5,000 Words is the Optimal Training Size for Low-Resource Morphology

Anonymous ACL submission

Abstract

Utilizing NLP to assist data annotation remains a challenge for low-resource languages. This study shows that the optimal minimal training size for morphology is about 5,000 labeled tokens. After that, annotation results in diminishing returns in terms of model performance. We assess performance improvement in relation to annotated dataset size for a neural Transformer model, a pre-trained llama3 model, and a non-neural CRF model, all of which are utilized in an active learning loop. We analyze corpus diversity (via type-token ratio) to better understand how sample diversity impacts active learning model improvement and find raw TTR scores indicate peak performance.

1 Introduction

Creating initial digital resources in a language requires primarily manual methods that cannot scale to the pace of language endangerment threatening roughly half the world's 7,100+ languages (Eberhard et al., 2025; Seifart et al., 2018) nor to the volume of annotated data demanded for high-performing natural language processing (NLP) models. Even though optimal performance does not wholly depend on data set size (Liu and Prud'hommeaux, 2022; Palmer, 2009), NLP must begin with some amount of annotated data. Unfortunately for low resource languages, "low resource" does not refer just to limited digital language data (Liu et al., 2022). Underfunded academic and community linguists must budget their resources and carefully determine when they have the minimal amount of annotated data that will make optimal contribution towards their goals. The question then is: how much training data should we start with?

This explores how to determine the minimal optimal training data size if no annotated data yet exists. To this end, we use four computational morphology tasks (see 4.1) as the test cases, leveraging *documentary collections from linguistic fieldwork* in six diverse endangered languages. We assume that at some point model performance becomes only marginally better than a model trained on less data. At this point, model performance begins to flatten and annotation efforts yield diminishing returns (e.g., the plateauing trends in Figure 1). Adopting an active learning framework (Settles, 2009), we ask:

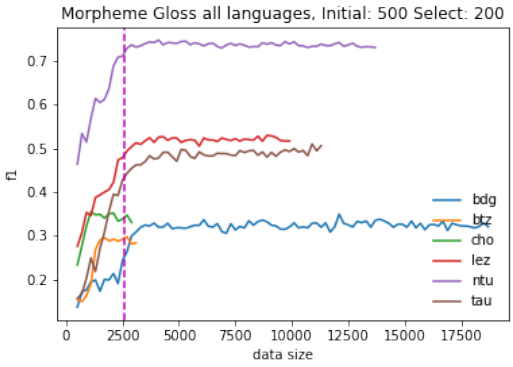

Figure 1: Annotating training data for morphology quickly reaches a point of diminishing returns in terms of model performance. For morpheme glossing with the Transformer the minimal optimal training size is about 2,000 tokens. On average across models, languages, and tasks, the rate of model improvement begins to flatten as total training size grows to about 5,000 tokens.

Q1: How quickly do manual annotation efforts yield diminishing returns in terms of model performance? Moreover, in an active learning scenario, what impact does initial size and selection size have on model improvement?

Q2: What is the minimal optimal size of training data for morphology?

Q3: How might the minimal optimal size be identified by the lexical diversity of training data?

Since large scale annotation is cost prohibitive for minority language communities, identifying the

earliest point of diminishing returns–the minimal optimal training size–guides field linguists who wish to give back to their host language community by making optimal contribution to technological development in the language.

2 Related Work

Investigations into the impact of data size on NLP offer little concrete guidance for budgeting manual annotation or the optimal initial size when developing early digital resources in a new language. If novices search the internet, they will find the most popular answers are "The more the better!" or "It depends".[1][2] Model performance benefits from lots of relevant examples but the ideal number cannot be calculated *a priori* because so many factors can determine how many examples are enough. Popular guidelines to estimate optimal dataset size are premised on the needs of the model, rather than on available resources (Dorfman, 2022; gaumutrapremi, 2025). The "10x Rule" considers computing resources, not available data resources, and recommends having 10 times as many training examples as parameters. Though correct, these answers are unhelpful for low-resource languages.

In low-resource contexts, NLP models and techniques should adjust to available data, not the other way around. Successful adjusted techniques include synthetic data augmentation (Silfverberg et al., 2017; Bergmanis and Goldwater, 2019; Anastasopoulos and Neubig, 2019), or cross-lingual transfer learning (Baumann and Pierrehumbert, 2014; Gu et al., 2018), and active learning (Palmer, 2009). Some research (Vieira et al., 2024; Le Scao and Rush, 2021; Debnath et al., 2021) looks not at minimal training size, but at the most effective number of examples for few-shot prompting, finetuning, or transfer learning from multilingual models. For example, Vieira et al. (2024) found a minimum of 5k examples is optimal for fine-tuning large language models (LLMs) for machine translation. Unfortunately, transfer learning or fine-tuning can yield unreliable results for low-resource languages because even the largest multilingual models fall far short of representing the world's language diversity (Debnath et al., 2021; Costa-Jussà et al., 2022). Additionally, such techniques are best applied only after some minimal training data exists in the language or a closely related language, which is not the case for many communities.

