Towards Unsupervised Morphological Analysis of Polysynthetic Languages

Sujay Khandagale[1] **Yoann Léveillé**[2] **Samuel Miller**[3],
Derek Pham[1] **Ramy Eskander**[1] **Cass Lowry**[4] **Richard Compton**[3],
Judith Klavans[3] **Maria Polinsky**[3] **Smaranda Muresan**[1]
[1]Columbia University, {sk4746,dp3081,rnd2110,smara}@columbia.edu
[2]Université du Québec à Montréal, {leveille.yoann,compton.richard}@uqam.ca
[3]University of Maryland, {samm,jklavans,polinsky}@umd.edu
[4]The Graduate Center, City University of New York, clowry@gradcenter.cuny.edu

Abstract

Polysynthetic languages present a challenge for morphological analysis due to the complexity of their words and the lack of high-quality annotated datasets needed to build and/or evaluate computational models. The contribution of this work is twofold. First, using linguists' help, we generate and contribute high-quality annotated data for two low-resource polysynthetic languages for two tasks: morphological segmentation and part-of-speech (POS) tagging. Second, we present the results of state-of-the-art unsupervised approaches for these two tasks on Adyghe and Inuktitut. Our findings show that for these polysynthetic languages, using linguistic priors helps the task of morphological segmentation and that using stems rather than words as the core unit of abstraction leads to superior performance on POS tagging.

1 Introduction

Polysynthetic languages are highly synthetic languages, where a single multi-morpheme verbal complex can express what would be a whole sentence in English. For example, in Inuktitut, "tusaatsiarunnanngittualuujunga" corresponds to the English sentence "I cannot hear very well" (Klavans, 2018). These languages pose two main challenges for computational models. First, they are often characterized by a significant number of morphemes per word and a high degree of ambiguity of their roots with respect to the part-of-speech specification (Baker, 1996). Second, these languages are low-resource, lacking large scale annotated datasets needed to build computational models.

We focus on surface-level morphological segmentation and part-of-speech tagging for two polysynthetic languages: Adyghe and Inuktitut. Progress in morphological analysis of polysynthetic languages has been made possible by two efforts: morphological segmentation frameworks that move away from rule-based methods to unsupervised machine learning models, which crucially are able to include linguistic priors to guide the learning process (Sirts and Goldwater, 2013; Mager et al., 2018; Eskander et al., 2021; Le and Sadat, 2021), and the growth in corpora for some of these languages (Farley, 2009; Sorokin, 2020; Micher, 2019; Arkhangelskiy and Medvedeva, 2016; Arkhangelskiy and Lander, 2015). A particularly fruitful line of work has been the use of *unsupervised models based on Adaptor Grammars* (Johnson et al., 2007), such as *MorphAGram* (Eskander et al., 2020a) that enables the use of *linguistic priors, either through grammar definition or linguist-provided affixes* (Eskander et al., 2021; Le and Sadat, 2021). We investigate whether linguistic priors in *MorphAGram* help the task of morphological segmentation for Adyghe and Inuktitut.

POS tagging for polysynthetic languages, on the other hand, is in its infancy. We investigate whether unsupervised approaches based on cross-lingual projection developed for low-data scenarios (Yarowsky et al., 2001; Agić et al., 2015; Das and Petrov, 2011; Buys and Botha, 2016; Täckström et al., 2013; Eskander et al., 2020b) could be useful for POS tagging of polysynthetic languages. These methods rely on the use of parallel data (e.g., the Bible) to project POS tags from a source language for which a POS tagger is accessible onto a target language across word-level alignments. The projected tags then become the basis to train a POS model for the target language. Eskander et al. (2022) have recently proposed an approach for cross-lingual projection in low-data scenarios, where the unit of abstraction could be either the word or the stem, thus exploring either word-level or stem-level alignments for projection [1]. We show that for Adyghe and Inuktitut, using stems as the unit of abstraction improves the results for POS tagging. We contribute Adyghe and Inuktitut evalu-

[1]See Eskander (2021) for broad experimentation in several monolingual and multilingual settings.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 334–340
November 20–23, 2022.

ation datasets both for morphological segmentation and POS tagging.

