated in Figure 1, there is a clear pattern of a logarithmic curve (with an upper bound) from the AL process, which increases rapidly at the beginning, then gradually slows before reaching a plateau, remaining stable afterwards. In this case, the F1 score starts at 0.71 with 1,000 tokens, reaches 0.85 with a training set of 4,500 tokens; around 12,000 tokens we arrive at an F1 score of above 0.90, which slowly increases to over 0.93 when training set exceeds 25,000 tokens, then plateaus. The learning curve for AL is also visually above that from random sampling until approaching the tail of the curve.

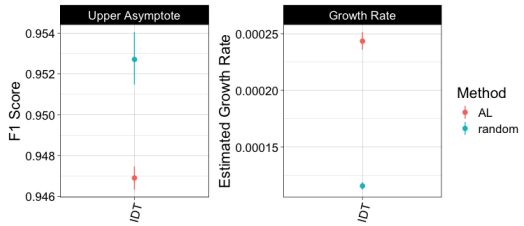

Figure 2: Upper asymptote and growth rate estimates of the Bayesian growth curve fits to the weighted F1 scores of UD_Irish-IDT.

Figure 2 shows the Bayesian growth curve estimates for upper asymptotes and growth rates of UD_Irish-IDT. The upper asymptote estimates show the maximum F1 scores that the models would reach if the training continued indefinitely and the growth rate shows how fast F-1 scores approach this maximum score. For the UD_Irish-IDT treebank, the upper asymptote estimate is higher ($\alpha = 0.9527, 95\%CI = (0.951, 0.954)$) than that of the AL method ($\alpha = 0.9469, 95\%CI = (0.946, 0.947)$). On the other hand, the growth rate estimates suggest that the AL method reaches higher F1-scores much faster and with fewer training data sets ($\gamma = 0.000243, 95\%CI = (0.000235, 0.000251)$) than the random sampling method ($\gamma = 0.000115, 95\%CI = (0.000112, 0.000118)$).

To gain more insight into the learning curve patterns, we analyze the POS tag distributions of training sets from AL iterations by measuring the distance between each training set and the test set. We use Kullback–Leibler (KL) divergence (Csiszar, 1975) as an approximation of how *divergent* a training set (Q) is from the test set (P), the latter of which is the reference distribution (Eq 2). We anticipate that the more different the training and the test sets are, the lower the F1 score will be.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \quad (2)$$

Results (Figure 5 in Appendix D) show that as training size increases, the KL divergence value mostly decreases, while showing more fluctuations with much larger training sets. To assess this relationship, we fit a linear regression predicting KL divergence as a function of the training size at every AL iteration. There appears to be a weak yet significant negative effect for training size ($\beta = -4.185e-08, p < 0.001$), meaning the distributions of the training and test sets become closer with larger training sets. As a result, we also find a pronounced negative relationship between KL divergence and the F1 score ($\beta = -5.35, p < 0.001$), confirming that a training set distributionally more similar to the test set yields better performance.

Results from Figure 1 are aggregated over all POS tags from the designated test sets of UD_Irish-IDT. Now we consider the learning curves of individual tags. To address this matter, we compute the F1 score for each of the 17 POS tags from UD; the learning curves for most POS tags (Figure 6 in Appendix E) from AL correspond to that of a logarithmic curve similar to the observations averaged across all tags (Figure 1).

We offer three conjectures for the observed learning curve patterns of individual POS tags. The first pertains to the probability of the tag in the training set (Tag_prob), which we expect to have a positive effect on the F1 scores of the tag across the AL iterations. The second conjecture involves the distribution of words with a given tag (e.g., all words tagged as NOUN). On one hand, we anticipate that more variation in word distribution might help a model be more robust to variation in the test set; on the other hand, it could also introduce more infrequent patterns that might not help the model

[5] https://en.wikipedia.org/wiki/Expanded_Graded_Intergenerational_Disruption_Scale

learn more effectively. To evaluate this conjecture, we measure the amount of variation in word distribution for a certain tag (`Tag_word_entropy`) with entropy (Shannon, 1948). For instance, we take all words tagged as `NOUN` in a treebank, derive a probability distribution for these words (e.g., {'dog': 0.4, 'cat': 0.3, 'cheese': 0.3}), then calculate entropy with this probability distribution.