Some work does not assume an initial amount of data exists provide a few insights. Pérez-Mayos et al. (2021) conducted a cost-benefit analysis of unsupervised dataset size on downstream syntactic tasks. Their conclusion simply states "more data is better". In machine translation Maillard et al. (2023) found that across 39 languages if the data is high quality, then diminishing returns being after 6,000 sentence pairs. High quality data is not always obtainable in documentary settings, where field collections tends to be noisy. This number is close to our findings. Looking at unsupervised learning, Eiselen and Gaustad (2023) found that when applying embedding models to part of speech tagging, morphologically complex languages need substantially more data than languages with simpler word structures. This pattern holds fairly well in our experiments, with the polysynthetic languages generally requiring more annotated data.

3 Languages and Data

Our data sets are extracted from realistic, resource-constrained collections of interlinearized glossed texts (IGTs) compiled by documentary linguists in six low-resource languages and shared for research purposes. Table 1 summarizes the original data These IGTs represent some of the only digital resources available in these languages. We do not publish the data on our Github to reduce possible risks to the minority language communities who provided the data. Most the data is publicly available elsewhere or by contacting the authors. We use only locally run models to reduce the risk that the data may be harvested and used for commerical purposes.

To create our datasets, we first constructed the test set by randomly extracting 200 morphologically annotated types (unique word forms), after removing punctuation and digits (unless the digits represent phonemes in the writing system). The training sets contain all remaining morphologically annotated tokens (non-unique words) after all copies of the test words were removed. There is no overlap between the training and test sets. This leads to an overestimation of the data because in realistic linguistic fieldwork, naturalistic language will contain many repeated words. We believe it is better to annotate too much data when chances of organizing additional annotation efforts may be

[1] www.quora.com/How-much-data-is-required-for-machine-learning

[2] www.geeksforgeeks.org/how-much-data-are-sufficient-to-train-my-machine-learning-model

slim. Relevant information and our code is available on Github.[3]

Batak Alas-Kluet [btz] is an vigorous Austronesian language spoken by 100,000 people on the Indonesian island of Sumatra (Eberhard et al., 2025). The selected data is an unpublished corpus from the Alas-Kluet dialect and features reduplication, infixation, and circumfixation.

Bonggi [bdg] is a Western Austronesian language spoken by some 1,500 people in on Banggi and Balambangan islands of Malaysia (Eberhard et al., 2025). It morphologically marks nominal case and verbal agreement (Boutin, 1996).

Choctaw [cho] is spoken by less than 10,000 people primarily in Mississippi and Oklahoma. It has complex morphology but the orthographic representation in the selected data breaks long words into shorter orthographic units (Broadwell, 2006). The original documents are in the United States National Archive and are available for research purposes per the usage policy at Oklahoma Historical Society (OHS).[4] The annotated data can be made available upon request.

Lezgi [lez] (Lezgian) is a Nakh-Daghestanian (Northeast Caucasian) language spoken by over 500,000 speakers in Russia and Azerbaijan (Eberhard et al., 2025). The corpus used is from the Qusar dialect in Azerbaijan (Donet, 2014). The public parts of the corpus are archived in ELAR (The Endangered Languages Archive).[5] Lezgi is a highly agglutinative language with overwhelmingly suffixing morphology (Haspelmath, 1993).

Natügu [ntu] belongs to the Reefs-Santa Cruz group in the Austronesian family. It is spoken by about 5,000 people in the Temotu Province of the Solomon Islands. It has primarily agglutinative morphology with complex verb structures (Næss and Boerger, 2008). The selected corpus contains transcribed narratives and a large written text.[6]

Upper Tanana [tau] (Nee'aaneegn', Nabesna, Tabesna) belongs to the Northern Dene (Athabaskan) family. It is an official language of Alaska, spoken in the eastern interior of Alaska and has been taught at a school in the one village in the Yukon Territory of Canada. Despite its status, it is critically endangered, with about 100 speakers (Lovick, 2020). Its complex morphology features non-continuous lexical, derivational, and inflectional prefixes on verbs. The selected corpus was collected in 2006-2019 in Alaska mostly representing speakers of the Tetlin and Northway dialects. The primary data is preserved at the Alaska Native Language Archive.[7]

4 Experiments

4.1 Four morphological tasks

Four morphological tasks add crucial information for early scientific documentation of a language and automating the tasks supports technological development and the creation language maintenance materials. The tasks are:

(1) Canonical Morpheme Segmentation (Mott et al., 2020): decomposing a word into its component underlying morphemes
input: churches → **output:** church+s

(2) Morpheme Glossing (Okabe and Yvon, 2023): converting a sequence of component morphemes to technical translation (glosses)[8]
input: church+s → **output:** church+PL

(3) Joint segmentation and glossing (Moeller and Hulden, 2021) treats (1) and (2) as a single task, imitating common practice by linguists.
input: churches →
output: church#religious.body + s#PL

(4) Inflection (McCarthy et al., 2019): providing only glosses for the inflectional affixes
input: churches → **output:** PL

4.2 Models

We compare a neural, non-neural, and large language model (LLM). We focus on the Transformer model because it is expected to give lowest performance with the most data. This focus safely

