Journal of Intelligent & Fuzzy Systems 34 (2018) 3081–3087
DOI:10.3233/JIFS-169492
IOS Press

Probabilistic Finite-State morphological segmenter for Wixarika (huichol) language[1]

Manuel Mager*, Diónico Carrillo and Ivan Meza
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México Circuito Escolar, Sin Número, Ciudad Universitaria, Ciudad de México, México

Abstract. In this work, we present a morphological segmenter for the Mexican indigenous language Wixarika. Segmentation is fundamental for rich morphological languages, a common aspect of the native American languages, to improve other tasks like machine translation, dialogue systems, summarization, etc. On top of the agglutinative nature of the language, the low amount of resources and the lack of an orthographic standard among dialects add to the challenge. Our proposal is based on a probabilistic finite-state approach that exploits regular agglutinative patterns and requires little linguistic knowledge. We show that our approach outperforms unsupervised and semi-supervised methods in a low-resource context. The dataset used in this work was openly released for future work by the community.

Keywords: Morphology, low resources, finite-state transducer, Wixarika, endangered languages

1. Introduction

Indigenous languages face several challenges in the new technological context. Many of them are endangered due a new technological reality based on other spoken languages that are more culturally dominant. People from all ages now use different computing technology (mobile phones, smart phones, laptops, computers) with their own language, however it is common to find among indigenous native speakers code switching (mix of native language with dominant language), or that they use the dominant language. Natural Language Processing (NLP) can contribute to preserve and vitalize these languages by providing ways these can be analyzed and used in high level NLP applications such as: translation, dialogue systems, summarization, etc.

Wixarika is a language spoken in the Mexican states of Jalisco, Nayarit, Durango and Zacatecas (central west of Mexico). It is approximately spoken by fifty thousand people. Like most South and North American indigenous languages, Wixarika has complex verbal morphology [2]. For instance, the word *nep+ka'ukats+k+*, which can be translated into English as "I don't have a dog" is segmented into the morphs *ne|p+ |ka|'u|ka|ts+k+*. It is important to notice that in Wixarika the symbol + is used to denote one of the vowels in the language; for this reason, we will use | symbol to delimit its morphemes. Notice that although this word is a verb form, its agglutinative nature makes it a full sentence. In this example *ts+k+* is the stem and means "dog", *ne* is a first person possessive, *ka* negation, *'u* refers to a visual object and *ka* is the second part of the negation. The study of the Wixarika language from the point of view of technologies is difficult, first because NLP resources that allow the computationally process of the language are few and second the morphological richness makes it harder to adapt tools from more common studied languages such as English or Spanish. Since the first step in many of the NLP tasks depends in

[1]We thank the guidance, feedback and support of Professor Jason Eisner as a mentor of this work.

*Corresponding author. Manuel Mager, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México Circuito Escolar, Sin Número, Ciudad Universitaria, Ciudad de México, 04510 México. E-mail: mmager@turing.iimas.unam.mx.

Table 1
Wixarika alphabet normalization. The symbols a, e, h, i, k, m, o, p, r, t, u y appear in all versions

Used symbol	Other representations
+	ü, ʌ, ï, cu, ɨ
k	c
kw	k^w
ts	tz, ch,
w	v
x	rr,
’	?, ϵ

a correct segmentation of the language sentences, in this work we present a probabilistic finite-state morphological analyzer for the Wixarika language[2]. The goal of the segmenter is to identify the boundaries among morphs of the language, a good segmentation will have a great effect on the performance of other tasks such: machine translations, summarization, etc.

Different linguistic studies have recorded Wixarika in written form, but its spelling is still not standardized. The most common spelling in practice by native speakers is an alphabet of 18 symbols: $\Sigma = \{a, e, h, i, +, k, m, n, p, r, t, s, u, w, x, y, ’\}$, as proposed in Gómez [7] and Iturrio and Gómez López [13]. In this work we follow this convention for the description of the work and the resources created. When the language imports words from other languages, like Spanish, unused symbols can be added to the enumerated alphabet. Although, Spanish words are also often adapted to the existing alphabet, e.g. the name Jesus uses *j* and can be transformed to *kets+*. Table 1 shows a normalization of the symbols for other alphabets conventions.