The third conjecture concerns the syntactic environment of a POS tag, i.e., what other tags a given POS tag can co-occur with and how often they co-occur. In this case, we consider the bigram distributions of POS tags. Take `NOUN` as an example. We first collect all tag bigrams where `NOUN` appears, turn them into a probability distribution (e.g., {'NOUN VERB': 0.3, 'VERB NOUN': 0.4, 'DET NOUN': 0.3}), then use entropy to measure the distribution of these bigrams (`Tag_syntax_entropy`).

For each treebank, we use mixed-effect linear regression to probe the roles of the aforementioned three conjectures. We include `Tag_prob`, `Tag_word_entropy`, and `Tag_syntax_entropy` as fixed effects and the POS tag as the random effect; the regression model predicts the F1 score from each AL iteration. Our analysis shows a notable positive effect for `Tag_word_entropy` ($\beta = 0.036, 95\%CI = (0.031, 0.041)$), suggesting that the more variable the word distributions of a given POS tag are, the better the model performance will be. On the other hand, `Tag_syntax_entropy` exhibits the opposite effect ($\beta = -0.05, 95\%CI = (-0.07, -0.03)$), indicating that F1 scores tend to be higher when the syntactic environments of the POS tag are less diverse. In contrast to these two factors, there is no notable effect for `Tag_prob` ($\beta = -0.17, 95\%CI = (-0.51, 0.17)$).

4.3.1 Overall results

We carry out the same analysis for other 60 languages and their associated UD treebanks as we did for Irish. We first examine the learning trajectory of model performance from AL. Across treebanks mostly with large training data pool (see Fig. 3 for contrasts between selected treebanks from different language families), the learning curve for F1 score increases comparatively rapidly until the training set reaches approximately 4,500-5,500 tokens. Growth then begins to decelerate substantially, with absolute F1 increases of less than 0.001 for every additional 500 to 1,000 tokens after the training set reaches 20,000 tokens approximately.

We compare the growth rates estimated from Bayesian growth curve modeling for AL vs. random sampling in Figure 4. We observe that for all the treebanks with reliable growth rate estimates, the AL models reach the estimated upper asymptote faster than the models that used the random sampling method. For upper asymptote estimates, however, the growth curve models do not show a clear advantage for either method (Figure 7 in Appendix F). For most treebanks, the upper asymptote estimates overlap, suggesting that both methods are predicted to reach the same maximal F1 score with continued training. Putting these results together, while models from AL approach their maximum F1 scores faster, the upper asymptotes of the F1 scores are not generally significantly higher than those from random sampling.

Note that the seemingly differing observations between growth rates and upper asymptotes are not in contradiction to each other, as the two parameters are estimating different properties of the learning curves. We consider the discrepancies to actually have important implications for dataset design. If one cares more about deriving a model with reasonable performance at a faster pace – for example, to get a working model to perform automatic annotation in the pipeline to speed up the process of manually correcting machine output – AL would be preferred. Alternatively, if achieving the highest score possible is the main concern, resources permitting, random sampling would be optimal.

5 Discussion & Future Directions

With existing datasets for 60 languages from 12 language families, our study makes the following recommendations for building POS tagging training sets for new languages that possibly face limitations in available annotation efforts. In scenarios where it is safe, ethical, and compatible with speaker community values to expose language data through an API, in-context learning with `GPT-4.1-mini` using a small randomly sampled training set of 1,000 tokens in the LLM prompt can deliver desirable results that are mostly better than those derived from AL and random sampling (at least for the languages investigated). In our experiments, API calls per treebank cost no more than $4 US dollars; this amount, together with the expense for obtaining manual annotations for 1,000 tokens, would be more economical than acquiring manual annotations for thousands of tokens for AL.