[3] withheld in review version because it is not anonymized
[4] gateway.okhistory.org/help/faq/rights
[5] The direct link is http://hdl.handle.net/2196/439f3247-66f3-4b76-b5e1-8f07f0d7b5a1
[6] A Natqgu grammar and much of the text data are available at https://www.langlxmelanesia.com/tilp
[7] https://www.uaf.edu/anla
[8] 3SG ← third person singular, PRES ← present tense

Language	ISO 639-3	Family	Typology	Tokens
Alas-Kluet	btz	Austronesian	Agglutinative	3,839
Bonggi	bdg	Austronesian	Non-concatenative	21,616
Choctaw	cho	Muskogean	Agglutinative	7,793
Lezgi	lez	Nakh-Daghestanian	Agglutinative	13,953
Natügu	ntu	Austronesian	Agglutinative	16,544
Upper Tanana	tau	Athabaskan	Polysynthetic	14,099

Table 1: Languages in our study; tokens are non-unique words with morphological annotation.

overestimates the minimal optimal amount of training data.

Transformer. For all settings, we employ the Transformer architecture, TRANSFORMER_TINY[9], from the Python package `fairseq`. This model architecture has two encoder-decoder layers and two self-attention heads. The embedding dimension size and that of the feed-forward layer are both 64. This is motivated by Liu and Dorr (2024) which show TRANSFORMER_TINY to be the most effective sequence-to-sequence (seq2seq) model for morphological segmentation in low-resource scenarios across 19 typologically distinct languages. All models were trained using the default parameters, maintaining consistency across experimental settings. All training was performed with an NVIDIA A100 GPU with a single CPU core and 8GB of RAM.

CRF. As a comparison to the encoder-decoder architecture, we adopt conditional random field (CRF), a non-neural log-linear model (Lafferty et al., 2001) that has been applied to morphological segmentation (Cotterell et al., 2016b), inflection (Cotterell et al., 2016a), and glossing (Okabe and Yvon, 2023). Prior work has shown that CRF can perform well for morphological segmentation (Liu and Dorr, 2024), often the task of interest was *surface* segmentation, where the linear concatenation of morphemes refers only to the orthography. Surface segmentation misses important meanings (e.g. that the prefix of the following words is underlyingly identical: il+legal, im+possible, in+complete), as opposed to *canonical* segmentation, which ignores surface orthographic changes during word formation that may be due to phonological processes (e.g. the rendering of the plural suffix without the epenthetic *e* in example (1)). Since our morphological data only consists of canonical segmentation and since seq2seq models have gradually become the default architectures for morphological inflection (Kodner et al., 2022), we apply the CRF only to morphological glossing. We employ first-order CRF, which treats glossing as a sequence labeling task: given a sequence of morphemes in a word, the model predicts each morpheme's English translation. Our hand-curated feature set includes its character-level and morpheme-level unigrams and bigrams, the linear position of the morpheme in the word, and whether the word itself is monomorphemic. All CRF models are implemented with `python-crfsuite` using a single CPU core and 8GB of RAM.[10]

LLM. We additionally assess the tasks on `llama3` (Dubey et al., 2024), an open-source LLM. We used the `ollama` package to run llama3-8b locally with an NVIDIA GeForce RTX 2080 Ti, thereby avoiding any potential data contamination that could result from relying on a closed-source model. The LLM system prompt is stated in Appendix B.

4.3 Machine-in-the-Loop

Machine-in-the-loop methods have been shown to be effective for digital language resources in underdocumented languages by leveraging a less accurate but still helpful NLP model to assist manual annotation (Palmer, 2009; Bird and Yibarbuk, 2024; Moeller and Arppe, 2024). We simulate a machine-in-the-loop approach to morphological analysis by applying active learning (AL), which derives better model performance with less training data by selecting additional batches for manual annotation in informed ways (Settles, 2009; Tharwat and Schenck, 2023). First, a model is trained on available annotated data and used to pre-annotate the remaining unannotated data. Information from the model is assigned to each pre-annotated example and that estimate is used to select a sampling from the data to present to humans experts for manual work. Our selection strategy is least confidence uncertainly sampling (Culotta and McCallum, 2005) which se-

[9] https://fairseq.readthedocs.io/en/latest/models.html

[10] https://python-crfsuite.readthedocs.io/en/latest/

lects the N least confident examples based on the model's entropy or marginal probability estimate. In our case, the new annotated instances are retrieved and added to training set from the IGT to simulate human annotation after each AL cycle.

To determine the relationship between combined training size and model performance, the Transformer and CRF models are trained with different initial training sizes ({500, 1,000, 1,500, 2,000, 2,500}) which represent all annotation efforts before the AL process begins. To determine whether the model improvement is affected by the number of new tokens selected by the least confidence strategy, different selection sizes ({50, 100, 200}) are added to the training set at each AL cycle.[11] The total training size is the sum of the initial training size and selected data from all previous AL iterations. If the initial training size is 500 and the selection size is 200, the initial model is trained on 500 tokens, the second on 700, and so on.

