2019 SAUPEC/RobMech/PRASA Conference
Bloemfontein, South Africa, January 28-30, 2019

Towards an unsupervised morphological segmenter for isiXhosa

Lulamile Mzamo, Albert Helberg
Faculty of Engineering
North-West University
Potchefstroom, South Africa
lula_mzamo@yahoo.co.uk, albert.helberg@nwu.ac.za

Sonja Bosch
Department of African Languages
UNISA
Pretoria, South Africa
boschse@unisa.ac.za

***Abstract*—In this paper, branching entropy techniques and isiXhosa language heuristics are adapted to develop unsupervised morphological segmenters for isiXhosa. An overview of isiXhosa segmentation issues is given, followed by a discussion on previous work in automated segmentation, and segmentation of isiXhosa in particular. Two unsupervised isiXhosa segmenters are presented and compared to a random minimum baseline and Morfessor-Baseline, a standard in unsupervised word segmentation. Morfessor-Baseline outperforms both isiXhosa segmenters at 79.10% boundary identification accuracy. The IsiXhosa Branching Entropy Segmenter (XBES) performance varies depending on the segmentation mode used, with a maximum of 73.39%. The IsiXhosa Heuristic Maximum Likelihood Segmenter (XHMLS) achieves 72.42%. The study suggests that unsupervised isiXhosa morphological segmentation is feasible with better optimization of the current attempts.**

Keywords—natural language processing; unsupervised machine learning; morphological segmentation; isiXhosa.

I. INTRODUCTION

Human language resources and applications currently available in South Africa are of a very basic nature. According to [1] this can be attributed to the dependence on Human Language Technology (HLT) expert knowledge, scarcity of data resources, lack of market demand for African languages, and how the particular language relates to other more resourced languages. Morphological analysis is one of the basic tools in the natural language processing (NLP) of agglutinating languages. The work detailed in this paper is the development of morphological analysers for one such language, namely isiXhosa.

IsiXhosa is one of the South African official languages belonging to the Bantu language family which are classified as "resource scarce languages" [1]. Although some work has been done in computational linguistic tools for isiXhosa, it is of a limited nature [2]. IsiXhosa is the second largest language in South Africa with 8.1 million mother-tongue speakers (16% of the South African population), second only to isiZulu [3].

IsiXhosa is closely related to other Nguni languages such as isiZulu, Siswati and isiNdebele and therefore work done in it could easily be bootstrapped to these languages as has been shown in [4].

II. MORPHOLOGICAL SEGMENTATION FOR ISIXHOSA

A. Morphological Segmentation

Language is productive, meaning that it produces new words from existing words. Morphology, therefore, is the study of the internal structure of words [5], and the processes involved in language productivity [6]. The major types of morphological processes are inflection, derivation and compounding.

Morphological analysis splits one token, a word, into several [6], e.g. the segmentation of a word into its constituent morphemes, and classification thereof. Morphemes are the smallest meaning bearing component of a word [5]. In languages with rich systems of inflection and derivation, morphological analysis is needed in information retrieval, translation etc.

Hammarstrom and Borin [7] differentiate between morphological segmentation, which splits words into constituent morphemes, and analysis, which also classifies the identified morphemes, a differentiation originated by Zwicky [8]. The task handled in this paper is morphological segmentation.

B. Morphological segmentation in isiXhosa

IsiXhosa is an agglutinating and polysynthetic language in that it usually has many morphemes per word [5]. It is also fusional/inflectional because morpheme boundaries are sometimes fused and difficult to distinguish.

IsiXhosa words are composed of a root, prefixes, suffixes and circumfixes that attach to the root which is the meaning carrying constituent of the word. A circumfix is the "simultaneous affixation of a prefix and suffix to a root or a stem to express a single meaning" [5]. An example of a circumfix in isiXhosa is the combination "*a…ang..*" in isiXhosa negation, e.g. ***a-****ka-hamb-****ang-****a (he/she* ***did not*** *go).*

Each of the affixes (i.e. prefixes, suffixes or circumfixes) is made up of one or more morphemes. Morphemes follow one another in an order prescribed for each word type [9]. In isiXhosa, most roots are however bound morphemes, meaning that they never appear independently as words which are independently meaningful [10]. They at least appear as stems, which are word roots suffixed with a termination vowel [9].