There are, however, many Indigenous and en-

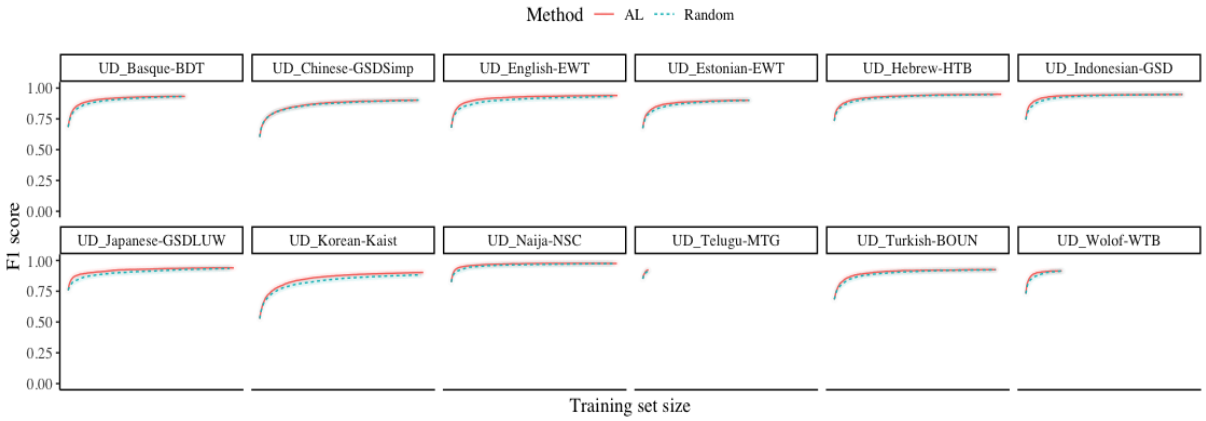

Figure 3: Snapshot of learner curves for selected treebanks from AL and random sampling; *x*-axis represents the training set sizes and *y*-axis corresponds to the weighted F1 scores.

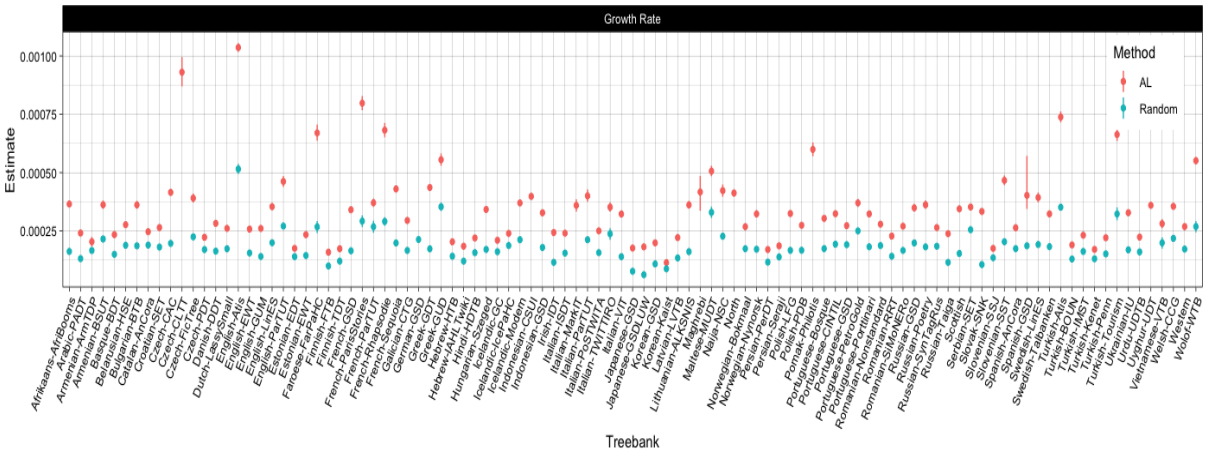

Figure 4: Growth rates with 95% credible intervals from Bayesian growth curve modeling for each treebank.

dangered speech communities for whom data sovereignty is a sensitive issue (Kukutai and Taylor, 2016). For these cases, AL for POS tagging with a statistical model that runs locally can be effective. In our AL experiments, the learning curves largely grow rapidly until reaching a training size of 4,500-5,500 tokens, after which growth diminishes and begins to approach an asymptote. Our statistical model further reveals that across the languages tested, while AL and random sampling might reach comparable maximum F1 scores eventually, models from AL iterations arrive at the upper asymptote faster. This means that for resource-constrained scenarios with limited bandwidth for manual annotations, AL is the better choice. We hope that our research methodology can be adopted in future work on AL and dataset design more broadly.