LLM. For the pre-trained LLM strategy approach, we use a slightly adjusted active learning strategy that leverages the MiniLM-L8 embeddings with furthest-point sampling (FPS), an explore-only batch acquisition function from the botorch package (Balandat et al., 2020). This differs from the explore-only AL approach with the CRF and transformer models, which leverage the statistical uncertainty metrics inherent to their training and evaluation. We use in-context learning (ICL) to present llama3-8b with the desired task (eg. morpheme glossing) and relevant training examples, and iteratively assess F1 accuracy for each sample in the test set. For training dataset sizes consisting of greater than 100 examples, we find that the LLM context window can easily be exceeded. We therefore use retrieval augmented generation (RAG) (Lewis et al., 2020) with a MiniLM-L6 embedding model (Wang et al., 2020) to select the 100 most relevant training examples (according to embedding cosine distance) to use for ICL on each test instance. We assess our LLM+RAG approach on the morpheme glossing task across available languages, with an initial randomly selected set of 500 samples. We use our FPS acquisition function to acquire subsequent samples in batches of 200, assessing to a maximum of 2500 tokens. Note that the size of the ICL samples remains fixed at 100, but the samples provided to the LLM change and improve as more tokens are iteratively acquired.

[11] The sizes were chosen based on a pilot study that tested initial training sizes of 50, 100, and 500 tokens and a selection size of 25 tokens and found that an initial training size less than 500 is counterproductive. As the total combined initial+selection size grows past 500, the model performance can actually be lower compared to a model trained on the same total tokens but starting with 500 tokens. We chose the other selection sizes based on what seemed reasonable for an average person to annotate in an hour or less.

4.4 Optimal minimal training size

We hypothesize that for probabilistic supervised models an certain amount of training data exists after which the costs to annotate data are no longer justified because annotation efforts yield diminishing in terms of model performance. Correspondingly, we define the *optimal minimal amount of training data* (hereafter, optimal minimal size) as the point where the rate of model improvement starts to decrease as the amount of annotated data increases monotonically. We estimate the optimal minimal size in three steps. First, we plot the training size x for all combinations of initial and selection sizes and the corresponding model performance y. Second, we find the point of maximum curvature in the plotted line, i.e. the "knee", using the `kneefinder`[12] algorithm, which identifies the point that has the maximum distance from a line passing between the first and last $[x, y]$ coordinates. In reality, the knee is not a single point but an area where the curve begins to flatten. So as the third step, we average the knees across the three selection sizes of each initial size; this average provides a better approximation of the area of maximum curvature since the selection size has minimal impact on the $[x, y]$ curvature (see sec. 5.1). We do this for the Transformer and the CRF results. For LLM assessment, we identify knees on the single selection size and acquisition rate.

4.5 Data diversity

The method of calculating the minimal optimal size described above requires already having more than the minimal optimal size. A more generalizable method would identify the optimal minimal size for any given dataset as annotation efforts are ongoing. In search of such a method, we examine training set diversity with the Transformer and CRF. As a measure of diversity, we adapt a simple measure of lexical diversity, the type to token ratio (TTR). In corpus linguistics, "type" refers to a distinct word form and "token" refers to all word forms in a data set (including duplicates). TTR is calculated by

[12] https://github.com/vlavorini/kneefinder

dividing the total number of types by the total number of tokens and multiplying by 100 to achieve a percentage. For example, the TTR of the sentence "They saw eye to eye." is 80% because eye is repeated twice, giving ,000 types in a sentence with 5 tokens. We calculate the TTR of each combined training set after every AL cycle and compare the curve trajectories to the model performance trajectory. We measured TTR on the input and output for each task in examples (1)-(4), butthe TTR of outputs did not provide useful information, so we consider only the TTR of the inputs in 5.4.

5 Results

Having compared all combinations of initial and selection data sizes in an AL scenario for all four tasks and all languages, we present the results and discuss each research question.

5.1 Q1: Annotation yields diminishing returns

As Figure 1 indicates, annotation efforts rather quickly yield diminishing returns in model performance across every language and nearly every task. This tendency towards a logarithmic relationship (with an upper bound) between data size and model performance holds less strongly for inflection and for the two smaller datasets. Alas-Kluet [btz] and Choctaw [cho] do not seem to contain their optimal minimal size, particularly for inflection. This is more clearly seen with the CRF which tends to take longer to show leveling of performance (and yields slightly higher peak F1 scores).

Initial training size With AL, the amount of initial training data only seems to impact how quickly optimal model performance is reached.For example, starting Alas-Kluet glossing at 500 tokens (and a selection size of 100), the model tops out after ten AL cycles (1,500 total training tokens). Starting with 2000 tokens requires only three iterations to reach the same performance, but more total training tokens.[13]

Selection size Overall, smaller selection size have a slight positive impact on the rate of model improvement. With the Transformer a selection sizes of 50 reaches the highest overall model performance about 30% of the time and occasionly levels out slightly earlier. With the CRF, however, the smaller selection sizes almost always achieve best performance and level out earlier (about 65% and 45% of of the experiments, respectively).

[13]Figures showing all settings and training combinations can be found on our GitHub repository.