2 Languages and Data Annotation

Adyghe, also known as West Circassian, is a member of the Northwest Caucasian language family with about 118K speakers. Adyghe is characterized by a complex encoding of clausal arguments in the verb form; person markers appear in the preverbal position, and in addition to subject and object markers include markers of additional arguments introduced by applicative morphemes in the verbal paradigm. Many researchers have noted the difficulty of distinguishing between inflection and derivation in the verbal morphology (Kimmelman, 2010; Arkadiev and Maisak, 2018). **Eastern Canadian Inuktitut** (Inuit-Yupik-Unangan) is spoken in the Canadian Arctic by about 40K speakers. The degree of polysynthesis in terms of the number of morphemes per word is high. The language possesses closed classes of verbs that obligatorily trigger either noun incorporation or verb incorporation. The language makes extensive use of category-changing morphology (Johns, 2014), including what Mattissen (2017) calls "ping-pong recategorization", whereby the category of a word switches back and forth due to the presence of multiple verbalizers and nominalizers. Another challenge for morphological segmentation is that the morphemes are relatively short and the phoneme inventory is small, leading to a fair amount of homophony and a high number of potential parses.

2.1 Morphological Segmentation

To create the evaluation datasets we had to decide the relevant level of granularity for morphological analysis and to include all plausible segmentations.

Adyghe. To build our training and evaluation datasets, we rely on an electronically annotated corpus, which allows searching based on specific morphological information (Arkhangelskiy and Medvedeva, 2016; Arkhangelskiy and Lander, 2015). To build the training dataset for *MorphA-Gram*, we select 50K unsegmented words by randomly sampling according to the logarithmic distribution of words' POS tags, with weighting for word frequency in the corpus. The gold-standard dataset contains 1000 words together with their morphological segmentation from the original corpus, which was automatically obtained. The segmentations are manually verified and corrected by a trained linguist with knowledge of Adyghe to ensure accuracy. Among the 1000 words, there are 208 verbs, 177 nouns, 167 adjectives, and 146 adverbs.

Inuktitut. For training the segmentation models, we collect the 50K most frequent words (unsegmented) from the Inuktitut Wikipedia, the Nunavut Hansard (NH) corpus, and the Bible. The primary data for the gold standard is collected from the UQAILAUT Project (Farley, 2009) and consists of 1094 words and their associated segmentations. Most words contain only one possible segmentation in this original dataset. Two trained linguists working on Inuktitut reviewed and corrected the dataset, including: regularizing inconsistencies in how inflectional morphology is segmented, regularizing lexicalized stem inconsistencies and segmenting spurious dual and plural morphemes, excluding sequences of words that were accidentally fused due to a missing space in the source data, and providing alternative segmentations, when appropriate. This corpus contains mostly nouns (85.4%). As verbs generally exhibit a higher degree of polysynthesis in Inuktitut, we collect an additional set of 100 words from the Nunavut Hansard corpus that consists of nouns (22), verbs (66), and participles (12), and that is manually segmented by two trained linguists. Our gold Inuktitut dataset contains words that have alternative segmentations (Table 1).

2.2 POS tagging

Adyghe. For training the POS tagger, we extract the available parallel Bible data (Russian-Adyghe) from the corpus introduced by Arkhangelskiy and Medvedeva (2016); Arkhangelskiy and Lander (2015). For the gold-standard dataset, a simple random sample of 200 sentences with well-formed data is extracted from the entire corpus and verified by a linguist, all after mapping the POS tags to the UD POS schema. The final distribution of POS is: VERB (31.9%), NOUN (27.8%), PUNCT (23.8%), ADJ (6.4%), PRON (5.1%), ADV (3.6%), NUM (0.7%), CCONJ (0.5%) and ADP (0.1%).

Inuktitut. For training the POS tagger, we collect the English-Inuktitut Bible data. For evaluation, we annotate a small dataset containing 124 sentences: 50 are extracted from the Nunavut Hansard and 74 are taken from three articles in Inuktitut Magazine. Word forms are manually tagged by a master student specializing in Inuktitut morphosyntax following the UD POS tagging conventions. The distribution of the tags in the

Word	Full Segmentation	Partial Segmentation
kiinaujalirijikkunnut *'of finance'*	kiina-u-ja-liri-ji-kkun-nut *face*-BE-PSV.PART-*work.on*-AG.NZ-ASSOC-PL.ALL	kiinauja-liri-ji-kkun-nut *money-work.on*-AG.NZ-ASSOC PL.ALL
kiinaujait *funds*	kiina-u-ja-it *face*-BE-PSV.PART-PL	kiinauja-it *money*-PL
titiraqtautsiarunnaqullugit *'so that they can be spelled correctly'*	titi-raq-tau-tsia-runna-qu-llu-git *mark*-REP-PASS-*well-can-so.that*-CTMP-3SG	titiraq-tau-tsia-runna-qu-llu-git *write*-PASS-*well-can-so.that*-CTMP-3SG

Table 1: Examples of *full* and *partial* segmentations from the Inuktitut gold dataset, where AG.NZ = agent nominalizer; ALL = allative case; ASSOC = associative; CTMP = contemporative mood; PASS = PASSIVE; PS.PART = passive participial; PL = plural; REP = repetitive; SG = singular

gold dataset is: NOUN (46.0%), VERB (27.3%), PUNCT (18.6%), CCONJ (4.6%), PROPN (3.1%), PRON (0.4%) and ADV (0.1%).

