Unsupervised Morphological Segmentation for Low-Resource Polysynthetic Languages

Ramy Eskander
Columbia University
Dept. of Computer Science
rnd2110@columbia.edu

Judith L. Klavans
University of Maryland
UMIACS
jklavans@umd.edu

Smaranda Muresan
Columbia University
Data Science Institute
smara@columbia.edu

Abstract

Polysynthetic languages pose a challenge for morphological analysis due to the root-morpheme complexity and to the word class "squish". In addition, many of these polysynthetic languages are low-resource. We propose *unsupervised* approaches for morphological segmentation of low-resource polysynthetic languages based on Adaptor Grammars (AG) (Eskander et al., 2016). We experiment with four languages from the Uto-Aztecan family. Our AG-based approaches outperform other unsupervised approaches and show promise when compared to supervised methods, outperforming them on two of the four languages.

1 Introduction

Computational morphology of polysynthetic languages is an emerging field of research. Polysynthetic languages pose unique challenges for computational approaches, including machine translation and morphological analysis, due to the root-morpheme complexity and to word class gradations (Homola, 2011; Mager et al., 2018d; Klavans, 2018a). Previous approaches include rule-based methods based on finite state transducers (Farley, 2009; Littell, 2018; Kazeminejad et al., 2017), hybrid models (Mager et al., 2018b; Moeller et al., 2018), and supervised machine learning, particularly deep learning approaches (Micher, 2017; Kann et al., 2018). While each rule-based method is developed for a specific language (Inuktitut (Farley, 2009), or Arapaho (Littell, 2018; Moeller et al., 2018)), machine learning, including deep learning approaches, might be more rapidly scalable to many additional languages.

We propose an *unsupervised* approach for morphological segmentation of polysynthetic languages based on Adaptor Grammars (Johnson et al., 2007). We experiment with four Uto-Aztecan languages: Mexicanero (MX), Nahuatl (NH), Wixarika (WX) and Yorem Nokki (YN) (Kann et al., 2018). Adaptor Grammars (AGs) are nonparametric Bayesian models that generalize probabilistic context free grammars (PCFG), and have proven to be successful for unsupervised morphological segmentation, where a PCFG is a morphological grammar that specifies word structure (Johnson, 2008; Sirts and Goldwater, 2013; Eskander et al., 2016, 2018). Our main goal is to examine the success of Adaptor Grammars for unsupervised morphological segmentation when applied to polysynthetic languages, where the morphology is synthetically complex (not simply agglutinative), and where resources are minimal. We use the datasets introduced by Kann et al. (2018) in an unsupervised fashion (unsegmented words). We design several AG learning setups: 1) use the best-on-average AG setup from Eskander et al. (2016); 2) optimize for language using just the small training vocabulary (unsegmented) and dev vocabulary (segmented) from Kann et al. (2018); 3) approximate the effect of having some linguistic knowledge; 4) learn from all languages at once and 5) add additional unsupervised data for NH and WX (Section 3). We show that the AG-based approaches outperform other unsupervised methods — $Morfessor$ (Creutz and Lagus, 2007) and $MorphoChain$ (Narasimhan et al., 2015)) —, and that for two of the languages (NH and YN), the best AG-based approaches outperform the best supervised methods (Section 4).

2 Languages and Datasets

Typically, polysynthetic languages demonstrate holophrasis, i.e. the ability of an entire sentence to be expressed as what is considered by native speakers to be just one word. To illustrate, consider the following example from Inuktitut (Kla-

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 189–195
Florence, Italy. August 2, 2019

vans, 2018b), where the morpheme -tusaa- is the root and all the other morphemes are synthetically combined with it in one unit:

tusaa-tsia-runna-nngit-tu-alu-u-jung
hear-well-be.able-NEG-DOE-very-BE-PT.1S
I can't hear very well.

Another example from WX, one of the languages in the dataset for this paper (from (Mager et al., 2018c)) shows this complexity:

yu-huta-me ne-p+-we-iwa
an-two-ns 1sg:s-asi-2pl:o-brother
I have two brothers.

In linguistic typology, the broader gradient is: isolating/analytic to synthetic to polysynthetic. Agglutinating refers to the clarity of boundaries between morphemes. This more specific gradation is: agglutinating to mildly fusional to fusional. Thus a language might be characterized overall as polysynthetic and agglutinating, i.e. generally a high number of morphemes per word, with clear boundaries between morphemes and thus easily segmentable. Another language might be characterized as polysynthetic and fusional, so again, many morphemes per word, but many phonological and other processes so it is difficult to segment morphemes.