The output of the morphological segmenter is a list of substrings called morphs given a w word in Wixarika. Past research has focused on unsupervised methods, but they can only be applied to languages for which there exists a sufficiently large corpus of words [19]. This task can be done on the surface level in high agglutinative languages, and for those languages we do not need to infer any fusioned morpheme type. In this work, we propose a semi-supervised approach in which a seed of morphological labeling interact with a set of rules in a hybrid fashion to produce the corresponding segmentation.

For indigenous languages like Wixarika with scarcity of digital available resources defines a bound on the performance of these methods. This scarcity broad for most Mexican indigenous languages. For instance, efforts to gather large collections of digital texts for Yutonahua languages exist only for Nahuatl [10]. In the case of Wixarika some prior work on Statistical Machine Translation has been done [17] but the set of examples still is limited.

On the other hand, rule-based automatic morphological analyzers require deep knowledge of the language or the expensive support of linguists [4]. Rule-based morphological analyzers have been developed for Quechua, Toba [18] and Aymara languages [12]. However, for a poorly studied language it is difficult to create such resources because the research switches from the computational aspects to the linguistic properties of the language, although this type of research are also necessary for all indigenous languages in this work we choose to focus on the computational aspect.

Our approach to the morphological segmentation of Wixarika deals with the scarcity of linguistic knowledge and a large digital formatted corpora, since we propose a hybrid system that combines morphological knowledge from descriptive grammars of the language with a probabilistic model learned from supervised data (previously seen segmented words).

There are also efforts for NLP tasks for other languages, such as Machine Translation (MT) from/to such languages, which has usually been done with rule-based approaches. For example the Quechua - Spanish translator developed by [1] uses the rule-based Apertum Translation System [5], and Zapoteco-Spanish translation application [8].

Our contribution is the construction of the first morphological analyzer for Wixarika, using hand-specified lists of legal stems and affixes together with an n-gram model that describes sequences. This hybrid method can achieve good performance for a morphologically rich language with scarce resources and low grammatical knowledge.

2. Method

Wixarika belongs to the family of Yutonahua languages, such as Nahua, Nayeri, Raramuri, etc. These languages have agglutinative morphology, using prefixation as well as suffixation around the verb stem. Nouns can be used to act as stems in verbs. The affix “p+” serve as the verbification morpheme. The agglutination is almost strictly concatenative, and each morpheme must be realized at a specific position in the word. The same string in a different position conveys a different meaning: e.g., the prefix *ne-* in

[2]The analyzer is open source and can be download it from: https://github.com/pywirrarika/smtwixes/tree/master/wixnlp