3 Approach

3.1 Morphological Segmentation

To conduct the experiments for morphological segmentation, we use *MorphAGram* [2] (Eskander et al., 2020a), a state-of-the-art, publicly available framework for unsupervised morphological segmentation that is based on Adaptor Grammars (AGs) (Johnson et al., 2007). AGs are nonparametric Bayesian models that utilize probabilistic context-free grammars (PCFGs). An AG is composed of two main components: a PCFG and an adaptor that adapts the probabilities of individual subtrees and acts as a caching mechanism. In the case of morphological segmentation, a PCFG represents a morphological grammar that specifies word formation, where the purpose is to learn latent tree structures of morphological segments given a list of unsegmented words.

While *MorphAGram* was originally developed for learning in a fully unsupervised manner, it also allows the use of linguistic priors to enhance morphological segmentation in a minimally supervised fashion. Eskander et al. (2021) introduce two methods for including linguistic priors: *grammar definition* and *linguist-provided affixes*. In the former, a linguist tailors the language independent grammars used by *MorphAGram* to more accurately model the word structure of the target language. In the latter, an expert in the target language compiles a list of affixes and seeds it into the grammars using the *Scholar-Seeded* learning setting (Eskander et al., 2016). For all of our experiments and languages in this paper, we apply the second approach where linguist-provided affixes are used.

We follow Eskander et al. (2021) by applying their on-average best performing grammar, namely *PrStSu+SM* [3], in which a word is modelled as a sequence of prefixes, a stem and a sequence of suffixes, Additionally, both prefixes and suffixes are recursively defined to allow for affix compounding, and the morphemes are further split into non-linguistically driven sub-morphemes that allow for better utilization of the generated latent subtrees (See Eskander et al. (2021) for more details). For Inuktitut, we use the affixes from Inuktitut Tusaalanga Grammar[4].

3.2 Part-of-Speech Tagging

To conduct the experiments for POS tagging, we use a publicly available fully unsupervised cross-lingual POS tagger that projects the annotations across some parallel text between a source language and the target one [5] (Eskander et al., 2020b, 2022). First, we utilize the Bible as the source of parallel data to train bidirectional alignment models between the source and target languages using GIZA++ (Och and Ney, 2003). We then tag the source side for POS using an off-the-shelf tagger. In our study, we use English as the source language and utilize Stanza (Qi et al., 2020) to tag the English text for the Universal-Dependencies

[2] https://github.com/rnd2110/MorphAGram

[3] https://github.com/rnd2110/MorphAGram/blob/master/data/georgian/grammar/standard/grammar1.txt

[4] https://tusaalanga.ca/grammar

[5] https://github.com/rnd2110/unsupervised-cross-lingual-POS-tagging

	AG-LI			AG-SS		
Language	All	Noun	Verb	All	Noun	Verb
Adyghe	66.1	69.3	56.5	78.9	70.8	69.4
Inuktitut	58.1	64.4	50.3	60.4	67.6	49.2

Table 2: Morphological segmentation results (BPR F1) on the entire test sets (All), Nouns and Verbs. AG-LI is the *MorphAGram* Standard language-independent model, while AG-SS is the model using linguistic priors.

	Adyghe			Inuktitut		
Alignment Type	All	Noun	Verb	All	Noun	Verb
Word-Based	62.4	49.2	67.1	57.3	62.3	39.6
Stem-Based	70.4	66.4	73.2	64.6	68.8	44.5

Table 3: POS tagging results for word-based and stem-based alignment and projection.

(UD) POS tagset [6]. The English tags are then projected onto the target side across the intersecting bidirectional alignments, while a target word that is not part of an alignment or part of an alignment in one direction but not the other receives a $NULL$ POS assignment. This is followed by a refinement phase in which we couple both token and type constraints and only consider highly scoring sentences, where sentence score is defined as the harmonic mean of its projection density and alignment confidence. Finally, we learn a neural Bi-LSTM model (Hochreiter and Schmidhuber, 1997) given the induced annotations. The model exploits both word embeddings and affix embeddings that represent n-gram prefixes and suffixes, where $n \in \{1, 2, 3, 4\}$. Additionally, we utilize hierarchical Brown-cluster (Brown et al., 1992) embeddings that we learn by applying the Percy Liang's implementation of Brown clustering [7] on the Bible data of the target languages (See Eskander et al. (2020b) for more details).