C. Automated Morphological segmentatin of isiXhosa

One of the earliest reports on automated morphological segmentation of South African languages is that of [11] on the automatic acquisition of a Directed Acyclic Graph (DAG) to model the two-level rules for morphological analysers and generators. The algorithm was tested on English adjectives, inflection of isiXhosa noun locatives and Afrikaans noun plurals, with a 100% accuracy for isiXhosa noun locatives inflection.

Bosch et al. [4] bootstrapped an existing isiZulu morphological analyser [12] to other Nguni languages including isiXhosa. The study reported that 93.30% of the words (181) were analysed.

Eiselen and Puttkammer [13] presented work on the development of text resources for ten South African languages, including a morphologically analysed corpus for isiXhosa. That morphological segmentation corpus is used in this study. The corpus is rated at an accuracy of 84.66%.

The most recent work for isiXhosa segmentation is that of [14] for the development of a rule-based noun stemmer for isiXhosa. The exercise reported a segmentation accuracy rate of 91% for noun segmentation.

III. UNSUPERVISED MORPHOLOGICAL SEGMENTATION

The last works done for morphological segmentation for isiXhosa are reported in [13] and [14]. Both works were based on linguistic rules. In contrast, the work presented in this paper uses unsupervised machine learning in the morphological segmentation of isiXhosa. This is attractive because it bypasses the need for expensive linguistic experts or annotated training data.

A. Unsupervised Machine Learning

There are three modes of training a machine learning model, i.e. supervised, semi-supervised and unsupervised. [6, p. 232]. In supervised learning, the training data contains solution examples that the model must generalise from. Data in unsupervised training is devoid of such, but only models from raw input data as provided. Semi-supervised systems use anything in between, from using limited supervised data with large amounts of unannotated data to unannotated data with heuristics built into the model.

The work presented in this paper involves looking for unsupervised ways of morphologically segmenting isiXhosa.

B. Unsupervised Morphological Segmentation works

The earliest works in unsupervised morphological segmentation used a form of accessor variety, where a morpheme boundary is identifiable by the possible number of letters that may follow a sequence of letters [15], [16]. This evolved to using mutual information [17], [18], and different forms of Branching Entropy [17], [19].

Minimum Description Length (MDL) [20] has seen extensive use in unsupervised morphological segmentation, primarily as a measure of fit of the training data to heuristic models and statistical models [21], [22]. The comparative standard used in this study, Morfessor-Baseline [23], uses MDL and Maximum likelihood estimation.

Clustering and paradigmatic models have also been used. These involve clustering related words using a similarity measure, identifying the stem, and considering the rest as sequences of affixes [24], [25]. The similarity measures used are Latent Semantic Analysis [26], Dice and Jaccard coefficients [6], Ordered Weighted Aggregator operators [25] and affixality measurements [27]. Word context is also another technique that is used to identify similar words [28], [29].

A number of non-parametric Bayesian techniques have also shown promise, including Pitman-Yor process based models [30], [31] and adaptor grammars [32]. These use Markov Chain Monte Carlo (MCMC) simulation with Gibbs Sampling [33] for inference. Contrastive Estimation [34], [35] is another non-parametric model that is showing elegance and promising results.

A number of studies have used a combination of the above techniques and measures [14], [29].

C. Choice of unsupervised segmenter for comparison

To place this work amongst other segmenters, a standard in morphological segmentation was chosen for comparison. The segmenter had to be publicly available and had to have been used for highly agglutinative languages like isiXhosa.

The Morfessor-Baseline segmenter [23] was chosen because it is a good benchmark that has been used extensively and is freely available. A number of studies referred to in the previous section use Morfessor-Baseline as a benchmark.

As a minimum baseline a random segmenter that randomly decides whether a point in a word is a boundary of a segment or not was implemented.

IV. UNSUPERVISED ISIXHOSA SEGMENTERS

This study investigates the use of different variants of Branching Entropy as detailed in [36] and propose novel isiXhosa heuristics with maximum likelihood estimation.

This paper presents two unsupervised morphological segmenters, one based on variants of branching entropy, the IsiXhosa Branching Entropy Segmenter (XBES), and one based on isiXhosa heuristics and multi-gram frequencies, the IsiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS).

