From phonemes to morphemes: relating linguistic complexity to unsupervised word over-segmentation

Georgia Loukatou
Laboratoire de sciences cognitives et de psycholinguistique, Département d'études cognitives
ENS, EHESS, CNRS, PSL University
georgialoukatou@gmail.com

Abstract

Previous work documented variation in word segmentation performance across languages, with a trend to yield lower scores for languages with elaborate morphological structure. However, segmenting smaller chunks than words, "oversegmenting", is reasonable from a computational point of view. We predict that oversegmentation would be encountered more often in complex languages. In this work in progress, we use a dataset of 9 languages varying in complexity and focus on cognitively-inspired infant word segmentation algorithms. Complexity is defined by Compression-based, Type-Token Ratio and Word Length metrics. Preliminary results show that a possible relation between morphological complexity and oversegmentation cannot be predicted exactly by none of these metrics, but may be best approximated by word length.

1 Introduction

The issue of word segmentation is open in the NLP community (e.g., Harris (1955)). Its implementations include processing languages with no orthographic word boundaries, such as Chinese and Japanese. It is also a key problem humans face when acquiring language.

Previous work documented variation in the success with which languages can be segmented and a trend to yield lower scores for languages with elaborate morphological structure. This is true for both cognitively inspired (Johnson, 2008; Fourtassi et al., 2013; Loukatou et al., 2018)) and other models (Mochihashi et al., 2009; Zhikov et al., 2013; Chen et al., 2011). Evaluation is conventionally done based on orthographic word boundaries. Do these models manage to learn more linguistic structure, that what is described in their accuracy scores?

Segmenting smaller meaningful chunks than words is reasonable from a computational point of view: morphologically complex languages often feature multimorphemic words, and algorithms might break words up into component morphemes, treating frequent morphemes as words. Finding out morphemes might be useful for later linguistic analysis, and such morphemes could be used as cues to further bootstrap segmentation. Thus, a "useful" error in segmentation could be oversegmentation (Gervain and Erra, 2012; Johnson, 2008), the percentage of word tokens returned as two or more subparts in the output.

We predict that oversegmentation might be encountered more often in complex languages. To test this, we need data from languages varying in complexity. However, there is no standard measurement of linguistic complexity. For this study, we use two metrics to define complexity: the Moving Average Type-token Ratio (500-word window) (Kettunen, 2014), and two versions of compression-based complexity (Szmrecsanyi, 2016)[1]. The two metrics are normalized (0=least complex, 1=most complex) and their average score attributed to each language. We also look at word length, since longer words could attract more division.

2 Methods

We use the ACQDIV database (Moran et al., 2016) of typologically diverse languages, with transcriptions of infant-directed and -surrounding speech recordings, from Inuktitut (Allen, 1996), Chintang (Stoll et al., 2015), Turkish (Küntay et al., Unpublished), Yucatec (Pfeiler, 2003), Russian (Stoll and Meyer, 2008), Sesotho (Demuth, 1992), Indone-

[1] 1st metric: the size of compressed corpus (gzip) divided by the size of raw corpus. 2nd metric: systematic distortion of morphological regularities, so as to estimate the role of morphological information in the corpus. Each word type is replaced with a randomly chosen number. The size of the distorted compressed corpus is then divided by the size of the originally compressed corpus.

lang	compr.	MATTR	w length	% over	% corr	% total
Inu	1	0.90	8.56	51	22	73
Chi	0.56	0.87	4.39	44	24	68
Tur	0.44	0.86	4.92	39	26	65
Yuc	0.42	0.92	3.80	31	27	58
Rus	0.41	0.91	4.47	46	19	65
Ses	0.31	0.86	4.28	44	25	69
Ind	0.28	0.85	4.11	42	25	67
Jap	0.14	0.87	3.94	37	25	62
Eng	0.02	0.39	3.04	6	51	57

Table 1: Complexity scores for the three metrics are given in the first 3 columns. Percentage of average oversegmented, correct word tokens and their sum are also given per language.

sian (Gil and Tadmor, 2007) and Japanese (Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2010). In order to compare with a previously studied language, we included the English Bernstein corpus (MacWhinney, 2000).

Several models have been proposed as plausible strategies used by learners retrieving words from input. We used a set of these strategies (Bernard et al., 2018). Two baselines were Base0, treating each sentence as a word, and Base1, treating each phoneme as a word. DiBS[2] (Daland, 2009) implements the idea that unit sequences often spanning phrase boundaries probably span word breaks. FTP[3] (Saksida et al., 2017) measures transitional probabilities between phonemes and cuts depending on a local threshold (relative, FTPr) or a global threshold (absolute, FTPa). Adaptor Grammar (AG) (Johnson, 2008) assumes that learners create a lexicon of minimal, recombinable units and use it to segment the input. AG implements the Pitman-Yor process. Finally, PUDDLE[4] (Monaghan and Christiansen, 2010) is incremental, and learners insert in a lexicon an utterance that cannot be broken down further, and use its entries to find subparts in subsequent utterances. Before segmentation, spaces between words were removed, leaving the input parsed into phonemes, with utterance boundaries preserved.

