Resisting the Lure of the Skyline: Grounding Practices in Active Learning for Morphological Inflection

Saliha Muradoğlu[♠] **Michael Ginn**[♣] **Miikka Silfverberg**[♦] **Mans Hulden**[♣]
[♠]The Australian National University (ANU) [♣]University of Colorado Boulder
[♦]University of British Columbia
Firstname.Lastname@ {[♠]anu.edu.au, [♣]colorado.edu, [♦]ubc.ca}

Abstract

Active learning (AL) aims to reduce the burden of annotation by selecting informative unannotated samples for model building. In this paper, we explore the importance of conscious experimental design in the language documentation and description setting, particularly the distribution of the unannotated sample pool. We focus on the task of morphological inflection using a Transformer model. We propose context motivated benchmarks: a baseline and skyline. The baseline describes the frequency weighted distribution encountered in natural speech. We simulate this using Wikipedia texts. The skyline defines the more common approach, uniform sampling from a large, balanced corpus (UniMorph, in our case), which often yields mixed results. We note the unrealistic nature of this unannotated pool. When these factors are considered, our results show a clear benefit to targeted sampling.

1 Introduction

Active learning (AL) (Cohn et al., 1996) is a data annotation approach, where the aim is to direct annotation effort at examples that are maximally helpful for model performance. Most active learning work in NLP involves **pool-based active learning** (McCallum et al., 1998) where a small seed training set is used to create an initial model, and additional examples are selected and annotated from a large pool of unannotated data. Several selection strategies exist, including confidence-based (Lewis, 1995; Cohn et al., 1996; Muradoglu and Hulden, 2022), diversity-based (Brinker, 2003; Sener and Savarese, 2018; Yuan et al., 2020) and committee-based approaches (Liere and Tadepalli, 1997; Farouk Abdel Hady and Schwenker, 2010); these approaches aim to outperform a uniform random selection baseline.

AL is often advocated as a method to rapidly improve model performance in low-resource settings (Baldridge and Palmer, 2009; Ambati, 2012; Grießhaber et al., 2020), where limited annotation capacity needs to be directed intelligently. Nevertheless, AL performance is inconsistent in practice and both success-stories and failures are reported in the literature (Settles et al., 2008; Baldridge and Palmer, 2009; Althammer et al., 2023), demonstrating that it is non-trivial to beat a uniform random selection baseline.

Language documentation is a natural application for active learning. Approximately half the world's languages face the grim forecast of extinction, with

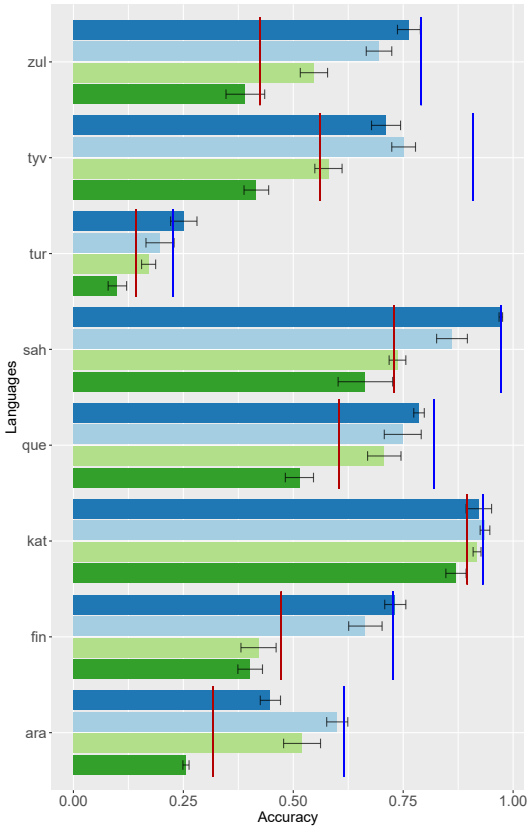

Figure 1: Accuracies reported across the eight languages considered, for the seed, LMC(LEMMA,MSD), LMC(WORDFORM, MSD) and TMC experiments. The maroon lines mark the Frequency Stratified (Baseline) accuracy, and the blue lines mark the Uniform sampling (Skyline) accuracy.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 47–55

around 35–42% of these still substantially undocumented (Krauss, 1992; Wurm, 2001; Bianco, 2002; Crystal, 2002; Austin and Sallabank, 2011; Seifart et al., 2018). However, data for training automated systems is often limited, and additional annotation bears a high opportunity cost, limited not only by resources but also native speaker availability.