A. How XBES works

XBES is based on the work of [36], implementing all the variants of branching entropy segmentation detailed therein.

Branching entropy techniques are based on the understanding that as one moves through the letters in a word, the predictability of the next letter increases, meaning the uncertainty reduces. An increase in uncertainty, measured as entropy, implies a morpheme boundary. Branching Entropy is calculated according to (1), which shows Right Branch Entropy and Left Branching Entropy across the X_k and X_{k-1} letters.

$$H\left(X_{k;k-1}\middle|x_{k-1:k-n}: x_{k:k+n-1}\right) =$$

$$h_{\rightarrow}(x_{0..n}) + h_{\leftarrow}(x_{0..n}) =$$

$$-\sum_{x\in\chi} P(x|x_{k-1:k-n}) log P(x|x_{k-1:k-n})$$
$$-\sum_{x\in\chi} P(x|x_{k:k+n-1}) log P(x|: x_{k:k+n-1}) \quad (1)$$

where, $x_{0..n}$ denotes a string of length n, $X_{k;k-1}$ is the word location between letters X_k and X_{k-1}, $x_{i:j}$ is the word substring from the i-th letter to the j-th letter, χ is the vocabulary of letter in the language, and $h_{\rightarrow}$ and $h_{\leftarrow}$ are the right branching and left branching entropies respectively.

Variation of Branching Entropy (VBE), (2), takes branching entropy further, by assuming that branching entropy generally decreases as one goes along the word [37]. A difference that is above zero (0) indicates a boundary:

$$\delta h_{\rightarrow}(x_{0..n}) = h_{\rightarrow}(x_{0..n}) - h_{\rightarrow}(x_{0..n-1})$$

$$VBE(x_{0..n}) = \delta h_{\rightarrow}(x_{0..n}) + \delta h_{\leftarrow}(x_{0..n}) \quad (2)$$

Normalised Variation of Branching Entropy (NVBE) is an extension of VBE where the VBE is dampened with the word location average VBE, and/or the VBE standard deviation:

$$N_u VBE\ (x) = VBE(x) - \mu_k$$

$$N_z VBE\ (x) = \frac{VBE(x) - \mu_k}{\sigma_k} \quad (3)$$

Magistry [36] continues to define an n-gram autonomy *a(x)* as:

$$a(x) = \delta h_{\leftarrow}(x) + \delta h_{\rightarrow}(x) \quad (4)$$

where the *δh*'s, the Variation of Branching Entropies, are measured at the edges of the n-gram *x*. The higher the autonomy of a word segment the higher the likelihood that it is a morpheme.

This n-gram autonomy can then be used in scoring a word segmentation, such that the segmentation to choose is:

$$\arg\max_{S \in Seg(s)} \sum_{w_i \in S} a(w_i).len(w_i) \quad (5)$$

where *Seg(s)* is a list of possible segmentations of string *s*, *S,* a particular segmentation, contains the list of the segments of word, w_i is the i-th segment of segmentation *S*, *a(.)* is the autonomy measure, and *len(.)* is the string length.

XBES implements all the above segmentation variants.

XBES is trained with character level multi-grams, which are character n-grams of different lengths, and stores multi-gram frequencies. The current implementation of the multi-gram frequencies is not smoothed, and that would be the next development in enhancing XBES.

B. How XHMLS works

The isiXhosa Heuristics Maximum Likelihood Segmenter (XHMLS) that is proposed is based on the directed graphical model of isiXhosa (Fig. 1) and isiXhosa heuristics.

In Fig. 1 and (6), *w* is a word, *S* is a word segmentation, *r* is a word root, *c* is a circumfix, *p* is a prefix, *s* is a suffix, p_i is the i-th prefix morpheme, s_j is the j-th suffix morpheme, and $p_{i-1:1}$ and $s_{j-1:1}$ are preceding sequences of prefix and suffix morphemes respectively.