Thus, morphological analysis of polysynthetic languages is challenging due to the root-morpheme complexity and to word class gradations. Linguists recognize a gradience in word classes, known as "squishiness", a term first discussed in Ross (1972) who argued that, instead of a fixed, distinct inventory of syntactic categories, a quasi-continuum from verb, adjective and noun best reflects most lexical distinctions. The root-morpheme complexity and the word class "squish" makes developing segmented training data with reliability across annotators difficult to achieve. Kann et al. (2018) have made a first step by releasing a small set of morphologically segmented datasets although even in these carefully curated datasets, the distinction between affix and clitic is not always indicated. We use these datasets in an unsupervised fashion (i.e., we use the unsegmented words). These datasets were taken from detailed descriptions in the Archive of Indigenous Languages collection for MX (Canger, 2001), NH (de Suárez, 1980), WX (Gómez and López, 1999), and YN (Freeze, 1989). They were constructed so they include both segmentable as well as non-segmentable words to ensure that methods can correctly decide against splitting up single morphemes. However, as noted above, there is a gradation of polysynthesis, so the delineation of language types is not clear-cut. For these four languages, the more agglutinative is WX; Leza (2004) has observed 20 morphemes per word for this language.

	Mexicanero	Nahuatl	Wixarika	Yorem N.
train	427	540	665	511
train_{Bible}	-	14.7K	16.6K	-
dev	106	134	176	127
test	355	449	553	425

Table 1: Number of words in train, dev, test splits from Kann et al. (2018) + additional Bible data

Each training, development and test example consists of one word. Table 1 contains the count of words in the training, development and test. Unlike Kann et al. (2018), for training we do not use the segmented version of the data (our approach is unsupervised). In addition to the datasets, for NH and WX we also have available the Bible (Christodouloupoulos and Steedman, 2015; Mager et al., 2018a), which we consider for one of our experimental setups as additional training data. In the dataset from (Kann et al., 2018), the maximum number of morphemes per word for MX is seven with an average of 2.13; for NH, six with an average of 2.2; for WX, maximum of ten with an average of 3.3; and for YN, the maximum is ten, with an average of 2.13.

3 Using Adaptor Grammars for Polysynthetic Languages

An Adaptor Grammar is typically composed of a PCFG and an adaptor that adapts the probabilities of individual subtrees. For morphological segmentation, a PCFG is a morphological grammar that specifies word structure, where AGs learn latent tree structures given a list of words. In this paper, we experiment with the grammars and the learning setups proposed by Eskander et al. (2016), which we outline briefly below.

Grammars. We use the nine grammars from Eskander et al. (2016, 2018) that were designed based on three dimensions: 1) how the grammar models word structure (e.g., prefix-stem-suffix vs. morphemes), 2) the level of abstraction in non-terminals (e.g., compounds, morphemes and sub-morphemes) and 3) how the output boundaries are specified (see Table 2 for a sample grammars). For example, the PrStSu+SM grammar models the

Grammar	Main Representation	Compound	Morph	SubMorph	Segmentation Level
Morph+SM	Morph+	No	Yes	Yes	Morph
PrStSu+SM	Prefix+Stem+Suffix	No	Yes	Yes	Prefix-Stem-Suffix
PrStSu+Co+SM	Prefix+Stem+Suffix	Yes	Yes	Yes	Prefix-Stem-Suffix

Table 2: Sample grammar setups used by Eskander et al. (2018, 2016). Compound = Upper level representation of the word as a sequence of compounds; Morph = affix/morpheme representation as a sequence of morphemes. SubMorph (SM) = Lower level representation of characters as a sequence of sub-morphemes. "+" denotes *one or more*.

word as a complex prefix, a stem and a complex suffix, where the complex prefix and suffix are composed of zero or more morphemes, and a morpheme is a sequence of sub-morphemes. The boundaries in the output are based on the prefix, stem and suffix levels.

Learning Settings. The input to the learner is a grammar and a vocabulary of unsegmented words. We consider the three learning settings in (Eskander et al., 2016): Standard, Scholar-seeded Knowledge and Cascaded. The Standard setting is language-independent and fully unsupervised, while in the Scholar-seeded-Knowledge setting, some linguistic knowledge (in the form of affixes taken from grammar books) is seeded into the grammar trees before learning takes place. The Cascaded setting simulates the effect of seeding scholar knowledge in a language-independent manner by first running an AG of high precision to derive a set of affixes, and then seeding those affixes into the grammars.