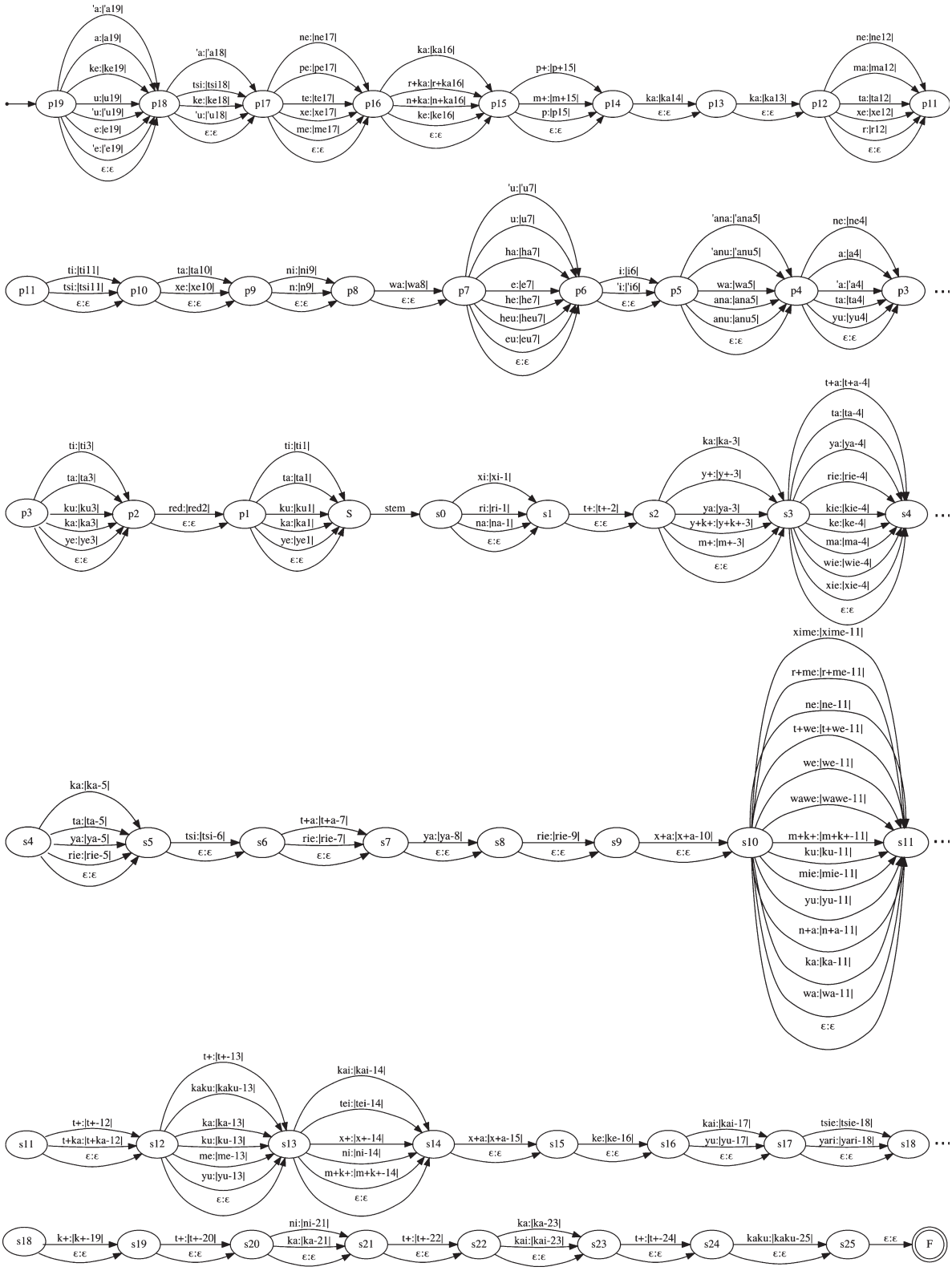

Fig. 1. Final State Transducer (FST) for the wixarika language. The “stem” arc stands for a collection of 374 arcs representing different stems.

position 16 acts as a pronominal morpheme, but in position 3 it is a possessive morpheme [7]. There are 18 such prefix positions and 23 suffix positions identified by Iturrio and Gómez López [13], where each position allows a certain set of morphemes (or can be left empty).

This description of the language can be used to construct a finite-state transducer (FST) from a list of legal morphemes at each position. Although there are more complex rules that govern sequences of morphemes, we will assume that the only condition is that each position allows only morphemes from its list. The errors introduced by this assumption will be corrected later by the n-gram model.

The stem is not defined by any rule and it can be based on words from other languages (e.g., Spanish). For the present study, however, we limited the possible stems to a tuple of 374 strings learned from examples. The list of sets used for affixes was taken from the linguistic work of Iturrio and Gómez López [13], which is a revision of an earlier study [9] and contain 131 morphemes.

A finite-state transducer can accept any string w in the language that it defines, and returns a set of accepting paths. The complete automaton for Wixarika verbs is shown in Fig. 1; its different accepting paths for a word w correspond to different morphological analyses of w. The FST has 47 states the prefix states are marked as p while the suffix are marked with s and it is composed by 374 different arcs.

In practice, there are few enough analyses that we can enumerate all of them. To choose the most probable analysis from among these, we used a simple n-gram model with Kneser-Ney smoothing [15], where each gram is a morph (a surface string associated with some morpheme). This model scores the sequence of non-empty strings (morphs) without considering their absolute positions. As a result, it can be trained simply from a segmented corpus.

As the automaton already found all possible segmentations, we only need to evaluate the probabilities for each n-gram and return the best-ranked segmentation. This was done by enumerating all of the segmentations and scoring them separately with the n-gram model.

It is also possible to place weights on the FSA arcs, but placing weights on each arc would only allow a 1-gram model. For an n-gram model, we need to split states, so that each state remembers the previous $n - 1$ morphs; that would result in a much bigger machine. Irregular agglutinations and unknown stems can mislead the automaton, so it sometimes fails to recognize an input word. If this happens, we can fall back to an unsupervised method to analyze this word. Usually an unsupervised analyzer under-performs with scarce resources, but it can improve the final segmentation in practice.