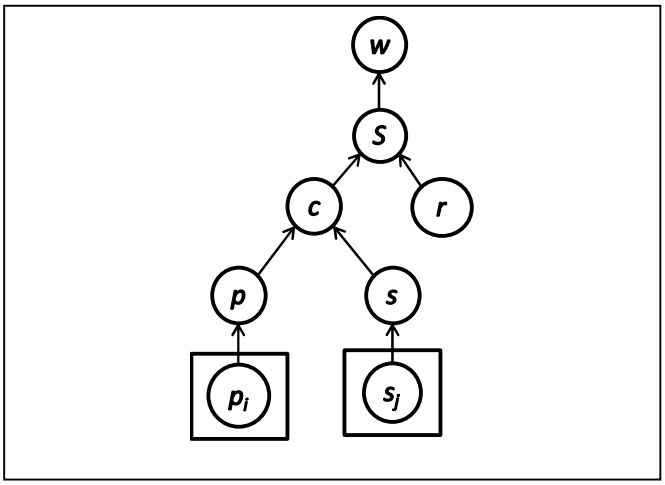

Fig. 1. XHMLS's Probabilistic Graphical Model

A word is modelled as a segmentation which consists of a root and a circumfix. The circumfix is made up of a prefix and a suffix. A prefix is a sequence of prefix morphemes and a suffix a sequence of suffix morphemes. The probability of a word *w* is modelled as:

$$P(w) = P(r|c) \cdot P(c|p) \cdot P(c|s) \cdot P(p) \cdot P(s)$$

$$P(p) = \prod_{i}^{|p|} P(p_i|p_{i-1:1})$$

$$P(s) = \prod_{j}^{|s|} P(s_j|s_{j-1:1}) \quad (6)$$

To minimise the search space the following isiXhosa word heuristics are used (e.g. *u-ya-ndi-phek-el-a* [s/he is cooking for me]:

- The first vowel is always a morpheme;
- The last vowel is always a morpheme (terminal vowel);
- Prefix morphemes are complete syllables, except for 'm' which has a silent vowel when followed by a consonant;
- Suffix morphemes start with a vowel and end in a consonant except for the terminal vowel and for 'w' which has a silent preceding vowel when following a consonant;
- roots start with a consonant.

The above heuristics are not linguistic, but are meant to simplify isiXhosa segmentation.

During training each training word is split into all possible combinations of the above heuristics and statistics are kept in three statistical models, i.e. root–circumfix, prefix n-grams and suffix-n-gram models.

As of publication the models are not smoothed and the intention is to smooth with modified Kneser-Kney smoothing [38].

V. EVALUATION

This section details the initial evaluation that was done on XBES and XHMLS.

A. Data Sources[1]

Raw unannotated isiXhosa data was compiled, to 1.45 million tokens from the isiXhosa version of the South African Constitution [39], isiXhosa stories on the internet and the IsiXhosa Genre Classification Corpus [40]. This text is named the training corpus.

For testing purposes the NCHLT IsiXhosa Text Corpus (29 511 tokens) was used.

B. Data Splits

For development 10000 tokens from the training corpus were used for training, and the first 1000 tokens from the NCHLT IsiXhosa Text Corpus were used for validation.

For evaluation training purposes the entirety of the training corpus was used.

For evaluation purposes a subset of the NCHLT corpus was used. Because the NCHLT corpus was generated with a rule based morphological analyser, the solutions are not all strictly segmentations, others include linguistic morphemes. Excluding such entries resulted in an evaluation corpus of 17574 tokens.

C. Experiment setup

Training was performed for all four segmenters, i.e. the random segmenter, XBES, XHMLS, and Morfessor-Baseline, using the entirety of the training corpus, and tested against the testing corpus.

XBES provides an option of using the minimum between the right branching entropy and left branching entropies or sum of the two.

Evaluation of the segmentations was measured as boundary tagging accuracy, where, in a word, the morpheme boundary location is tagged 1 and everything else 0. Accuracy measures how many boundaries and non-boundaries the segmenter identified correctly.

D. Results

Table 1 shows results from the different morphological segmenters evaluated.

TABLE I. BOUNDARY TAGGING ACCURACY

Method	% Accuracy
Random	60.72
XBES-BE	66.90
XBES-VBE	73.39
XBES-N_tVBE - minimum	67.15
XBES-N_tVBE - sum	69.47
XBES-N_zVBE - minimum	69.72
XBES-N_zVBE - sum	73.32
XHMLS	72.46
Morfessor-Baseline	79.10

[1] The IsiXhosa Genre Classification Corpus and NCHLT IsiXhosa Text Corpus are available at the South African Language Resource Management Agency, (http://rma.nwu.ac.za/index.php)

From Table 1 one notes a minimum accuracy requirement of 60.72% from the random segmenter. This implies that any segmenter below this threshold actively degrades segmentation.