3.1 AG Setups for Polysynthetic Languages

We experimented with several setups using AGs for unsupervised segmentation.

Language-Independent Morphological Segmenter. LIMS is the best-on-average AG setup obtained by Eskander et al. (2016) when trained on six languages (English, German, Finnish, Estonian, Turkish and Zulu), which is the Cascaded PrStSu+SM configuration. We use this AG setup for each of the four languages. We refer to this system as AG_{LIMS}.

Best AG Configuration per Language. In this experimental setup, we consider all nine grammars from Eskander et al. (2016) using both the Standard and the Cascaded approaches and choosing the one that is best for each polysynthetic language by training on the training set and evaluating on the development set. We denote this system as AG_{BestL}.

Using Seeded Knowledge. To approximate the effect of Scholar-seeded-Knowledge in Eskander et al. (2016), we used the training set to derive affixes and use them as scholar-seeded knowledge added to the grammars (before the learning happens). However, since affixes and stems are not distinguished in the training annotations from Kann et al. (2018), we only consider the first and last morphemes that appear at least five times. We call this setup $AG^{Scholar}_{BestL}$.

Multilingual Training. Since the vocabulary in Kann et al. (2018) for each language is small, and the languages are from the same language family, one data augmentation approach is to train on all languages and test then on each language individually. We call this setup AG_{Multi}.

Data Augmentation. In this setup, we examine the performance of the best AG configuration per language (AG_{BestL}) when more data is available. We merge the training corpus with unique words in the New Testament of the Bible (train$_{Bible}$). We run this only on NH and WX since the Bible text is only available for these two languages. We denote this setup as AG_{Aug}.

4 Evaluation and Discussion

We evaluate the different AG setups on the blind test set from Kann et al. (2018) and compare our AG approaches to state-of-the-art unsupervised systems as well as supervised models including the best supervised deep learning models from Kann et al. (2018). As the metric, we use the segmentation-boundary F1-score, which is standard for this task (Virpioja et al., 2011).

Evaluating different AG setups. Table 3 shows the performance of our AG setups on the four languages. The best AG setup learned for each of the four polysynthetic languages (AG_{BestL}) is the PrStSu+SM grammar using the Cascaded learning setup. This is an interesting finding as the Cascaded PrSTSu+SM setup is in fact AG_{LIMS} — the best-on-average AG setup obtained by Eskander et al. (2016) when trained on six languages (English, German, Finnish, Estonian, Turkish and Zulu). This achieves F1-scores of 0.775, 0.744, 0.768 and 0.820 on MX, NH,

Language	AG_{LIMS}	AG_{BestL}	AG_{Multi}	$AG^{Scholar}_{BestL}$	AG_{Aug}	$Morfessor$	$Morphochain$
Mexicanero	0.775	0.775	0.770	**0.798**	-	0.528	0.283
Nahuatl	0.744	0.744	0.723	0.742	**0.759**	0.505	0.259
Wixarika	0.768	0.768	0.746	**0.787**	0.783	0.709	0.283
Yorem Nokki	**0.820**	**0.820**	0.775	0.804	–	0.549	0.351

Table 3: AG systems compared to unsupervised baselines. Bold indicates best scores

Language	$BestAG$	$S2S$	CRF	$BestMTT$	$BestDA$
Mexicanero	0.798	0.862	0.864	**0.879**	0.868
Nahuatl	**0.759**	0.727	0.749	0.739	0.732
Wixarika	0.787	0.796	0.793	0.802	**0.816**
Yorem Nokki	**0.820**	0.773	0.774	0.808	0.792

Table 4: Best AG results compared to supervised approaches from Kann et al. (2018). Bold indicates best scores.