3 Results

Statistics regarding corpora and results are presented in Table 1. In general, English oversegmented less than other languages, which had similar oversegmentation scores (ranging from 31% to 51%), and did not exactly follow their complexity ranking. Performance difference across languages decreased when considering oversegmented tokens as correct.

4 Discussion

Word length had the best prediction of oversegmentation compared to other metrics (compression and MATTR). This shows that longer words mean more alternative parses, and this might explain oversegmentation results better than other properties inherent to morphologically complex languages. That said, a possible relation between morphological complexity and oversegmentation, could not be *exactly* explained by none of these complexity metrics.

It was also observed that, there was no absolute ranking of complexity across languages; on the contrary, it would change according to the feature studied. In general, cross-linguistic differences were small for this typologically distinct dataset of languages. Further research might shed light on whether this behavior is due to linguistic properties which are common across languages, or a confound (e.g. corpus size).

Moreover, finding meaningful units are of particular importance for language acquisition models, such as the ones implemented here. Infant word segmentation algorithms are plausible only if they are cross-linguistically valid and offer useful insights to learn all linguistic structures. It would also be interesting to compare performance if these models to state-of-the-art NLP algorithms, such as HPYLM (Mochihashi et al., 2009) or ESA (Chen et al., 2011).

Finally, the current implementation of WordSeg does not separate oversegmentation cases resulting in meaningful, morpheme-like sub-parts from other cases. A next step would be to focus on reasonable oversegmentation errors, even though not all of these corpora have morpheme annotations.

Measuring reasonable errors such as oversegmentation could shed light on the segmentability of morphologically complex languages and the cross-linguistic model applicability. Further research might include over-, but also undersegmentation errors, when two or more words in the input returned as a single unit in the output.

[2]Diphone Based Segmentation algorithm

[3]Forward Transitional Probabilities algorithm

[4]Phonotactics from Utterances Determine Distributional Lexical Elements

References

Shanley E. M. Allen. 1996. *Aspects of argument structure acquisition in Inuktitut*. Benjamins, Amsterdam.

Mathieu Bernard, Roland Thiolliere, Amanda Saksida, Georgia Loukatou, Elin Larsen, Mark Johnson, Laia Fibla Reixachs, Emmanuel Dupoux, Robert Daland, Xuan Nga Cao, and Alejandrina Cristia. 2018. Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior research Methods*.

Songjian Chen, Yabo Xu, and Huiyou Chang. 2011. A simple and effective unsupervised word segmentation approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Robert Daland. 2009. *Word segmentation, word recognition, and word learning: A computational model of first language acquisition*. Ph.D. thesis, Northwestern University.

Katherine A. Demuth. 1992. Acquisition of sesotho. In Dan Isaac Slobin, editor, *The crosslinguistic study of language acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates, Hillsdale, NJ.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. WhyisEnglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10.

Judit Gervain and Ramón Guevara Erra. 2012. The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125(2):263–287.

David Gil and Uri Tadmor. 2007. The mpi-eva jakarta child language database. a joint project of the department of linguistics, max planck institute for evolutionary anthropology and the center for language and culture studies, atma jaya catholic university.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

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. Association for Computational Linguistics.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Aylin C. Küntay, Dilara Koçbaş, and Süleyman Sabri Taşçı. Unpublished. Koç university longitudinal language development database on language acquisition of 8 children from 8 to 36 months of age.

Georgia Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2018. Modeling infant segmentation of two morphologically diverse languages. *TALN*.

Brian MacWhinney. 2000. *The CHILDES project: tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Susanne Miyata and Hiro Yuki Nisisawa. 2010. *MiiPro - Tomito Corpus*. Talkbank, Pittsburgh, PA.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(3):545–564.

Steven Moran, Robert Schikowski, D Pajović, Cazim Hysi, and Sabine Stoll. 2016. The ACQDIV database: Min (d) ing the ambient language. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, pages 4423–4429.

Hiro Yuki Nisisawa and Susanne Miyata. 2010. *MiiPro - ArikaM Corpus*. Talkbank, Pittsburgh, PA.

Barbara Pfeiler. 2003. Early acquisition of the verbal complex in yucatec maya. *Development of verb inflection in first language acquisition*, pages 379–399.

Amanda Saksida, Alan Langus, and Marina Nespor. 2017. Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20(3):1–11.

Sabine Stoll, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski, and Balthasar Bickel. 2015. Audiovisual corpus on the acquisition of chintang by six children.

Sabine Stoll and Roland Meyer. 2008. Audio-visional longitudinal corpus on the acquisition of russian by 5 children.

Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, 57:71.

Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. 2013. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. *Information and Media Technologies*, 8(2):514–527.