Simulated active learning The gold standard of active learning experiments for language documentation is the use of human annotators in a genuine low-resource setting, as in studies such as Baldridge and Palmer (2009). However, for practical reasons, most AL research uses **simulated active learning**, where a small seed training set is sampled from a large existing annotated dataset, and the remaining annotated examples represent the pool from which new examples are selected. While this approach allows for experimentation without costly manual annotations, it introduces a number of confounding factors which can complicate interpretation of results.

Baldridge and Palmer (2009) note that *unit annotation cost* is generally assumed in simulated active learning experiments, but this approach can be unrealistic when selection strategies tend to choose ambiguous examples that are harder, and therefore slower, to annotate. In a similar vein, Margatina and Aletras (2023) argue that in simulations, the unannotated pool tends to be carefully curated and preprocessed (as it is formed from an existing annotated training set). These pools often display unrealistic distributions of classes and lexical and structural diversity, which can be a highly inaccurate reflection of data in the wild, where noise, irrelevant examples and repetitions abound. To ensure validity of the results of simulated active learning experiments (particularly for low-resource settings), it is important to mimic a setting with limited lexical diversity and characteristic class imbalance, as is present in natural language datasets.

Active learning for morphology In this paper, we analyze pool-based active learning for language documentation, focusing on models for morphological inflection. We first argue that existing type-level morphological resources (such as Unimorph, Batsuren et al. 2022) are a poor representation of a realistic unannotated pool in language documentation settings, unless some notion of lexical frequency is injected into the data. We then present experiments on morphological inflection, which demonstrate that the composition of the unannotated pool is highly influential for performance in simulated active learning experiments.

We employ two selection criteria: **transformer model confidence** as previously investigated by Muradoglu and Hulden (2022) and a novel **language model-based selection criterion**. Given a carefully designed, frequency stratified, pool of unannotated examples mimicking naturalistic text, these methods can beat a uniform random baseline by a sizable margin. However, given a naïvely constructed, unannotated pool (based on the Unimorph database), neither of the methods confers an advantage over the baseline.

2 Data

We conduct experiments on the UniMorph database of inflection tables (Batsuren et al., 2022)[1] on a typologically diverse set of eight languages: Arapaho (arp), Finnish (fin), Georgian (kat), Quechua (que), Sakha (sah), Turkish (tur), Tuvan (tyv) and Zulu (zul). Our choice of languages is motivated by a balance between morphological complexity, data availability (both UniMorph and Wikipedia) and endangerment classification according to UNESCO Atlas of the World's Languages in Danger (Moseley, 2010). Where possible, we have attempted to maximise the diversity of our subject languages. Across 8 languages, 6 language families[2]. Further, three of the languages considered (sah, tyv and arp) are considered endangered. We exclusively include adjectives and nouns in our experiments.[3] This simplifies analysis while still representing substantial morphological diversity as nouns make up a sizable portion of text cross-linguistically (Hudson, 1994; Liang and Liu, 2013).

To model word frequencies, we extract the Wikipedias for each language and form the intersection of word types present in UniMorph (U) and Wikipedia (W): U ∩ W. We also retain the much larger part of the UniMorph database U \ W, representing types not found in the Wikipedia. Data sampling is visualized in Figure 2.[4] Our development and initial seed training set are formed by sampling (without replacement) 500 and 1,000

[1] Released under the CC BY-SA 3.0 license

[2] Uralic, Kartvelian, Turkic (South Siberian, North Siberia, Western Oghuz), Quechuan, Algonquian, Bantu.