Table 2
Results for the morphological segmentation task on Wixarika using direct comparison to the gold segmentation: Edit distance (ED) and accuracy (Ac.). Results for the morphological segmentation task on Wixarika using EMMA metric. M. stands for Morfessor, P for precision, R for recall and F for the F-measure

Method	ED	Ac.	P	R	F
Best M. WND	1.788	0.272	0.503	0.564	0.532
Best M. WLD	1.784	0.262	0.589	0.618	0.603
Best M. WSD	1.112	0.326	0.646	0.656	0.650
M. Viterbi	1.020	0.376	0.669	0.662	0.665
WixNLP	1.002	0.478	0.666	0.724	0.694
WixNLP 2-grams	0.933	0.485	0.598	0.623	0.610
WixNLP 3-grams	0.803	0.580	0.727	0.758	0.742
Hybrid 2-grams	0.738	0.565	0.743	0.777	0.760
Hybrid 3-grams	**0.639**	**0.600**	**0.782**	**0.808**	**0.795**

3. Results

For our experiment we collected two Wixarika corpora as shown in Table 4. The first is a high-quality segmented text taken from a grammar [7] containing 1,079 type words, which we used as our gold standard. We randomly selected 400 words from these words, to be used as a test set, and the rest were used for the training of 51 semi-supervised Morfessor models and our n-gram model trained on the probabilities on morph tokens [23].

The second source of words is a translation of Hans Christian Andersen's classic fairy tales[3] to Wixarika containing 17,131 non segmented word types, used for the training of the unsupervised based on Morfessor model. It is important to notice that both resources although in Wixarika language, they are different dialects.

Evaluating morphological segmentations is difficult since for a single word there are several valid segmentations. There are two types of metrics for morphologies: those that directly compare the hypotheses against the gold standard and those that perform the comparison indirectly "by measuring the strength of an isomorphic like relationship between the proposed and answer morphemes" [20].[4]

[3]The dataset is available from https://github.com/pywirrarika/wixarikacorpora

[4]For a comparison among the various metrics see [22].

Table 3
Results for the morphological segmentation task using different semi-supervised Morfessor setting on Wixarika. P stands for precision, R for recall and F for the F-measure

Algorithm		With segmented d. WSD			With nonsegmented d. WND			With large d. WLD		
	CW	P	R	F	P	R	F	P	R	F
Recursive	0.0	0.0732	0.1385	0.0958	0.0732	0.1385	0.0958	0.0697	0.1337	0.0916
	0.1	0.3918	0.5309	0.4509	0.3918	0.5309	0.4509	0.5892	0.6157	0.6022
	0.5	0.4713	0.5625	0.5129	0.4713	0.5625	0.5129	0.5948	0.5866	0.5907
	1.0	0.4980	0.5896	0.5400	0.4980	0.5896	0.5400	0.5856	0.5547	0.5697
	1.5	0.5502	0.5993	0.5737	0.5502	0.5993	0.5737	0.5474	0.5057	0.5257
	2.0	0.6105	0.6360	0.6230	0.6105	0.6360	0.6230	0.5561	0.5143	0.5344
	2.5	0.6364	0.6565	0.6463	0.6364	0.6565	0.6463	0.5619	0.5158	0.5379
	3.0	0.6222	0.6409	0.6314	0.6222	0.6409	0.6314	0.5558	0.5097	0.5318
	3.5	0.6312	0.6401	0.6356	0.6312	0.6401	0.6356	0.5293	0.4887	0.5082
	4.0	0.6368	0.6483	0.6425	0.6368	0.6483	0.6425	0.5313	0.4932	0.5115
	4.5	0.6363	0.6434	0.6398	0.6363	0.6434	0.6398	0.5504	0.5135	0.5313
	5.0	0.6399	0.6518	0.6458	0.6399	0.6518	0.6458	0.5647	0.5259	0.5446
	5.6	0.6455	0.6552	0.6503	0.6455	0.6552	0.6503	0.5224	0.4856	0.5033
	6.0	**0.6471**	**0.6565**	**0.6518**	0.6471	0.6565	0.6518	0.5403	0.5035	0.5212
	6.5	0.6431	0.6513	0.6472	0.6431	0.6513	0.6472	0.5524	0.5105	0.5307
	7.0	0.6333	0.6396	0.6365	0.6333	0.6396	0.6365	0.5567	0.5176	0.5364
Viterbi	–	0.6687	0.6617	0.6652	0.4173	0.4557	0.4356	0.3333	0.2692	0.2979

In this work, we used both types of metrics. For direct comparison we follow Kann et al. [14], using the accuracy and the edit distance of morphs between the hypothesis and the golden standard. For the indirect evaluation we used EMMA [20], which produces precision, recall and F-measure scores.