Morfessor-Baseline outperformed other segmenters by a good margin. VBE and N_zVBE–sum seemed to be the best modes for unsmoothed XBES for isiXhosa with accuracy rates of 73.39% and 73.32% respectively.

For Normalised Variation of Branching Entropy modes, the sum of the left and right branching measures performed better than the minimum of the two, implying that a smoothing effect is better, as the sum is a form of averaging the two branching directions.

XHMLS provided very good accuracy (72.46%) from an unsmoothed language model, and showed promise.

VI. CONCLUSIONS

In this paper two possible solutions to the morphological segmentation of isiXhosa were proposed. Both techniques use unsupervised machine learning.

The IsiXhosa Branching Entropy Segmenter (XBES) uses an adaptation of branching entropy techniques to isiXhosa.

The use of novel heuristics that limit the search space in the unsupervised segmentation isiXhosa combined with maximum likelihood estimation was proposed in the isiXhosa Heuristic Maximum Likelihood Segmenter (XHMLS).

Both approaches performed well, considering that their language models still needed to be smoothed, compared to the standard in morphological segmentation, Morfessor-Baseline. XBES's boundary identification accuracy was measured at 73.39% and XHMLS at 72.46% compared to Morfessor-Baseline's 79.1%.

The results show promise in the presented techniques although more work needs to be done.

Going forward the intention is to focus on character language model smoothing, optimise against overfitting and modelling of prefix-suffix morpheme correlations as they are prevalent in isiXhosa.

Acknowledgment

The authors thank the South African Language Resource Management Agency (http://rma.nwu.ac.za/index.php/) for providing a central source of data and resources for Natural Language Processing work.

References

[1] H. J. Groenewald, "Using Technology Transfer to Advance Automatic Lemmatisation for Setswana," in *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 32–37, Athens, Greece, 31 March 2009.*, 2009, pp. 32–37.

[2] A. Sharma Grover, G. B. van Huyssteen, and M. Pretorius, "HLT profile of the official South African languages.," in *2nd AFLaT workshop at the Seventh International Conference on Language Resources and Evaluation (LREC) 2010*, 2010, pp. 3–7.

[3] Statistics South Africa., "Census 2011: Census in brief," 2012.

[4] S. Bosch, L. Pretorius, and A. Fleisch, "Experimental Bootstrapping of Morphological Analysers for Nguni Languages," *Nord. J. African Stud.*, vol. 17, no. 2, pp. 66–88, 2008.

[5] I. M. Kosch, *Topics in Morphology in the African Language*

Context. Pretoria: Unisa Press, 2006.

[6] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, vol. 26, no. 2. MIT Press, 1999.

[7] H. Hammarström and L. Borin, "Unsupervised learning of morphology," *Comput. Linguist.*, vol. 37, no. 2, pp. 309–350, 2011.

[8] F. Zwicky, "Entdecken, erfinden, forschen: im morphologischen Weltbild," *Muenchen: Droemer*, 1966.

[9] J. A. Louw, R. Finlayson, and S. C. Satyo, *Xhosa Guide 3 for XHA100-F*. Pretoria: University of South Africa, 1984.

[10] H. W. Pahl, *IsiXhosa*. King Williams Town: Educum Publishers, 1982.

[11] P. Theron and I. Cloete, "Automatic acquisition of two-level morphological rules," in *Proceedings of the fifth conference on Applied Natural Langauge Processing*, 1997, pp. 103–110.

[12] L. Pretorius and S. E. Bosch, "Finite-State Computational Morphology: An Analyzer Prototype For Zulu," *Mach. Transl.*, vol. 18, no. 3, pp. 195–216, Jul. 2005.

[13] R. Eiselen and M. J. Puttkammer, "Developing Text Resources for Ten South African Languages," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2014, pp. 3698–3703.

[14] M. Nogwina, "Development of a Stemmer for the IsiXhosa Language," University of Fort Hare: MSc. Dissertation, 2016.

[15] H. Déjean, "Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora," in *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Natural Language Learning*, 1998, pp. 295–299.