[3] If these inflect identically, we combine them into a category of nominals. See Table 4 for details.

[4] All data and code will be made available at https://github.com/michaelpginn/active-learning-for-morphology/. Code released under the MIT license.

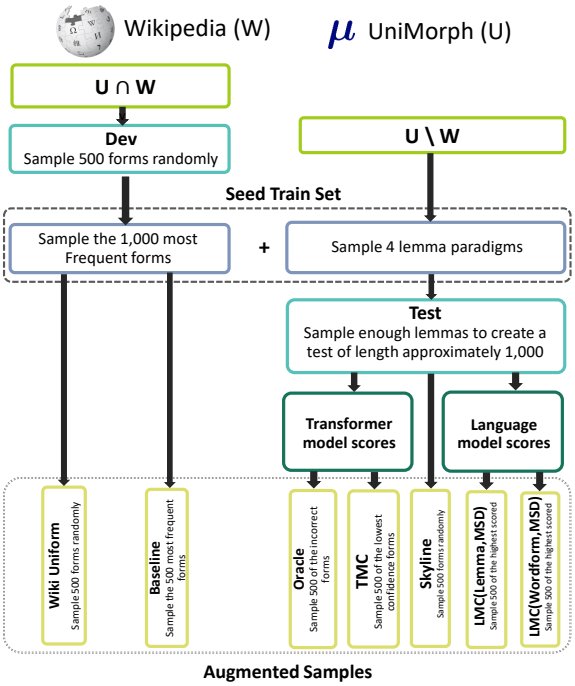

Figure 2: Overview of data sampling, where U and W notes the UniMorph and Wikipedia databases respectively, U ∩ W denotes the intersection and U \ W the difference. Arrows note a sampling without replacement.

forms, respectively, from U ∩ W (with their UniMorph lemmata and MSDs). For each language, we additionally supplement the seed training set with four complete inflection paradigms extracted from U \ W to ensure that all inflections are covered by the seed training data[5]. From the remaining types in U \ W, we then sample 1,000 for testing. Thus, we ensure that there is no overlap between the data splits.

Using each of the different active learning selection strategies presented below in Section 3, we sample an additional 500 training examples. For our baseline method, those are sampled from U ∩ W, while for all the other methods, additional data comes from U \ W.

3 Experimental Setup

We perform experiments on the word inflection task (Cotterell et al., 2016; Goldman et al., 2023) with datasets consisting of triplets ⟨lexeme, MSD, inflected form⟩, e.g. ⟨`smile, V;PST, smiled`⟩. Models are trained to predict the correct inflected form based on the lemma and MSD. We train transformer (Vaswani et al., 2017) inflection models using *fairseq* (Ott et al., 2019).[6] In all experiments, we apply data augmentation using the lemma-copy mechanism (Liu and Hulden, 2022). We initially train models on the seed training set and use various sampling strategies to select 500 additional examples from the unused pool, evaluating the change in inflection performance when training on the augmented set. The test and development sets, disjoint with all training data, remain unchanged through this process.

We experiment with the following strategies:

Frequency Stratified (Baseline) We use word frequencies from Wikipedia to perform weighted random sampling from the pool U ∩ W. This method serves as a linguistically motivated, realistic baseline, accounting for the Zipfian nature of language, and approximating realistic lexical diversity and the naturalistic distribution of inflected forms.

Wiki Uniform We additionally report results on a baseline which samples from U ∩ W without frequency weighting.

Uniform sampling (Skyline) Our second baseline (which we call *Skyline*, as it is near-unbeatable) uses uniform sampling without word frequency information from U \ W. This setting is unrealistic in a language documentation setting—due to the lexical diversity and balanced class distribution of the samples, rare paradigm slots are over-represented.

Oracle Inspired by Muradoglu and Hulden (2022), we sample forms which the model fails to inflect correctly. Since this requires knowledge of gold standard forms, the method can only be used for comparison. This strategy mimics feedback from a linguist or language expert. In many cases, there are more than 500 incorrectly inflected forms to choose from. When this happens, we select maximally erroneous examples, that is, the examples with the greatest Levenshtein distance to the gold standard form.[7] In contrast, when there are fewer than 500 incorrectly inflected forms, we augment the set using correctly inflected forms with the lowest confidence.