We conduct the experiments using two different approaches for alignment and projection as introduced recently by Eskander et al. (2022): (1) word-based; and (2) stem-based. In the word-based approach, we utilize the parallel text to train models that align the source and target sides at the word level. After generating the POS annotations for the source language, these annotations are then projected onto the target across the word-level alignments. In the stem-based approach, we perform both alignment and projection in the stem space. In this setup, we first conduct stemming for the source and target texts using *MorphAGram* and learn stem-based alignment models between the two sides. We then apply the source annotations to the underlying stems and project them onto the stemmed target across the stem-level alignments. Finally, we replace each tagged target stem by its corresponding word so that we can train the neural POS tagger at the word level. We experiment with both approaches for Adyghe and Inuktitut.

It is worth noting that MorphAGram performs surface-level morphological segmentation in which the stem is automatically specified without supervision, where starting and ending frequent morphemes are highly likely to receive an affix assignment.

4 Results and Error Analysis

4.1 Morphological Segmentation.

The performance of *MorphAGram* segmentation models is shown in Table 2. Adding scholarly seeded affixes improves the BPR F1-score (Virpioja et al., 2011) by 19.4% for Adyghe and 4.0% for Inuktitut. Table 2 also shows the segmentation performance for noun and verbs. While for Adyghe the linguistic priors help substantially for verbs, for Inuktitut we do not see this effect, indicating that more care needs to be given to the linguist-provided affixes related to verbal constructions and/or exploring linguistic priors as grammar definition (Le and Sadat, 2021).

[6] https://universaldependencies.org/u/pos/

[7] https://github.com/percyliang/brown-cluster

Language	Model	Example Sentence
Adyghe	Gold	Ау_CCONJ джырэ_ADV нэс_ADP а_PRON къулыкъум_NOUN Іоф_NOUN зыщиш-Iэн_VERB унэ_NOUN тэрэз_ADJ иIагъэп_VERB ._PUNCT
	Word-Based	Ау_CCONJ джырэ_NOUN нэс_PROPN а_VERB къулыкъум_VERB Іоф_NOUN зыщиш-Iэн_VERB унэ_NOUN тэрэз_VERB иIагъэп_VERB ._PUNCT
	Stem-Based	Ау_CCONJ джырэ_ADV нэс_ADP а_CCONJ къулыкъум_NOUN Іоф_NOUN зыщиш-Iэн_VERB унэ_NOUN тэрэз_VERB иIагъэп_VERB ._PUNCT
Inuktitut	Gold	taima_NOUN ,_PUNCT qaujigumavunga_VERB itsivautaaq_NOUN ,_PUNCT minista_NOUN uqarunnarmangaaq_VERB qanuq_NOUN pilirivingit_NOUN piliriaqaqattarmangaata_VERB taimaittunik_NOUN qimaavit_NOUN matutuinnariaqaliraimmata_VERB nunalinni_NOUN kiinaujaqtuutairutuaramik_VERB ._PUNCT
	Word-Based	taima_ADV ,_PUNCT qaujigumavunga_NOUN itsivautaaq_VERB ,_PUNCT minista_NOUN uqarunnarmangaaq_VERB qanuq_NOUN pilirivingit_NOUN piliriaqaqattarmangaata_NOUN taimaittunik_NOUN qimaavit_NOUN matutuinnariaqaliraimmata_NOUN nunalinni_NOUN kiinaujaqtuutairutuaramik_VERB ._PUNCT
	Stem-Based	taima_ADV ,_PUNCT qaujigumavunga_VERB itsivautaaq_NOUN ,_PUNCT minista_NOUN uqarunnarmangaaq_VERB qanuq_PRON pilirivingit_NOUN piliriaqaqattarmangaata_NOUN taimaittunik_NOUN qimaavit_NOUN matutuinnariaqaliraimmata_NOUN nunalinni_NOUN kiinaujaqtuutairutuaramik_NOUN ._PUNCT

Table 4: POS tagging comparison between ground truth, word-based, and stem-based models for Adyghe and Inuktitut. Green indicate correctly identified POS tags, while red indicates incorrect POS tags.