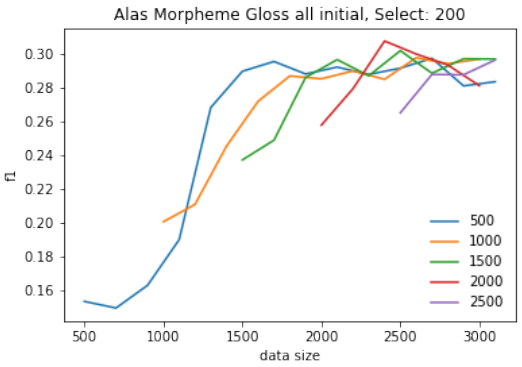

Figure 2: All initial training sizes achieve similar model performance.

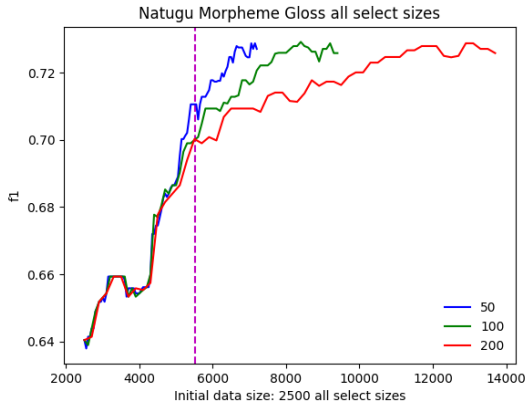

Figure 3: The rate of model improvement diverges slightly with the CRF depending on selection size during AL but does not impact peak performance.

5.2 Q2: Optimal minimal size is 5k

Considering all combined sizes with the Transformer and CRF, the optimal minimal training size ranges 1,000-12,000 (see Appendix A). The higher amounts are due to larger initial training sizes. The average optimal minimal size is just under 4,000 training tokens with the Transformer and around 5,000 tokens with the CRF. The non-neural model also yields slightly higher F1 scores which is typical of low-resource settings, unless/until strategies such as data augmentation are applied. Alas-Kluet and Choctaw have datasets smaller than these averages and may not contain the optimal minimal size. However, even without those two languages, the average point of diminishing returns from annotation increases slightly to about 4,500 tokens (Transformer) and 6,000 (CRF) tokens. Overall,

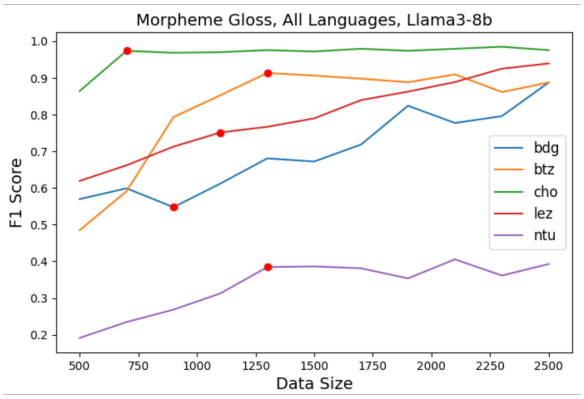

Figure 4: Llama3-8b performance on morpheme glossing with active learning selection and retrieval of relevant examples for ICL. Red points show the approximate region where a knee may be located.

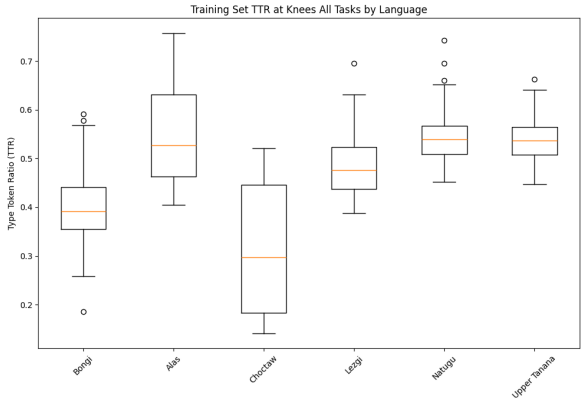

Figure 5: Average TTR scores at point of performance plateauing suggest a raw TTR score of 50% or below indicates the minimal optimal training size is reached.

the average optimal minimal training size is about 5,000 tokens.

5.3 Q1 & Q2: LLM results

Based on preliminary assessments with our pre-trained LLM, we find that for most languages the peak F1 score is significantly higher than the statistical or fine-tuned transformer strategies. Figure 4 shows the F1 scores across languages for the morpheme glossing task, ranging in dataset size between 500 and 2500 labeled tokens. Natügu is a notable exception, with F1 score peaking at around 0.78 for the Transformer model but at only around 0.40 for the LLM. We also find that, according to the `kneefinder` algorithm, the data size at the point of diminishing returns may be far less than 5,000 tokens for the morpheme glossing task. This is especially the case for Choctaw, where the pre-trained LLM appears saturated at an F1 score of around 0.95 after being provided only 700 labeled tokens. While the Transformer model also peaks at around 700 labelled tokens for the morpheme glossing task, the F1 score remains at only 0.35. In future efforts, we will assess the pre-trained LLM across additional tasks (segmentation and inflection) with larger data sizes to determine whether, for languages like Bonggi and Lezgi, the pre-trained LLM reaches a point of diminishing returns at a later point.