[16] H. Feng, K. Chen, X. Deng, and W. Zheng, "Accessor Variety Criteria for Chinese Word Extraction," *Comput. Linguist.*, vol. 30, no. 1, pp. 75–93, 2004.

[17] M. Sun, D. Shen, and B. K. Tsou, "Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data," in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, 1998, no. August 10, pp. 1265–1271.

[18] Y. Ye, Q. Wu, Y. Li, K. P. Chow, L. C. K. Hui, and S. M. Yiu, "Unknown Chinese word extraction based on variety of overlapping strings," *Inf. Process. Manag.*, vol. 49, no. 2, pp. 497–512, 2013.

[19] R. K. Ando and L. Lee, "Mostly-Unsupervised Statistical Segmentation of Japanese: Applications to Kanji," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 2000.

[20] J. Rissanen, "Modelling by the shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[21] C. Kit, "A Goodness Measure for Phrase Learning via Compression with the MDL Principle," in *Proceedings of the Third ESSLLI Student Session*, 1998, pp. 175–187.

[22] B. Golénia, S. Spiegler, and P. Flach, "Unsupervised Morpheme Discovery with Ungrade," in *Workshop of the Cross-Language Evaluation Forum for European Languages*, 2009, pp. 633–640.

[23] M. Creutz and K. Lagus, "Unsupervised Discovery of Morphemes," in *Procedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology*, 2002, no. July, pp. 21–30.

[24] E. Gaussier, "Unsupervised learning of derivational morphology from inflectional lexicons," in *Proceedings of ACL'99 Workshop: Unsupervised Learning in Natural Language Processing*, 1999, pp. 24–30.

[25] C. Chavula and H. Suleman, "Morphological Cluster Induction of Bantu Words Using a Weighted Similarity Measure," in *Proceedings of SAICSIT '17,* 2017, no. September 26–28, p. 9.

[26] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J. Am. Soc. Inf. Sci. Sep*, vol. 41, no. 6, pp. 391–407, 1990.

[27] C.-F. Méndez-Cruz, A. Medina-Urrea, and G. Sierra, "Unsupervised morphological segmentation based on affixality measurements," 2016.

[28] M. Belkin and J. Goldsmith, "Using eigenvectors of the bigram graph to infer morpheme identity," in *Proceedings of the ACL-02 workshop on Morphological and phonological learning*, 2002, vol. 6, no. July, pp. 41–47.

[29] P. Schone and D. Jurafsky, "Knowledge-Free Induction of Morphology Using Latent Semantic Analysis," in *Proceedings of CoNLL-2000 and LLL-2000*, 2000, pp. 67–72.

[30] S. Goldwater, T. L. Griffiths, and M. Johnson, "Interpolating Between Types and Tokens by Estimating Power-Law Generators," in *Advances in neural information processing systems*, 2005, pp. 459–466.

[31] K. Uchiumi, H. Tsukahara, and D. Mochihashi, "Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, no. July 26-31, pp. 1774–1782.

[32] K. Sirts and S. Goldwater, "Minimally-Supervised Morphological Segmentation using Adaptor Grammars," *Trans. Assoc. Comput. Linguist.*, vol. 1, pp. 255–266, 2013.

[33] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, 1984.

[34] N. A. Smith and J. Eisner, "Contrastive Estimation: Training Log-Linear Models on Unlabeled Data," in *Proceedings of the 43rd Annual Meeting of the ACL*, 2005, no. June, pp. 354–362.

[35] K. Narasimhan, R. Barzilay, and T. Jaakkola, "An Unsupervised Method for Uncovering Morphological Chains," *Trans. Assoc. Comput. Linguist.*, vol. 3, pp. 157–167, 2015.

[36] P. Magistry, "Unsupervised Word Segmentation and Wordhood Assessment: The case for Mandarin Chineses," Université Paris Diderot, 2013.

[37] Z. S. Harris, "Morpheme Alternants in Linguistic Analysis," *Langauge*, vol. 18, no. 3, pp. 169–180, 1942.

[38] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling," Cambridge, Massachusetts, 1998.

[39] South African Parliament, *UMgaqo-siseko weRiphablikhi yoMzantsi-Afrika ka-1996*. 1996.

[40] D. Snyman, G. B. Van Huyssteen, and W. Daelemans, "Cross-Lingual Genre Classification for Closely Related Languages," in *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, 2012, pp. 132–137.