We see a number of fruitful directions in our future work. First, we would like to explore diversity sampling (Bodó et al., 2011), an alternative sampling method that has been incorporated into AL in previous literature (Shi et al., 2021b). Second, we hope to learn more about the practical impact of our findings. For instance, while our results might suggest certain recommendations for building datasets, we do not know the relationship between improvements in POS tagging F1 for these languages and downstream tasks that rely on POS tags. In addition, because our work only *simulates* the process of AL, we do not yet fully understand the impact of putting AL into action on the required effort of real annotators (see also Baldridge and Palmer (2009)). Finally, we would like to carry out similar experiments with other NLP tasks, especially those that are helpful in the contexts of documenting endangered languages, such as automatic speech recognition (Prud'hommeaux et al., 2021) and morphological segmentation (Garrett, 2011).

6 Limitations

While we believe our work to be thorough and our choices well justified, we do acknowledge some potential limitations. First, we only experiment with a fixed size of 1,000 tokens as prompt examples for in-context learning; it is possible that even a smaller training sets might yield comparable performance. Second, our prompt examples are randomly sampled; future work can consider more informed data selection strategies for in-context learning or finetuning with LLM.

Finally, the UD project, while offering a wide range of language selections, is (as of now) heavily weighted toward IE languages, which are not necessarily representative linguistically of the world's languages at large, or Indigenous and endangered languages in particular. We choose UD for its relatively consistent annotation standards across languages, as well as the fact that it is open-access. Additionally, languages from UD with comparatively fewer resources can provide some idea of how well the approach would fare with endangered languages. Since an endangered language is often tied closely to its respective community's identity and cultural heritage, work involving unpublished data from an endangered language is best done in close partnership with and for the direct benefit of its community. Here, we focus on the methods that would enable such work. As more datasets annotated with (UD-style) POS tag information become publicly available, we hope our methods would be directly applicable to those cases.

References

David Ifeoluwa Adelani, A. Seza Doğruöz, André Coneglian, and Atul Kr. Ojha. 2024. Comparing LLM prompting with cross-lingual transfer performance on indigenous and low-resource Brazilian languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 34–41, Mexico City, Mexico. Association for Computational Linguistics.

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annotation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 9–16.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Thomas Berg. 2014. Boundary permeability: A parameter for linguistic typology. *Linguistic Typology*, 18(3):489–531.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 127–139, Sardinia, Italy. PMLR.

Paul-Christian Bürkner. 2017. brms: An R package for bayesian multilevel models using Stan. *Journal of statistical software*, 80:1–28.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Samuel Cahyawijaya, Ruochen Zhang, Jan Christian Blaise Cruz, Holy Lovenia, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2025. Thank you, Stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word senses. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3228–3250, Albuquerque, New Mexico. Association for Computational Linguistics.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Jie Chi and Peter Bell. 2024. Analyzing the role of Part-of-Speech in code-switching: A corpus-based study. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1712–1721, St. Julian's, Malta. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

I. Csiszar. 1975. *I*-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146 – 158.

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781.

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

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in AI tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 283–293, Torino, Italia. ELRA and ICCL.

Matthew S. Dryer and Martin Haspelmath. 2013. WALS Online (v2020.4). Available online at https://wals.info, Accessed on 2025-09-15.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Doha, Qatar. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Christopher D Erb, Laura Germine, and Joshua K Hartshorne. 2023. Cognitive control across the lifespan: Congruency effects reveal divergent developmental trajectories. *Journal of Experimental Psychology: General*, 152(11):3285.

Andrew Garrett. 2011. An online dictionary with texts and pedagogical tools: The Yurok language project at Berkeley. *International Journal of Lexicography*, 24(4):405–419.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3863–3867, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024b. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Neil Houlsby, José Miguel Hernández-Lobato, and Zoubin Ghahramani. 2014. Cold-start active learning with robust ordinal matrix factorization. In *International conference on machine learning*, pages 766–774. PMLR.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30(3):253–276.

Rebecca Hwa, Miles Osborne, Anoop Sarkar, and Mark Steedman. 2003. Corrected co-training for statistical parsers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*.

Qiuye Jin, Mingzhi Yuan, Shiman Li, Haoran Wang, Manning Wang, and Zhijian Song. 2022. Cold-start active learning for image classification. *Information sciences*, 616:16–36.

Tahu Kukutai and John Taylor. 2016. *Indigenous data sovereignty: Toward an agenda*. ANU press.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

David D Lewis. 1995. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA.