5.4 Q3: TTR indicates optimal training size

After examining changes in data diversity, we find that raw measures of TTR is a useful proxy for optimal minimal size with the Transformer and CRF. Each task follows a unique pattern of TTR change before showing a final steady decline and the patterns hold across languages and models. Except for segmentation, however, these patterns are not reliable indicators of optimal minimal size. However, raw TTR score is a reasonably reliable indicator that one approaching the optimal minimal size of any given dataset because this size has a very strong tendency to when the TTR score declines to 50% or less especially in the larger datasets, as can be seen in Figure 5. These tendency is more obvious after eliminating the two smallest datasets.

6 Conclusion

This work ascertains the minimum amount of labeled low-resource language data that ought to be acquired manually before expecting statistical, neural, and large language models to provide optimal linguistic assistance. Applying active learning to four morphology tasks in six diverse low-resource languages and using available annotated documentary data, we compare model performance of CRF, Transformer, and llama3. We find that the average minimal optimal training size is about 5,000 words. Our work suggests that raw TTR scores of 50% or less can indicate when the optimal minimal size has been reached during the AL process.

A note on Choctaw and Alas-Kluet: The two data sets under 10k available tokens reached performance plateau more quickly than the larger data sets with the Transformer or not all with the CRF. We conclude that 2-3,000 training tokens is not enough even though we see a leveling of model performance in those datasets. They showed less clear correlation between TTR scores and optimal minimal size. The larger datasets, on the other

hand, peaked fairly consistently shortly after 4,000 (Transformer) or 6,000 (CRF) training tokens and did not show marked increase even after another 10 thousand tokens.

Interestingly, 5,000 tokens corresponds nicely with the size of two IGT collections (btz and lez) created by graduate student language documentation projects. (Lezgi's annotation has since expanded.) Although too small to be conclusive we tentatively conclude that early documentary efforts may be producing the minimal optimal training data needed to support initial NLP development. We also recommend starting with an initial size of 500 and using an active learning approach with a smaller selection size of 50 or 100. We acknowledge smaller selection batches demands more frequent computing time and other logistical matters which could slow the journey to optimal performance. In realistic scenarios, the practicality of frequent retraining should be considered.

This paper does not answer a related question, namely "When with AI be useful for my language?" An optimal trade-off between data size and model performance does mean the model is yet useful. At this point we suggest academic and community linguist should cease annotation efforts and seek collaboration with experts in low-resource NLP. Our ongoing work is evaluating when a models becomes useful as an annotation assistant.

Limitations

Due to time constraints we were not able to run k-fold or average across random seeds. This means the numbers and scores may not hold well as an average performance. We were also not able to run the pre-trained LLM for larger dataset sizes and across segmentation and inflection tasks. Although the set of languages is diverse in morphological, geographical, and genealogical categories, six is a small number of languages. The conclusions we draw should be tested further on other *realistic documentary field data*.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the Limits of Low-Resource Morphological Inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. 2020. Botorch: A framework for efficient monte-carlo bayesian optimization. *Advances in neural information processing systems*, 33:21524–21538.

Peter Baumann and Janet Pierrehumbert. 2014. Using Resource-Rich Languages to Improve Morphological Analysis of Under-Resourced Languages. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Toms Bergmanis and Sharon Goldwater. 2019. Training Data Augmentation for Context-Sensitive Neural Lemmatizer Using Inflection Tables and Raw Text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4119–4128, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Bird and Dean Yibarbuk. 2024. Centering the Mixtec speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839.

Michael Boutin. 1996. Bonggi clitics. In *Pan-Asiatic Linguistics: Proceedings of the fourth International Symposium on languages and linguistics*, volume 1.

George Aaron Broadwell. 2006. *A Choctaw reference grammar*. Studies in the anthropology of North American Indians. University of Nebraska Press, Lincoln.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016a. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016b. Morphological segmentation inside-out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas. Association for Computational Linguistics.

Aron Culotta and Andrew McCallum. 2005. Reducing Labeling Effort for Structured Prediction Tasks:. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference*,

Pittsburgh, Pennsylvania, USA. American Association for Artificial Intelligence.

Arnab Debnath, Navid Rajabi, Fardina Fathmiul Alam, and Antonios Anastasopoulos. 2021. Towards more equitable question answering systems: How much more data do you need? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 621–629, Online. Association for Computational Linguistics.

Charles Donet. 2014. The Importance of Verb Salience in the Followability of Lezgi Oral Narratives. Master's thesis, Graduate Institute of Applied Linguistics, Dallas, TX.

Eugene Dorfman. 2022. How Much Data Is Required for Machine Learning? – PostIndustria.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? A case study of three linguistically distinct South African languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53, Dubrovnik, Croatia. Association for Computational Linguistics.

gaumutrapremi. 2025. What is the minimum amount of parallel corpora needed for Machine Translation of Extremely Low Resource Ancient Language.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Haspelmath. 1993. *A grammar of Lezgian*. Mouton de Gruyter, Berlin; New York.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, and 11 others. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

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.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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 and Emily Prud'hommeaux. 2022. Data-driven model generalizability in crosslinguistic low-resource morphological segmentation. *Transactions of the Association for Computational Linguistics*, 10:393–413.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. 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.