Table 3 presents a summary of the direct results using Moferssor under three scenarios: with segmented data (WSD), without segmented data (WND) and with large dataset (corpus) (WLD). In the first case, the segmented data, this is the 679 training words are passed twice to Morfessor first in a semi-supervised and the second as a fully superviside. In the second case, Morfessor only sees the collection of 679 words but without segmentation. Finally, in the third case add the 17, 131 words to the second case. As it can be appreciated there is not difference among first and second cases, we attribute this to dialect differences among the resources.For training the Morfessor model we varied the corpus weight value from 0.1 to 7 in steps of 0.5 for the recursive algorithm. In addition we also trained a model with the viterbi algorithm. In this last one case we see that viterbi is more susceptible to make errors without the segmented information. We also see the importance of the corpus weight, a small value produces a weak segmenter, however there is a good spot to define it.

Our proposed **WixNLP** system looks for all possible paths in the forward graph and chooses the shortest valid path. **WixNLP with n-grams** estimates the most probable segmentation among the valid paths. It is based on the FST presented in Fig. 1. We compared its performance against **Semi-Supervised Morfessor** [23] with a range of settings that can be analyzed in Table 3. Finally, the **Hybrid** methods use WixNLP when possible but fall back to Unsupervised Morfessor for words that have no valid paths.

Table 4
Description of the two data sets used to train the segmentation models. The small data set is a high quality phrase collection of segmented words. The large data set is a parallel corpus of a translation of Hans Christian Andersen's classic fairy tales, but has no segmentation available

	Small Data Set	Large Data Set
Tokens	2,537	56,037
Unique Words	1,079	17,131
Morphemes	3,870	–
Unique Morphs	241	–

Table 2 show that the experimental results using Morfessor suffer from the lack of examples in order to infer a good model. However, this can improve by using the segmented corpus as base of the model. In all trained models we included a semi-supervised setting using a list of words with their segmented version from the development set. We only changed the amount and form of the base data of the model: those trained with a non segmented corpus got a poorer results as those trained with only segmented data. Using a larger dataset helped, but was not the best version. On the other hand, our first model, WixNLP without maximizing the probabilities of paths improves Morfessor without even using probabilities for disambiguation. The criterion used for choosing a path was the path of minimum

length. WixNLP with 2-grams under performs normal WixNLP in indirect evaluation but obtains better results in accuracy and edit distance. The version using 3-grams improve the results notably in all metrics. The hybrid approach deals with the problem of unseen roots and suffixes, and thus achieves the best results in all metrics, particularly with a 3-gram model.

4. Related work

Morphological segmentation has a large history, Harris et al. in 1951 did the first research on the task [11]. Since then, many approaches have been developed. Two essential systems were used with good results for semisupervised segmentation: LINGUISTICS [6] and the extended version of MORFESSOR [16]. MORFESSOR also has an earlier unsupervised method [3]. But, rule-based language specific FSM has been developed for many languages archiving good results. For indigenous languages such morphological analyzers have been developed for Quechua, Toba [18] and Aymara languages [12]. These works differentiate from our approach since they try to model the complete underlying morphology of their languages, and WixNLP only uses a list of morphemes and infer the morphological rules. For the specific case of Wixarika, there was no previous attempt to implement any finite-state approach.

5. Conclusion

Morphological segmentation is an important task for language processing for the Wixarika language. In this work we presented the first Wixarika morphology analyzer, a finite-state transducer that exploits the agglutinative pattern of Yutonahua languages, with lists of stems and affixes, together with a n-gram model to estimate the best segmentation among multiple matches. We showed that for Wixarika our method improves on the Morfessor baselines.

We also created and publicly released a parallel Wixarika-Spanish dataset to encourage the community to study this language further. Together with these corpora we release a couple of NLP tools, such as a normalizer and tokenizer to handle Wixarika texts. This is an important tool due the lack of an orthographic standardization among the native speakers.