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique*, 55(2):103–120.

M Paul Lewis and Gary F Simons. 2017. *Sustaining Language Use*. SIL International.

Zoey Liu and Bonnie Dorr. 2024. The effect of data partitioning strategy on model generalizability: A case study of morphological segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2851–2864, Mexico City, Mexico. Association for Computational Linguistics.

Zoey Liu, Tiwalayo Eisape, Emily Prud'hommeaux, and Joshua K Hartshorne. 2022a. Data-driven crosslinguistic syntactic transfer in second language learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022b. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.

Mateus Machado and Evandro Ruiz. 2024. Evaluating large language models for the tasks of PoS tagging within the Universal Dependency framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 454–460.

Barbra A Meek. 2012. *We are our language: An ethnography of language revitalization in a Northern Athabaskan community*. University of Arizona Press.

Seyed Morteza Mirbostani, Yasaman Boreshban, Salam Khalifa, SeyedAbolghasem Mirroshandel, and Owen Rambow. 2023. Deep Active Learning for Morphophonological Processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 793–803, Toronto, Canada. Association for Computational Linguistics.

Saliha Muradoglu, Michael Ginn, Miikka Silfverberg, and Mans Hulden. 2024. Resisting the lure of the skyline: Grounding practices in active learning for morphological inflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 47–55, Bangkok, Thailand. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. How to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303.

Michael C Neale and John J McArdle. 2000. Structured latent growth curves for twin data. *Twin Research and Human Genetics*, 3(3):165–177.

Bruno Nicenboim, Daniel Schad, and Shravan Vasishth. 2021. An introduction to Bayesian data analysis for cognitive science. *Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series*.

Miles Osborne and Jason Baldridge. 2004. Ensemble-based active learning for parse selection. In *Proceedings of the human language technology conference of the north American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 89–96.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. Ph.D. thesis, University of Texas at Austin.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Michael J. Panik. 2014. *Growth curve modeling : theory and applications*, 1st ed. edition. John Wiley & Sons, Inc., Hoboken, New Jersey.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language documentation and conservation*, 15:187–192.

David A Ratkowsky. 1983. *Nonlinear regression modeling: a unified practical approach*. Marcel Dekker, New York.

Francis J Richards. 1959. A flexible growth function for empirical use. *Journal of experimental Botany*, 10(2):290–301.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 197–204.

Burr Settles. 2009. Active learning literature survey. *Tech. Rep. Computer Sciences Technical Report 1648, University of Wisconsin-Madison*.

Bhargav Shandilya and Alexis Palmer. 2025. Boosting the capabilities of compact models in low-data contexts with large language models and retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7470–7483, Abu Dhabi, UAE. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 623–656.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021a. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Tianze Shi, Adrian Benton, Igor Malioutov, and Ozan İrsoy. 2021b. Diversity-aware batch active learning for dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2616–2626, Online. Association for Computational Linguistics.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–243.

Karl Stratos and Michael Collins. 2015. Simple semi-supervised POS tagging. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 79–87, Denver, Colorado. Association for Computational Linguistics.

Rob van der Goot, Zoey Liu, and Max Müller-Eberstein. 2024. Enough is enough! a case study on the effect of data size for evaluation using Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6167–6176, Torino, Italia. ELRA and ICCL.

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Luby, Christian Luhmann, and H. Andrew Schwartz. 2023. Transfer and active learning for dissonance detection: Addressing the rare-class challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11923–11936, Toronto, Canada. Association for Computational Linguistics.

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-Start Data Selection for Better Few-shot Language Model Fine-tuning: A Prompt-based Uncertainty Propagation Approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2499–2521, Toronto, Canada. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy,

Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Tatiana Merzhevich, Paul Meurer, Niko Miekka, Emilia Milano, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos,

Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2024. Universal dependencies 2.14. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

Yu Zhu, Jinghao Lin, Shibi He, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. 2019. Addressing the item cold-start problem by attribute-driven active learning. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):631–644.

A Languages studied

We list the languages studied here and their language families in Table 2; language family information is taken from The World Atlas of Language Structures (Dryer and Haspelmath, 2013).