4.2 Part-of-Speech Tagging.

Table 3 shows our results for POS tagging using the word-level and stem-level alignment and projection for Adyghe and Inuktitut on all POS tags (All) as well as the performance on Nouns and Verbs. The stem-based approach outperforms the word-based one, which lends support that using the stem as the unit of abstraction for the POS tagging of polysynthetic languages is a fruitful avenue of research. In terms of accuracy, stem-based POS tagging outperforms word-based POS tagging by 8.0% for Adyghe, and 7.3% for Inuktitut across all POS tags. Moreover, we see substantial improvements on both nouns and verbs when using stem-based over word-based POS tagging (F1 metric). As an example for Adyghe, the stem-based model correctly tags the word къулыкъум as a noun, while the word-based model misclassifies it as a verb (Table 4). This shows that even though in Adyghe both verbs and nouns can end in –ъум, the stem-based model is able to determine that the word is a noun. For Inuktitut, the stem-based model correctly classifies the word *qaujigumavunga* as a verb, while the word-based model incorrectly labels it as a noun.

5 Conclusion

We contribute high-quality datasets for Inuktitut and Adyghe, both for morphological segmentation and POS tagging. We show that unsupervised approaches that consider linguistic priors are a promising avenue for tackling morphological segmentaters for polysynthetic languages. We also show that unsupervised cross-lingual projection approaches for POS tagging that use the stem as a unit of abstraction are a fruitful avenue of research on POS tagging for polysynthetic languages.

Acknowledgements

This research is based upon work supported by the National Science Foundation (awards #1941742 and #1941733). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, expressed or implied, of NSF or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Ethical Considerations

The annotations were done by linguists with appropriate compensation after educating them about the research purpose and the annotation process. The quality of the annotations was examined manually and empirically. The source code and the data will be released open-source. Finally, the limitations of the work lay within the reported performance. There should be no potential risks given these stated limitations.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272.

Peter Arkadiev and Timur Maisak. 2018. Grammaticalization in the north caucasian languages. page 116–145. Oxford University Press.

Timofey Arkhangelskiy and Yury Lander. 2015. Some challenges of the west circassian polysynthetic corpus. *SSRN Electronic Journal*.

Timofey Arkhangelskiy and Maria Medvedeva. 2016. Developing morphologically annotated corpora for minority languages of russia. In *CLiF*, pages 1–6.

Mark Baker. 1996. *The polysynthesis parameter*. Oxford University Press.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

Ramy Eskander. 2021. *Unsupervised Morphological Segmentation and Part-of-Speech Tagging for Low-Resource Scenarios*. Columbia University.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith L Klavans, and Smaranda Muresan. 2020a. Morphagram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7112–7122.

Ramy Eskander, Cass Lowry, Sujay Khandagale, Francesca Callejas, Judith Klavans, Maria Polinsky, and Smaranda Muresan. 2021. Minimally-supervised morphological segmentation using Adaptor Grammars with linguistic priors. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3969–3974, Online. Association for Computational Linguistics.

Ramy Eskander, Cass Lowry, Sujay Khandagale, Judith Klavans, Maria Polinsky, and Smaranda Muresan. 2022. Unsupervised stem-based cross-lingual part-of-speech tagging for morphologically rich low-resource languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4061–4072, Seattle, United States. Association for Computational Linguistics.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020b. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 900–910.

Benoit Farley. 2009. The Uqailaut Project. Accessed on 10 Jan 2019.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Alana Johns. 2014. Eskimo-aleut. In *The Oxford Handbook of Derivational Morphology*. Oxford University Press.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor Grammars: a Framework for Specifying Compositional Nonparametric Bayesian Models. In *Advances in Neural Information Processing Systems 19*, pages 641–648, Cambridge, MA. MIT Press.

Vadim Kimmelman. 2010. Auxiliaries in adyghe. In *Paper presented at the Workshop on Grammaticalization*, University of Amsterdam.

Judith L. Klavans. 2018. Proceedings of the workshop on computational modeling of polysynthetic languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Association for Computational Linguistics.

Ngoc Tan Le and Fatiha Sadat. 2021. Towards a first automatic unsupervised morphological segmentation for Inuinnaqtun. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 159–162, Online. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic Finite-State Morphological Segmenter for Wixarika (Huichol) Language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Johanna Mattissen. 2017. Sub-types of polysynthesis. In *The Oxford Handbook of Polysynthesis*. Oxford University Press.

Jeffrey Micher. 2019. Bootstrappin a neural morphological generator from morphological analyzer output for inuktitut. In *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages*. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1:255–266.

Alexey Sorokin. 2020. Getting more data for low-resource morphological inflection: Language models and data augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3978–3983, Marseille, France. European Language Resources Association.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.