WX and YN, respectively. Seeding affixes into the grammar trees ($AG^{Scholar}_{BestL}$) improves the performance of the Cascaded $PrStSu+SM$ setup only for MX and WX (additional absolute F1-scores of 0.023 and 0.019, respectively). However, it does not help for NH, while it even decreases the performance on YN. This occurs because AGs are able to recognize the main affixes in the Cascaded setup, while the seeded affixes were either abundant or conflicting with the automatically discovered ones. The multilingual setup (AG_{Multi}) does not improve the performance on any of the languages. This could be because the datasets are too small to generalize common patterns across languages. Finally, augmenting with Bible text in the cases of NH and WX leads to an absolute F1-score increase of 0.015 for both languages when compared to AG_{BestL}. There are two possible explanations for why we only see a slight increase when adding more data: 1) AGs are able to generalize from small data and 2) the added Bible data represents a domain that is different from those of the datasets we are experimenting with as only 4.8% and 9% of the words in the training sets from Kann et al. (2018) appear in the augmented data of NH and WX, respectively. Overall, AG_{BestL} is the best setup for YN, $AG^{Scholar}_{BestL}$ is the best setup for MX and WX, while AG_{Aug} is the best for NH.

Comparison with unsupervised baselines. We consider $Morfessor$ (Creutz and Lagus, 2007), a commonly-used toolkit for unsupervised morphological segmentation, and $MorphoChain$ (Narasimhan et al., 2015), another unsupervised morphological system based on constructing morphological chains. Our AG approaches significantly outperform both $Morfessor$ and $MorphoChain$ on all four languages, as shown in Table 3.

Comparison with supervised baselines. To obtain an upper bound, we compare the best AG setup to the best supervised neural methods presented in Kann et al. (2018) for each language. We consider their best multi-task approach (*BestMTT*) and the best data-augmentation approach (*BestDA*), using F1 scores from their Table 4 for each language. In addition, we report the results on their other supervised baselines: a supervised seq-to-seq model ($S2S$) and a supervised CRF approach. As can be seen in Table 4, our unsupervised AG-based approaches outperform the best supervised approaches for NH and YN with absolute F1-scores of 0.010 and 0.012, respectively. An interesting observation is that for YN we only used the words in the training set of Kann et al. (2018) (unsegmented), without any data augmentation. For MX and WX, the neural models from Kann et al. (2018) (BestMTT and BestDA), outperform our unsupervised AG-based approaches.

Error Analysis. For the purpose of error analysis, we train our unsupervised segmentation on the training sets and perform the analysis of results on the output of the development sets based on our best unsupervised models AG_{BestL}. Since there is no distinction between stems and affixes in the labeled data, we only consider the morphemes that appear at least three times in order to eliminate open-class morphemes in our statistics.

We first define the degree of ambiguity of a morpheme to be the percentage of times its sequence of characters does not form a segmentable morpheme when they appear in the training set. We also define the degree of ambiguity of a language as the average degree of ambiguity of the morphemes in that language. Table 5 shows the number of morphemes, average length of a morpheme (in characters) and the degree of morpheme

	Mexicanero	Nahuatl	Wixarika	Yorem Nokki
Number of Morphemes	343	479	434	424
Average Length of a Morpheme	3.17	3.16	3.19	3.40
Degree of Ambiguity	69.81%	73.97%	74.49%	58.67%

Table 5: Morpheme-based Statistics

Language	word	Gold Segmentation	AG_{BestL} segmentation
Mexicanero	tawanitika unipodero tikipiyal	tawani+ti+ka u+ni+podero ti+ki+piya+l	tawani+tika u+ni+pode+ro ti+ki+piya+l
Nahuatl	nannechtlatlaniliáke omokokowaya	nan+nech+tla+tlanilia+'ke o+mo+kokowa+ya	nan+nech+tlatla+nilia+'ke o+mo+kokowa+ya
Wixarika	nep@tiwarutiwawiriwa pep@netsiuta	ne+p@+ti+wa+r+u+ti+wawi+ri+wa pe+p@+ne+tsi+u+ta	ne+p@+ti+waru+ti+wawiriwa pe+p@+ne+tsi+u+ta
Yorem Nokki	ßohoßareka haikimsu'e	ßoho + ßa+ re+ ka haiki+m+su+'e	ßoho + ßare+ ka haiki+m+su+'e

Table 6: Examples of correct and incorrect segmentation

ambiguity in each language. Looking at the two languages where our models perform worse than the supervised models, we notice that MX has the least number of morphemes, and our unsupervised methods tend to oversegment; WX has the highest degree of ambiguity with a large number of one-letter morphemes, which makes the task more challenging for unsupervised segmentation as opposed to the case of a supervised setup. Analyzing all the errors that our AG-based models made across all languages, we noticed one, or a combination, of the following factors: a high degree of morpheme ambiguity, short morpheme length and/or low frequency of a morpheme.