Transformer model confidence (TMC) Again following Muradoglu and Hulden (2022), we train an initial inflection model on the seed training set. We use this model to make predictions and select the examples with the lowest confidence scores.

[5] In a language documentation setting, this information could be supplied by the linguist.

[6] Our model and training hyperparameters follow Liu and Hulden (2020), described in Appendix A.

[7] This can be thought of as maximizing the informativity of the examples.

Language model confidence scores (LMC) We train two character-level language models (LM) over lemma+MSD and wordform+MSD sequences (respectively) from the seed training set. This means that our LMs return probabilities for sequences like *walk+V+PAST* and *walked+V+PAST*. We use the LMs to select examples with low probability or, equivalently, high negative log-likelihood (NLL).[8] We experiment with using NLL from either the input lemma or the predicted inflected forms (not gold forms), and term these approaches LMC(LEMMA,MSD) and LMC(WORDFORM,MSD), respectively.

4 Results and Discussion

Experiment	Δ accuracy
Baseline	0.067
Wiki Uniform	0.122
LMC(lemma,MSD)	0.124
Oracle	0.193
LMC(Wordform,MSD)	0.230
TMC	0.247
Skyline	0.298

Table 1: Average change in accuracy observed across each sampling strategy.

Table 1 reports the average change in accuracy from the seed models for each sampling strategy. The two benchmarks provide upper and lower limits for sample selection. The baseline underperforms on average, an expected result given the Zipfian nature of language. As the sampling strategy is dependent on natural texts, the samples have less diverse lemmas and MSDs. Meanwhile, the skyline outperforms every other strategy for five of the eight languages; again, this result is unsurprising, as the UniMorph database provides highly diverse examples. However, it is nearly impossible to replicate this approach, which treats all words equally regardless of rarity, in a realistic setting.

While the WIKI UNIFORM strategy shows greater average improvements than the baseline, the results across languages are mixed[9]. For example, while Finnish shows a 28.8% accuracy gain, performance on Quechua decreases by 0.05%.

It is surprising that the oracle, intended to mimic a language expert, is outperformed by the either the TMC or LMC(WORDFORM, MSD) strategies for six of the languages considered. For almost all of the languages examined, the Levenshtein distance is the primary weighing factor[10]. The edit distance fails to consider the diversity of vocabulary or MSD. Compound words can also skew the Levenshtein distance significantly. For example, for the Turkish compound *otomatik bilet makinası* ("automatic ticket machine"), if the model does not capture the space between *otomatik* and *bilet*, though characters are merely shifted to the left, the Levenshtein distance is artificially high.

4.1 Edit Diversity

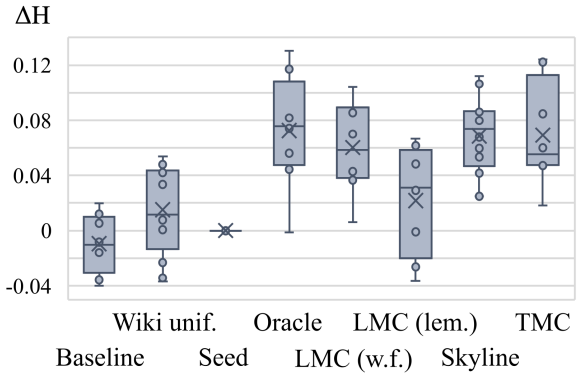

Figure 3: Change in edit diversity (H), compared to the base train set, for each sampling method. While the baseline method leads to reduced edit diversity, most of the sampling methods instead result in increased diversity.

We seek to understand the effects of the various sampling strategies by estimating the relative *edit diversity* for each sample. For each dataset, we enumerate the edits (insertion, deletion, or replacement of subwords) needed to transform each lexeme to the inflected word. We collect edits of the same type and subword to give an edit distribution. Using this distribution, we compute entropy, which is higher for a distribution with a more diverse set of edits, and lower when the dataset is dominated by a few frequent edits. We provide the entropy, relative to the base training set, in Figure 3.