Language Family	Language
Indo-European	Afrikaans, Armenian, Belarusian, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Faroese, French, Galician, German, Greek, Hindi, Icelandic, Irish, Italian, Latvian, Lithuanian, Low Saxon, Manx, Norwegian, Persian, Polish, Pomak, Portuguese, Romanian, Russian, Scottish, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Urdu, Vietnamese, Welsh, Western Armenian
Afro-Asiatic	Arabic, Hebrew, Maghrebi, Maltese
Isolate	Basque
Uralic	Erzya, Estonian, Finnish, Hungarian, North Sami
Niger-Congo	Wolof
Turkic	Turkish, Uyghur
English-based	Naija
Koreanic	Korean
Austronesian	Indonesian
Dravidian	Tamil, Telugu
Japonic	Japanese
Sino-Tibetan	Mandarin Chinese

Table 2: The 60 languages and 12 language families studied in our experiments.

B LLM Prompt

Table 3 provides the prompt we used for `GPT-4.1-mini`.

C CRF Implementation

CRF treats POS tagging as a sequence labeling task. We build first-order CRF models (Lafferty et al., 2001) throughout our experiments. All models are implemented with the Python library `crfsuite`. This decision was motivated by two factors. First, prior work has demonstrated CRF to be superior to neural sequence-to-sequence models for sequence tagging task such as morphological segmentation in low-resource settings for a variety of typologically diverse languages (Liu and Dorr, 2024). Second, CRF models, particularly those of lower orders

You are a helpful assistant who is an expert part of speech tagger that works with the Universal Dependencies part-of-speech tagset.
The user will give you a sentence in Irish to be tagged, with one token per line, where each line contains the token's index and the token.
You must provide the tagged sentence in the following format:
1 token_1 tag_1
2 token_2 tag_2
...
n token_n tag_n

You must use only the tags in the following tagset: PROPN NUM SYM ADJ NOUN PRON PUNCT DET INTJ ADV PART ADP X AUX SCONJ CCONJ VERB.

IMPORTANT: the sentence that the user provides has already been tokenized, and each sequence of characters separated by whitespace is a token. DO NOT further split the tokens, and DO NOT join tokens. Also, DO NOT change anything in the sentence provided.

Please provide only the tagged sentence, and nothing else. No explanation, no alternatives, only the tagger output in the format specified above.
SENTENCE:
1 Bhí
2 an
3 geata
4 dúnta
5 .

TAGGED SENTENCE:
1 Bhí VERB
2 an DET
3 geata NOUN
4 dúnta ADJ
5 . PUNCT

.....

Table 3: `GPT-4.1-mini` prompt for POS Tagging.

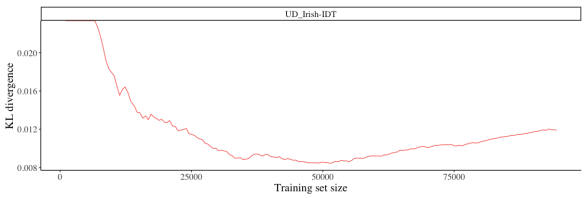

Figure 5: KL divergence between the test set and the training set from each AL iteration for `UD_Irish-IDT`.

(first-/second-order), are less computationally expensive to implement; all models are trained with a single CPU core and 8GB of RAM.[6]

D KL divergence between the training and the test set for `UD_Irish-IDT`

Figure 5 presents results for the KL divergence between the test set and the training set from each AL iteration for `UD_Irish-IDT`.

E AL learning curves for individual tags of `UD_Irish-IDT`

Figure 6 shows the learning curves of individual tags from AL iterations for `UD_Irish-IDT`.

F Growth Curve Model Implementation Results

Given that F1 scores are between 0 and 1, for all growth curve models, we set uniform priors in the same range for the parameters of the upper and lower asymptotes. Since the values for growth rates and delta are always positive, we also set uniform priors for these two parameters with the lower bound of 0. In order to help with model convergence we set 10 as a reasonably high growth rate as the upper bound of the uniform priors for these parameters. In practice the values for the growth rate and delta were below 1, rendering an upper bound of 10 to be suitably uninformative that can be helpful for deriving more objective estimates (Nicenboim et al., 2021). Each growth curve model ran for 4000 iterations. Models that did not converge were re-run for 8000 and 12000 iterations. We excluded the following three treebanks from our statistical analyses since their growth curve models did not converge with higher iterations due to their respective small number of analyzable data points: `UD_Dutch-Alpino`, `UD_Erzya-JR`, and `UD_Irish-TwittIrish`, `UD_Galician-TreeGal`, `UD_German-GSD`, `UD_Tamil-TTB`, `UD_Italian-ParlaMint`, `UD_Turkish-FrameNet`, `UD_Uyghur-UDT`, `UD_Pomak-Philotis`, `UD_Indonesian-CSUI`.