Olga Lovick. 2020. *A Grammar of Upper Tanana, Volume 1: Phonology, Lexical Classes, Morphology*. U of Nebraska Press, Lincoln. Google-Books-ID: jdrKDwAAQBAJ.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-loop with documentary and descriptive linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2021. Integrating automated segmentation and glossing into documentary and descriptive linguistics. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 86–95, Online. Association for Computational Linguistics.

Justin Mott, Ann Bies, Stephanie Strassel, Jordan Kodner, Caitlin Richter, Hongzhi Xu, and Mitchell Marcus. 2020. Morphological segmentation for low resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3996–4002, Marseille, France. European Language Resources Association.

Åshild Næss and Brenda H. Boerger. 2008. Reefs–Santa Cruz as Oceanic: Evidence from the Verb Complex. *Oceanic Linguistics*, 47(1):185–212.

Shu Okabe and François Yvon. 2023. Towards multilingual interlinear morphological glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

Alexis Mary Palmer. 2009. *Semi-automated annotation and active learning for language documentation*. PhD Thesis, University of Texas at Austin.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

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

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data Augmentation for Morphological Reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99.

Alaa Tharwat and Wolfram Schenck. 2023. A Survey on Active Learning: State-of-the-Art, Practical Challenges and Research Directions. *Mathematics*, 11(4):820.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA. Association for Machine Translation in the Americas.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

A Minimal Optimal Sizes

(see next page)

Table 2: The minimal optimal sizes for all four tasks with the Transformer; the minimal optimal size for each initial training size is the average of the knees (point of maximum curvature of plotted f1 scores) across all three selection sizes.

Language	ISO 639-3	Task	Initial Size	Optimal Size
Alas-Kluet	btz	morpheme segmentation	500	1666
Alas-Kluet	btz	morpheme segmentation	1000	1833
Alas-Kluet	btz	morpheme segmentation	1500	2333
Alas-Kluet	btz	morpheme segmentation	2000	2400
Alas-Kluet	btz	morpheme segmentation	2500	2700
Alas-Kluet	btz	morpheme glossing	500	1450
Alas-Kluet	btz	morpheme glossing	1000	1750
Alas-Kluet	btz	morpheme glossing	1500	2000
Alas-Kluet	btz	morpheme glossing	2000	2383
Alas-Kluet	btz	morpheme glossing	2500	2633
Alas-Kluet	btz	joint segmentation & glossing	500	1566
Alas-Kluet	btz	joint segmentation & glossing	1000	1866
Alas-Kluet	btz	joint segmentation & glossing	1500	2350
Alas-Kluet	btz	joint segmentation & glossing	2000	2583
Alas-Kluet	btz	joint segmentation & glossing	2500	2750
Alas-Kluet	btz	inflection	500	1500
Alas-Kluet	btz	inflection	1000	1666
Alas-Kluet	btz	inflection	1500	2100
Alas-Kluet	btz	inflection	2000	2500
Alas-Kluet	btz	inflection	2500	2750
Bonggi	bdg	morpheme segmentation	500	6583
Bonggi	bdg	morpheme segmentation	1000	6283
Bonggi	bdg	morpheme segmentation	1500	6300
Bonggi	bdg	morpheme segmentation	2000	6033
Bonggi	bdg	morpheme segmentation	2500	6200
Bonggi	bdg	morpheme glossing	500	3316
Bonggi	bdg	morpheme glossing	1000	3383
Bonggi	bdg	morpheme glossing	1500	3450
Bonggi	bdg	morpheme glossing	2000	3966
Bonggi	bdg	morpheme glossing	2500	4583
Bonggi	bdg	inflection	500	4116
Bonggi	bdg	inflection	1000	3200
Bonggi	bdg	inflection	1500	4083
Bonggi	bdg	inflection	2000	6000

Continued on next page

Table 2: The minimal optimal sizes for all four tasks with the Transformer; the minimal optimal size for each initial training size is the average of the knees (point of maximum curvature of plotted f1 scores) across all three selection sizes. (Continued)

Language	**ISO 639-3**	**Task**	**Initial Size**	**Optimal Size**
Bonggi	bdg	inflection	2500	6050
Bonggi	bdg	joint segmentation & glossing	500	6583
Bonggi	bdg	joint segmentation & glossing	1000	6316
Bonggi	bdg	joint segmentation & glossing	1500	**7016**
Bonggi	bdg	joint segmentation & glossing	2000	6366
Bonggi	bdg	joint segmentation & glossing	2500	5750
Choctaw	cho	morpheme segmentation	500	2816
Choctaw	cho	morpheme segmentation	1000	3233
Choctaw	cho	morpheme segmentation	1500	3766
Choctaw	cho	morpheme segmentation	2000	3833
Choctaw	cho	morpheme segmentation	2500	4550
Choctaw	cho	morpheme glossing	500	1016
Choctaw	cho	morpheme glossing	1000	1333
Choctaw	cho	morpheme glossing	1500	1783
Choctaw	cho	morpheme glossing	2000	2116
Choctaw	cho	morpheme glossing	2500	2716
Choctaw	cho	joint segmentation & glossing	500	1383
Choctaw	cho	joint segmentation & glossing	1000	1450
Choctaw	cho	joint segmentation & glossing	1500	2116
Choctaw	cho	joint segmentation & glossing	2000	2333
Choctaw	cho	joint segmentation & glossing	2500	2716
Choctaw	cho	inflection	500	**966**
Choctaw	cho	inflection	1000	1383
Choctaw	cho	inflection	1500	2216
Choctaw	cho	inflection	2000	2316
Choctaw	cho	inflection	2500	2750
Lezgi	lez	morpheme segmentation	500	5483
Lezgi	lez	morpheme segmentation	1000	5583
Lezgi	lez	morpheme segmentation	1500	5250
Lezgi	lez	morpheme segmentation	2000	5583
Lezgi	lez	morpheme segmentation	2500	5950
Lezgi	lez	morpheme glossing	500	2700
Lezgi	lez	morpheme glossing	1000	2966
Lezgi	lez	morpheme glossing	1500	3000