We observe that the strategies that sample from Wikipedia (which tend to be less successful) have lower entropy on average, while the Oracle, TMC, and skyline samples (which are more successful)

[8]This approach is inspired by the observation that novel words are often inflected based on analogy to know words (Skousen, 1990; Derwing and Skousen, 1994; Prasada and Pinker, 1993). The LMC approach aims to seek out examples which are not represented by the seed training set.

[9]See Table 5 for details.

[10]Since there are more than 500 incorrect predictions for the remaining U \ W dataset. The only exception is Georgian, with < 500 incorrect predictions.

have higher entropy. We also find correlations between lower *cross-entropy* with the test set and better performance (see section 4.1.1).

The distinction between the naïve UniMorph pool and the frequency stratified sampling is mirrored in the language documentation and description (LDD) community with the elicitation or naturalistic speech debate. Chelliah (2001) notes that *'language description based solely on textual data results in patchy and incomplete descriptions'*. Similarly, Evans (2008) highlights the necessity of both linguistic phenomena targeting elicitation and observed communicative events[11] (often narratives, conversations, etc.).

4.1.1 Cross-entropy and performance

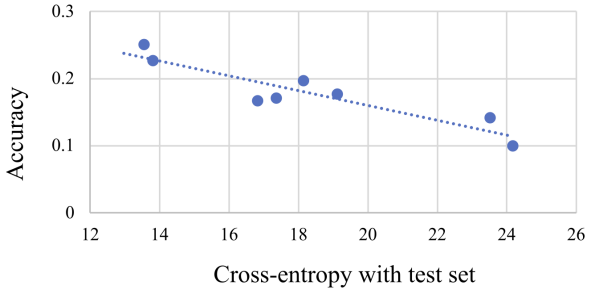

Figure 4: Regression between accuracy and cross-entropy for various sampling strategies on Turkish inflection.

We compute the cross-entropy between the test set edit distribution and each of the sampled sets. We find that across languages, increased cross-entropy, which indicates that the sampled set is more dissimilar from the test set, tends to correlate with decreased performance. For example, Figure 4 plots the performance and cross entropy for the various sampling strategies for Turkish.

This is an intuitive result, confirming the importance of sampling a training set that is similar in distribution to the target test set. We run linear regression for each language and report the slopes and R^2 values.

It is clear that in most cases, reducing cross-entropy by choosing a sampling strategy that approximates the test distribution is beneficial to performance. However, since the test distribution is not necessarily known in real-world active learning scenarios, this remains a difficult task to solve.

[11] Himmelmann (1998) distinguishes these categories further, with a third 'Staged communicative events'. This refers to tasks that are prompted for linguistic purposes, such as a picture task.

Language	Slope	R^2
arp	-0.25	0.381
fin	-0.03	0.148
kat	-0.05	0.731**
que	-0.49	0.512*
sah	-0.06	0.716**
tur	-0.01	0.842**
tyv	-0.05	0.321
zul	-0.08	0.847**

Table 2: Linear regressions for each language between cross-entropy of sampled sets with test sets (x) and accuracy on the test set (y). * indicates significance with $n = 8$ and $p < 0.05$, ** indicates significance with $p < 0.01$.

5 Conclusion

Computational methods can aid language documentation and description projects by processing and analyzing recorded data. Active learning approaches can greatly aid in the rapid development of robust automated systems by focusing annotation on highly beneficial samples, but existing research on simulated AL often makes unrealistic assumptions. We compare a standard approach (skyline), where data is sampled from unrealistic linguistic resources, an approach based on naturalistic word frequencies (baseline), and a number of strategies motivated by encouraging lexical diversity. Our skyline and baseline approaches serve as analogs to elicitation and naturalistic recording.

We find that the skyline approach is difficult to beat, but as few languages have sufficient corpora with complete, diverse paradigms, we argue this approach is an unrealistic baseline for AL. Meanwhile, we find clear benefits from targeted sampling strategies, with inflection model confidence (TMC) and character LM scores (LMC(WORDFORM, MSD)) yielding the greatest improvements.

