Cross-linguistic robustness of infant word segmentation algorithms: over-segmenting morphologically complex languages

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

Abstract

Infant word segmentation algorithms are plausible only if they are cross-linguistically valid. They are conventionally evaluated on the word level, tending to yield lower segmentation scores for morphologically complex languages. However, they might segment morphologically complex languages in smaller chunks e.g. morphemes. Such errors could be useful for later linguistic analysis. In this work in progress, a set of algorithms segment corpora from nine typologically diverse languages and oversegmentation errors are counted. The results of this study could inform on the segmentability of morphologically complex languages and the generalizability of segmentation strategies.

1 Introduction

The cross-linguistic application of unsupervised word segmentation strategies is an open issue in the NLP community (e.g., Harris (1955)). Several algorithms have been proposed as plausible strategies used by learners retrieving words from input. Infants, when learning language, need to break down the flow of input speech into word-like units (Saffran et al., 1996). Since they do not know which language(s) will be found in their environment at the beginning of development, they would be better off by using cross-linguistically robust strategies, offering useful insights to learn every linguistic structure.

Previous work assessed the applicability of segmentation algorithms across languages, typically concluding that morphologically complex languages tend to yield lower segmentation results than simpler ones (Johnson, 2008; Fourtassi et al., 2013; Loukatou et al., 2018)). Evaluation was conventionally done based on orthographic word boundaries. However, segmenting smaller meaningful chunks than words, such as morphemes, is reasonable from both a computational and an acquisition point of view: Languages with elaborate morphological structure often feature multimorphemic words, and algorithms might break words up into these component morphemes, treating frequent morphemes as words. Finding out morphemes can also be useful for later linguistic analysis, especially for languages with rich morphological systems (Phillips and Pearl, 2014), and infants seem to recognize functional morphemes of their language early on (Marquis and Shi, 2015), using them as cues to bootstrap segmentation.

lang	% over	% correct	% total
Inuktitut	51	22	73
Chintang	44	24	68
Turkish	39	26	65
Yucatec	31	27	58
Russian	46	19	65
Sesotho	44	25	69
Indonesian	42	25	67
Japanese	37	25	62
English	6	51	57

Table 1: Percentage of average oversegmented, correct word tokens and their sum, per language (by decreasing complexity).

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 cases of oversegmentation would be encountered more often in languages with complex morphology. Also, an algorithm generating reasonable errors like oversegmentation would be more useful for language learning, thus more plausible as a strategy of infant word segmentation. It could also account for previously documented low scores in morphologically complex languages.

algo	% over	% correct	% total
Base0	0	12	12
Base1	87	0	87
DiBS	3	31	34
FTPa	25	33	58
FTPr	41	34	75
AG	47	46	93
PUDDLE	59	31	90

Table 2: Percentage of average oversegmented, correct word tokens per algorithm and their sum. Languages are ordered by decreasing complexity.

2 Methods

We used 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), Indonesian (Gil and Tadmor, 2007) and Japanese (Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2010). In order to compare with a morphologically simple language, we included the English Bernstein corpus (MacWhinney, 2000). We used four metrics to measure morphological complexity: the order of verb synthesis (Stoll and Bickel, 2013), the Moving Average Type-token Ratio (500-word window) (Kettunen, 2014), and two measurements of compression-based complexity (Szmrecsanyi, 2016).[1] Metrics were normalized (0=least complex, 1=most complex) and an average score of the four was attributed to each language. The results, in order of decreasing complexity, are: Inuktitut 1, Chintang 0.56, Turkish 0.44, Yucatec 0.42, Russian 0.41, Sesotho 0.31, Indonesian 0.28, Japanese 0.14, English 0.02.

A set of plausible segmentation strategies was used (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, a stochastic process which reuses frequently occurring rules to build a lexicon. 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 and Discussion

Table 1 gives the average percentage of correct, oversegmented words and their sum for each language. Table 2 shows the average proportion of correct, oversegmented words and their sum for each algorithm. In general, large oversegmentation scores belong to languages with complex morphology (51% for Inuktitut, and 6 % for English). The findings show a possible relation between morphological complexity and oversegmentation, which could not be entirely explained by our complexity metric. The cross-linguistic performance difference ranged from 19% to 51%, but when considering oversegmented words as correct, the difference decreased (57% to 73%). The AG and PUDDLE algorithms segment after building a lexicon, and generate words partially based on their frequency. They largely oversegmented; since they find repeating units, they may have segmented out morphemes instead of words.

Measuring reasonable errors could thus shed light on the segmentability of morphologically complex languages and the cross-linguistic applicability of infant word segmentation strategies. 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.

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

[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*.

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.

Alexandra Marquis and Rushen Shi. 2015. *The Beginning of Morphological Learning: Evidence from Verb Morpheme Processing in Preverbal Infants*, pages 279–295. Springer International Publishing Switzerland.

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

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.

Lawrence Phillips and Lisa Pearl. 2014. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of the Cognitive Science Society*, pages 2775–2780.

Jenny R Saffran, Elissa L Newport, and Richard N Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4):606–621.

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 and Balthasar Bickel. 2013. Capturing diversity in language acquisition research. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson, and Alan Timberlake, editors, *Language typology and historical contingency: studies in honor of Johanna Nichols*, pages 195–260. Benjamins, Amsterdam. [pre-print available at http://www.psycholinguistics.uzh.ch/stoll/publications/stollbickel.sampling2012rev.pdf].

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.