Examples. Table 6 shows some examples of correctly and incorrectly segmented words by our models (blue indicates correct morphemes while red are wrong ones). For MX, our models fail to recognize *ka* as a correct affix 100% of the time due to its high degree of ambiguity (71.79%), while we often wrongly detect *ro* as an affix, most likely since *ro* tends to appear at the end of a word; our approaches tend to oversegment in such cases. On the other hand, our method correctly identify *ki* as a correct affix 100% of the time since it appears frequently in the training data. For NH, the morpheme *tla* has a high degree of ambiguity at 79.12%, which lead the model to fail in recognizing it as an affix (see an example in Table 6). On the other hand, NH has a higher percentage of correctly recognized morphemes, due to their less ambiguous nature and higher frequency (such as *ke*, *tl* or *mo*). For WX, a large portion of errors stem from one-letter morphemes that are highly ambiguous (e.g., *u*, *a*, *e*, *m*, *n*, *p* and *r*), in addition to having morphemes in the training set which are not frequent enough to learn from, such as *ki*,*nua* and *wawi* (see Table 6). Examples of correct segmentation involve morphemes that are more frequent and less ambiguous (*pe, p@* and *ne*). For YN, ambiguity is the main source of segmentation errors (e.g., *wa*, *wi* and *ßa*).slight

5 Conclusions

Unsupervised approaches based on Adaptor Grammars show promise for morphological segmentation of low-resource polysynthetic languages. We worked with the AG grammars developed by Eskander et al. (2016, 2018) for languages that are not polysynthetic. We showed that even when using these approaches and very little data, we can obtain encouraging results, and that using additional unsupervised data is a promising path.

Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), (contract # FA8650-17-C-9117) and the Army Research Laboratory (ARL). The views and conclusions herein are those of the authors and should not be interpreted as necessarily representing official policies, expressed or implied, of ODNI, IARPA, ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Una Canger. 2001. *Mexicanero de la Sierra Madre Occidental*, volume 24 of *Archivo de lenguas indígenas de México*. Centro de Estudios Linguísticos y Literarios, México.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3:1–3:34.

Ramy Eskander, Owen Rambow, and Smaranda Muresan. 2016. Automatically tailoring unsupervised morphological segmentation to the language. In *Proceedings of the Twenty-Sixth International Conference on Computational Linguistics (LREC)*, Osaka, Japan.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2018. Extending the use of adaptor grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of the Fifteenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Brussels, Belgium.

Benoit Farley. 2009. The uqailaut project. Accessed on 10 Jan 2019.

Ray Freeze. 1989. *May de los Capomos*. Sinaloa.

P. Gómez and P.G. López. 1999. *Huichol de San Andrés Cohamiata, Jalisco*. Archivo de lenguas indígenas de México. Colegio de México.

Petr Homola. 2011. Parsing a polysynthetic language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 562–567, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Mark Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio. Association for Computational Linguistics.

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.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-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 47–57.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18. Association for Computational Linguistics.

Judith L. Klavans. 2018a. Computational challenges for polysynthetic languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages (COLING 2018)*, pages 1–11. Association for Computational Linguistics.

Judith L. Klavans. 2018b. *On Clitics and Cliticization: The Interaction of Morphology, Phonology, and Syntax*, 2 edition. London, Routledge.

José Luis Leza. 2004. *Lenguas y literaturas indígenas de Jalisco*. Secretaría de Cultura, Gobierno Estatal de Jalisco. Colección: Las culturas populares de Jalisco, Guadalajara, Mexico.

Patrick Littell. 2018. Finite-state morphology for kwak'wala: A phonological approach. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages (COLING 2018)*, pages 21–30. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Diónico Carrillo, and Iván V. Meza-Ruíz. 2018b. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent and Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018c. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018d. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages (COLING 2018)*, pages 73–83.

Jeffrey Micher. 2017. Improving coverage of an inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106. Association for Computational Linguistics.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for arapaho verbs learned from a finite state transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages (COLING 2018)*, pages 12–20. Association for Computational Linguistics.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. In *Twelfth AAAI Conference on Artificial Intelligence*.

John R. Ross. 1972. Endstation hauptword: The category squish. *Chicago Linguistic Society*, 8:316–328.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. *Transactions of the Association for Computational Linguistics*, 1(May):231–242.

Yolanda Lastra de Suárez. 1980. *N´a]huatl Acaxochitlán, Hidalgo*, volume 10 of *Archivo de lenguas indígenas de México*. Colegio de México, México.

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.