6 Limitations

Three of our eight languages are members of the Turkic language family. Despite our best efforts, it was not possible to have a set of languages that covered a significant range of typological features, particularly pertaining to phonology and morphology. In most cases, either the existing Wikipedia was too small or there were issues with orthography that did not map neatly with the UniMorph database. This is a limitation of the study presented

and remains an intended future rectification for the authors.

It is important to note that the style and register of Wikipedia is limited. As such, certain MSDs are underrepresented or over-represented, compared with natural speech. Our experiments use Wikipedia articles to simulate texts/recordings of language, a limited approximation of the natural setting that does not cover a broad range of genres. However, constructing a representative corpus in the language documentation context is an almost impossible endeavour.

7 Ethics Statement

If our results do not hold across a wide variety of languages, our suggested AL approaches may result in annotator effort that is not beneficial to the model. This would be a significant opportunity cost, particularly in the case of languages which are considered critically endangered.

Automated systems for inflection and language documentation are limited in scope and carry some degree of error. While they can greatly aid in documentation projects, they should not be used to entirely replace human annotators and linguists in the documentation, study, and preservation of languages. Particularly for Indigenous and endangered languages, care should be taken to use data and automated systems in a way consistent with the desires of the language community (Schwartz, 2022).

Finally, training models carries an unavoidable environmental cost (Bender et al., 2021). While our research uses small models, we strive to ensure the benefits outweigh these costs.

References

Sophia Althammer, Guido Zuccon, Sebastian Hofstätter, Suzan Verberne, and Allan Hanbury. 2023. Annotating data for fine-tuning a neural ranker? current active learning strategies are not better than random selection. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 139–149.

Vamshi Ambati. 2012. *Active learning and crowdsourcing for machine translation in low resource scenarios*. Ph.D. thesis, Carnegie Mellon University.

Peter Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages*. Cambridge University Press.

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.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Joseph Lo Bianco. 2002. Real world language politics and policy. *Language policy: Lessons from global models. Monterey, California: Monterey Institute of International Studies*.

Klaus Brinker. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 59–66.

Shobhana L. Chelliah. 2001. *The role of text collection and elicitation in linguistic fieldwork*, page 152–165. Cambridge University Press.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. 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.

David Crystal. 2002. *Language Death*. Canto. Cambridge University Press.

Bruce L Derwing and Royal Skousen. 1994. Productivity and the english past tense. *The reality of linguistic rules, Amsterdam/Philadelphia, John Benjamins Publishing Company*, pages 193–218.

Nicholas Evans. 2008. Review of gippert, jost, nikolaus himmelmann and ulrike mosel (eds.), essentials of language documentation. *Language Documentation & Conservation*, 2:340–350.

Mohamed Farouk Abdel Hady and Friedhelm Schwenker. 2010. Combining committee-based semi-supervised learning and active learning. *Journal of Computer Science and Technology*, 25(4):681–698.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning bert for low-resource natural language understanding via active learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1158–1171.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–196.

Richard Hudson. 1994. About 37% of word-tokens are nouns. *Language*, 70(2):331–339.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

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.

Junying Liang and Haitao Liu. 2013. Noun distribution in natural languages. *Poznań Studies in Contemporary Linguistics*, 49(4):509–529.

Ray Liere and Prasad Tadepalli. 1997. Active learning with committees for text categorization. In *AAAI/IAAI*, pages 591–596. Citeseer.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 153–161, Online. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada. Association for Computational Linguistics.

Andrew McCallum, Kamal Nigam, et al. 1998. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Citeseer.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Sandeep Prasada and Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and cognitive processes*, 8(1):1–56.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. 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.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Royal Skousen. 1990. *Analogical Modeling of Language*. Springer Netherlands.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Stephen A Wurm. 2001. *Atlas of the World's Languages in Danger of Disappearing*. Unesco.

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.