Results for the upper asymptotes estimates with 95% credible intervals from growth curve analysis are presented in Figure 7.

[6]Code for our experiments is available at https://github.com/ufcompling/unlabeled_pos.

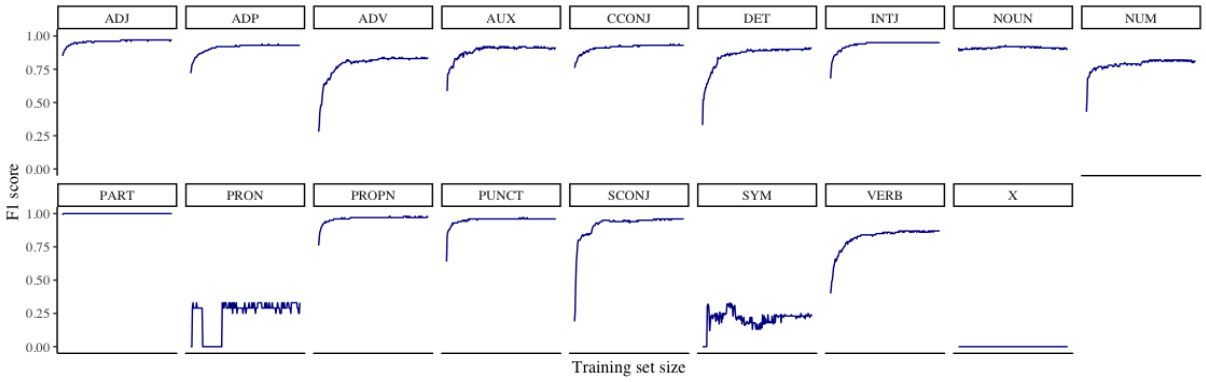

Figure 6: AL learning curves for individual tags of `UD_Irish-IDT`; the x-axes range from 1,000 to 100K tokens.

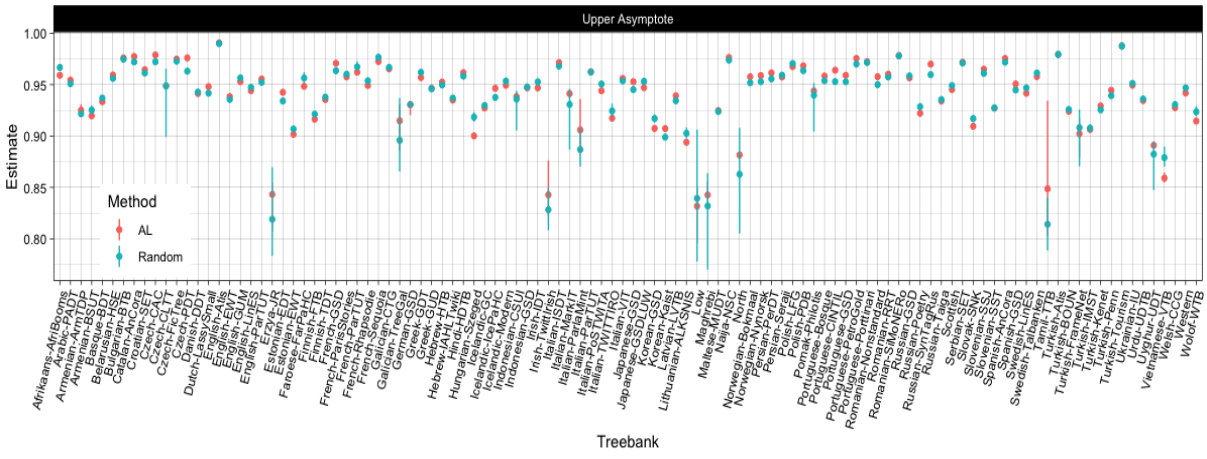

Figure 7: Upper asymptotes with 95% credible intervals from Bayesian growth curve modeling.