Continued on next page

Table 2: The minimal optimal sizes for all four tasks with the Transformer; the minimal optimal size for each initial training size is the average of the knees (point of maximum curvature of plotted f1 scores) across all three selection sizes. (Continued)

Language	ISO 639-3	Task	Initial Size	Optimal Size
Lezgi	lez	morpheme glossing	2000	3316
Lezgi	lez	morpheme glossing	2500	3783
Lezgi	lez	joint segmentation & glossing	500	4416
Lezgi	lez	joint segmentation & glossing	1000	4450
Lezgi	lez	joint segmentation & glossing	1500	4416
Lezgi	lez	joint segmentation & glossing	2000	4600
Lezgi	lez	joint segmentation & glossing	2500	4483
Lezgi	lez	inflection	500	3300
Lezgi	lez	inflection	1000	2566
Lezgi	lez	inflection	1500	2850
Lezgi	lez	inflection	2000	4016
Lezgi	lez	inflection	2500	4033
Natügu	ntu	morpheme segmentation	500	5700
Natügu	ntu	morpheme segmentation	1000	5950
Natügu	ntu	morpheme segmentation	1500	5566
Natügu	ntu	morpheme segmentation	2000	6000
Natügu	ntu	morpheme segmentation	2500	6083
Natügu	ntu	morpheme glossing	500	2516
Natügu	ntu	morpheme glossing	1000	3350
Natügu	ntu	morpheme glossing	1500	3633
Natügu	ntu	morpheme glossing	2000	3933
Natügu	ntu	morpheme glossing	2500	4683
Natügu	ntu	joint segmentation & glossing	500	5233
Natügu	ntu	joint segmentation & glossing	1000	5300
Natügu	ntu	joint segmentation & glossing	1500	4683
Natügu	ntu	joint segmentation & glossing	2000	5050
Natügu	ntu	joint segmentation & glossing	2500	5383
Natügu	ntu	inflection	500	3483
Natügu	ntu	inflection	1000	4533
Natügu	ntu	inflection	1500	5200
Natügu	ntu	inflection	2000	5283
Natügu	ntu	inflection	2500	5433
Upper Tanana	tau	morpheme segmentation	500	5716
Upper Tanana	tau	morpheme segmentation	1000	5616

Continued on next page

Table 2: The minimal optimal sizes for all four tasks with the Transformer; the minimal optimal size for each initial training size is the average of the knees (point of maximum curvature of plotted f1 scores) across all three selection sizes. (Continued)

Language	**ISO 639-3**	**Task**	**Initial Size**	**Optimal Size**
Upper Tanana	tau	morpheme segmentation	1500	5933
Upper Tanana	tau	morpheme segmentation	2000	5883
Upper Tanana	tau	morpheme segmentation	2500	5400
Upper Tanana	tau	morpheme glossing	500	3116
Upper Tanana	tau	morpheme glossing	1000	3066
Upper Tanana	tau	morpheme glossing	1500	3400
Upper Tanana	tau	morpheme glossing	2000	3383
Upper Tanana	tau	morpheme glossing	2500	4050
Upper Tanana	tau	joint segmentation & glossing	500	4083
Upper Tanana	tau	joint segmentation & glossing	1000	4616
Upper Tanana	tau	joint segmentation & glossing	1500	4066
Upper Tanana	tau	joint segmentation & glossing	2000	4483
Upper Tanana	tau	joint segmentation & glossing	2500	4766
Upper Tanana	tau	inflection	500	3600
Upper Tanana	tau	inflection	1000	2883
Upper Tanana	tau	inflection	1500	3266
Upper Tanana	tau	inflection	2000	3983
Upper Tanana	tau	inflection	2500	4633

B LLM Prompting

The llama3 system prompt is provided as follows, with curly braces representing where the language, task, and input/output pairs are specified.

You are a linguist specializing in the language {language}. You are given words and tasked with {task}. Here is a training set of {N} word inputs (left) and their expected {task} outputs (right):

{input A} {output A}

{input B} {output B}

...

Now, I will give you a list of words. For each word, provide only the {task} output on a separate line, in the same order as the input words. Do not include the input words in your response. Output one result only. It is absolutely critical that you only return the output, no other text or formatting.