A Model details

Across preliminary experiments and the runs listed in this paper, training took around 1,500 compute hours. We ran experiments on the UBC computing cluster and Google Colab. Models were trained with the hyperparameters listed in Table 3. Models had around 10M parameters.

Hyperparameter	Value
Encoder/Decoder layers	4
Encoder/Decoder attention heads	4
Optimization	Adam
Embedding size	256
Hidden layer size	1024
Learning rate	0.001
Batch Size	400
Label Smoothing	0.1
Gradient clip threshold	1.0
Warmup updates	1000
Max updates	6000

Table 3: Our hyperparameters follow the setup described by Liu and Hulden (2020).

B Data Composition

Information about the composition for each language is given in Table 4.

C Language-Specific Model Accuracies

Accuracy scores are reported in Table 5

Language	POS present	N=adj	Wikipedia sample	Four lemma tables size	COPY size	Total training set size	Test size
tyv	N	?	1000	336	10	1346	1008
ara	A,N	N	1000	320	10	1330	1147
kat	N	Y	1000	64	65	1129	1040
que	N	Y	1000	768	5	1773	1152
zul	A,N	N	1000	236	20	1256	992
sah	N	?	1000	350	10	1360	1092
tur	A,N	N	1000	216	30	1246	1068
fin	N	Y	1000	104	40	1140	1092

Table 4: Seed training and test set composition for each language. The wikipedia sample refers to the frequency weighted sample taken from Wikipedia. The four lemma table size describes the added full paradigms from the Unimorph database. Copy size denotes the number of unique lemma found in the test size. The test size varies for each language as the paradigm sizes differ (and thus the number of lemma).

Language	Seed	± std	Skyline	± std	Wiki Uniform	± std	Baseline	± std	Oracle	± std	TMC	± std
tyv	0.416	0.028	0.909	0.013	0.686	0.020	0.561	0.034	0.733	0.036	0.711	0.033
ara	0.256	0.007	0.616	0.031	0.336	0.010	0.318	0.019	0.556	0.023	0.448	0.023
kat	0.870	0.023	0.931	0.004	0.926	0.018	0.896	0.020	0.949	0.014	0.922	0.029
que	0.514	0.032	0.820	0.021	0.509	0.023	0.604	0.031	0.811	0.027	0.786	0.012
zul	0.391	0.044	0.791	0.025	0.400	0.016	0.424	0.019	0.576	0.025	0.763	0.026
sah	0.664	0.062	0.972	0.011	0.863	0.021	0.728	0.026	0.912	0.021	0.972	0.004
tur	0.100	0.021	0.227	0.026	0.177	0.026	0.142	0.015	0.167	0.018	0.251	0.030
fin	0.402	0.028	0.727	0.047	0.690	0.020	0.473	0.032	0.453	0.050	0.732	0.024

Language	Seed	± std	LMC (WF,MSD)	± std	LMC (Lem,MSD)	± std	LMC(WF)	± std	LMC(Lem)	± std
tyv	0.416	0.028	0.751	0.027	0.580	0.031	0.695	0.045	0.571	0.014
ara	0.256	0.007	0.600	0.024	0.520	0.042	0.498	0.040	0.372	0.009
kat	0.870	0.023	0.936	0.011	0.918	0.009	0.931	0.010	0.908	0.028
que	0.514	0.032	0.749	0.042	0.707	0.038	0.654	0.044	0.688	0.055
zul	0.391	0.044	0.695	0.029	0.547	0.031	0.625	0.019	0.571	0.021
sah	0.664	0.062	0.861	0.035	0.737	0.019	0.865	0.023	0.787	0.017
tur	0.100	0.021	0.197	0.032	0.171	0.016	0.197	0.037	0.154	0.025
fin	0.402	0.028	0.664	0.038	0.421	0.040	0.683	0.037	0.472	0.042

Table 5: Model accuracies for all sampling strategies considered. The reported standard deviation is calculated across five equal partitions on the test set. **TMC** = "Transformer Model Confidence", **LMC** = "Language model confidence", **WF** = "Wordform", and **Lem** = "Lemma".