For future work, we would apply this methodology to other Yutonahua languages. We also want to feed the morphological segmentation to a MT system. It is also desirable to find improved methodologies to combine unsupervised with supervised methods to address the scarce resource problem for agglutinative languages, including tagging each morph as in [21].

References

[1] H.D. Calderón, V.C.D. Mamani Calderón, F.C. Cárdenas Mariño and E.F. Mamani Calderón, Automatic translator in line Spanish a Quechua, based on free and open source platform apertium, *Revista Investig (Esc Post Grado)* **5**(3) (2009), ISSN 1997-4035.

[2] L. Campbell and V. Grondona, The indigenous languages of South America: A comprehensive guide, volume 2. Walter de Gruyter, 2012.

[3] M. Creutz and K. Lagus, Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pp. 21–30. Association for Computational Linguistics, 2002.

[4] M. Creutz and K. Lagus, Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Helsinki University of Technology, 2005.

[5] M.L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez and F.M. Tyers, Apertium: A free/opensource platform for rule-based machine translation, *Machine Translation* **25**(2) (2011), 127–144. ISSN 1573-0573.

[6] J. Goldsmith, Unsupervised learning of the morphology of a natural language, *Computational Linguistics* **27**(2) (2001), 153–198.

[7] P. Gómez, Huichol de San Andrés Cohamiata, Jalisco. Archivo de lenguas indígenas de México. Colegio de México, 1999. ISBN 968120851X.

[8] S. Gonzalo, App para traducir zapoteco – DIDXAZAPP. http://aprendezapoteco.blogspot.mx/2016/03/appparatraducir-zapoteco-didxazapp.html, 2016. [Online; accessed 2017-04-25].

[9] J. Grimes, Huichol Syntax. Mouton, The Hague; 1964.

[10] X. Gutierrez-Vasques, G. Sierra and I.H. Pompa, Axolotl: A web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

[11] Z.S. Harris, Methods in structural linguistics; 1951.

[12] P. Homola, Parsing a polysynthetic language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 562–567, Hissar, Bulgaria, September 2011. RANLP 2011 Organising Committee. URL http://www.aclweb.org/anthology/R11-1079

[13] J.L. Iturrio and P. Gómez López, Gramática Wixarika I. Archivo de lenguas indígenas de México. Lincom Europa; 1999.

[14] K. Kann, R. Cotterell and H. Schútze, Neural multi-source morphological reinflection. In *Proceedings of the 2017 Conference European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April 2017.

[15] R. Kneser and H. Ney, Improved backing-off for m-gram language modeling. In ICASSP; 1995.

[16] O. Kohonen, S. Virpioja, L. Leppánen and K. Lagus, Semisupervised extensions to morfessor baseline. In *Proceedings of the Morpho Challenge 2010 Workshop*, pp. 30–34, 2010.

[17] J.M. Mager Hois, C. Barron Romero and I.V. Meza Ruíz, Traductor estadístico wixarika - español usando descomposición morfológica, *COMTEL* (6), 2016.

[18] A.O. Porta, The use of formal language models in the typology of the morphology of amerindian languages. In *Proceedings of the ACL 2010 Student Research Workshop*, pp. 109–114, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-3019

[19] T. Ruokolainen, O. Kohonen, K. Sirts, S.-A. Grónroos, M. Kurimo and S. Virpioja, A comparative study of minimally supervised morphological segmentation, *Computational Linguistics* (2016).

[20] S. Spiegler and C. Monson, Emma: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics,* COLING '10, pp. 1029–1037, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[21] D.J. Spoustová, J. Hajič, J. Votrubec, P. Krbec and P. Květoň, The best of two worlds: Cooperation of statistical and rulebased taggers for czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, ACL '07, pp. 67–74, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1567545.1567558

[22] S. Virpioja, V.T. Turunen, S. Spiegler, O. Kohonen and M. Kurimo, Empirical comparison of evaluation methods for unsupervised learning of morphology, *TAL* **52**(2) (2011), 45–90.

[23] S. Virpioja, P. Smit, S.-A. Grónroos and M. Kurimo, Morfessor 2.0: Python implementation and extensions for Morfessor baseline. D4 julkaistu kehittámis- tai tutkimusraportti tai -selvitys, 2013. URL http://urn.fi/URN:ISBN:978-952-60-5501-5