Is word segmentation child's play in all languages?

Georgia R. Loukatou
Laboratoire de Sciences Cognitives et de Psycholinguistique,
Département d'études cognitives, ENS, EHESS, CNRS, PSL University
georgialoukatou@gmail.com

Steven Moran
Damián E. Blasi
Sabine Stoll
University of Zurich

Alejandrina Cristia
Laboratoire de Sciences Cognitives et de Psycholinguistique,
Département d'études cognitives, ENS, EHESS, CNRS, PSL University

Abstract

When learning language, infants need to break down the flow of input speech into minimal word-like units, a process best described as unsupervised bottom-up segmentation. Proposed strategies include several segmentation algorithms, but only cross-linguistically robust algorithms could be plausible candidates for human word learning, since infants have no initial knowledge of the ambient language. We report on the stability in performance of 11 conceptually diverse algorithms on a selection of 8 typologically distinct languages. The results are evidence that some segmentation algorithms are cross-linguistically valid, thus could be considered as potential strategies employed by all infants.

1 Introduction

Six-month-old infants can recognize recurrent words in running speech, even with no meaning available or with experimentally impoverished cues to wordhood (Saffran et al., 1996). Most words do not appear in isolation (Brent and Siskind, 2001), so infants would need to discover the form of words in their caregivers' input before attaching them to meaning. Since infants do not know which language(s) will be found in their environment at the beginning of development, they would be better off by using segmentation strategies that perform above chance for any language. In fact, despite the fact that languages vary widely in a number of dimensions affecting word segmentation, all human languages are learnable for infants (see Discussion for the question of the extent of variation in human learning).

1.1 Unsupervised bottom-up segmentation across languages

The problem of learners retrieving words in input has a long history in computational approaches (e.g., Harris 1955; Elsner et al. 2013; Lee et al. 2015). Most previous computational research has used as input texts representing phonologized language, that is, sequences of phonemes with no overt word boundaries, and the task is to retrieve these. Several algorithms inspired by laboratory research on infant word segmentation are currently represented in WordSeg, an open source package (Bernard et al., 2018).

Are such algorithms as robust to cross-linguistic variation as human infants are? Some previous work has assessed the generalizability of specific approaches across different languages, typically concluding that strong performance differences arise (Johnson 2008; Daland 2009; Gervain and Erra 2012; Fourtassi et al. 2013; Saksida et al. 2017; Loukatou et al. 2018, with the possible exception of Phillips and Pearl 2014a,b).

However, very little previous research compares the performance of a wide range of algorithms using diverse and cognitively plausible segmentation methods within a large set of typologically diverse languages and closely matched corpora, with unified coding criteria for linguistic units.

1.2 The present work

In this paper, we sought to fill this gap by employing a systematic approach that samples both over the space of algorithms and the space of human languages. We used 11 segmentation algorithms included in WordSeg, for improved reproducibility and transparency.

As for languages, we used the ACQDIV

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3931–3937
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

lang	#chi	#sent	#words	m.syn.	%s.com.
Inu	4	13,166	22,045	high	57
Chi	6	160,524	459,585	high	50
Tur	8	249,507	875,349	high	44
Rus	5	468,397	1,302,650	med.	43
Yuc	3	29,795	88,018	med.	51
Ses	4	23,539	62,024	low	55
Ind	10	399,606	1,179,505	low	46
Jap	7	242,774	741,594	low	51

Table 1: Number of children, sentences and word tokens for each language corpus. "m.syn." stands for morphological synthesis derived from sto: A language received a "high" here if nominal and verbal complexity were both listed as the highest in that work; and low if they were both in the lowest levels, and moderate otherwise. " % s.com." stands for syllable complexity, measured as average percentage of vowels per total phonemes for each word. Languages are represented by the first three letters of their names.

database (Moran et al., 2016), which contains a set of typologically diverse languages, as explained in Stoll and Bickel (2013). All corpora were gathered longitudinally and were ecologically valid, with transcriptions of child-directed and child-surrounding speech recordings (target children's age ranges from 6 months to 6 years).

ACQDIV contains data for eight languages with large enough data sets to allow for analyses of the type used here: Chintang (Stoll et al., 2015), Indonesian (Gil and Tadmor, 2007), Inuktitut (Allen, 1996, Unpublished), Japanese (Miyata, 2012b,a; Oshima-Takane et al., 1995; Miyata, 1992) , Russian (Stoll, 2001; Stoll and Meyer, 2008), Sesotho (Demuth, 1992, 2015), Turkish (Küntay et al., Unpublished), and Yucatec Mayan (Pfeiler, Unpublished).

The present study addresses the following questions:

1. **Do algorithms perform above chance level for all languages?** Algorithms that systematically perform at or below chance level would not be plausible strategies for infants.

2. **Is the rank ordering of algorithm performance similar across languages?** That is, is it the case that the same algorithms perform poorly or well across languages? If unsupervised word discovery algorithms pick up on general linguistic properties that are stable across this typologically diverse sample, then we expect the rank ordering to be rather stable. If, conversely, some algorithms pick up on cues that are useful in one language but noxious in another, then the rank ordering may change.

2 Methods

Phonemization was done using grapheme-to-phoneme rewrite rules adapted to each language (Moran and Cysouw, 2018). Only adult-produced speech was included.

The input to each algorithm was the phonemized transcript, with word boundaries removed. Sentence boundaries were preserved because infants are sensitive to them from before 6 months of age (Christophe et al., 2001; Shukla et al., 2011). Table 1 gives the number of children, sentences, and words across corpora, as well as a rough metric of morphological and phonological complexity.

For lack of space, we will only briefly describe the algorithms drawn from WordSeg (see Johnson and Goldwater 2009; Monaghan and Christiansen 2010; Lignos 2012; Daland and Zuraw 2013; Saksida et al. 2017; Bernard et al. 2018). All algorithms were used with their default parameters.

Baseline algorithms represent the simplest segmentation strategies possible. The first baseline, p=0, is a learner who treats each whole sentence as a unit, cutting at 0% of possible points. The second baseline is a learner (innately) informed about average word duration, cutting at a probability level of average word length. Since in the reduced lexicon expected for child-surrounding speech, words average 6 phonemes in length in several languages (Shoemark et al., 2016), p=1/6 was used.

The Diphone Based Segmentation algorithm (DiBS) is based on phonotactics, and implements the idea that phoneme sequences that span phrase boundaries also span word breaks (Daland and Pierrehumbert, 2011; Daland, 2009). The learner decides whether there is a boundary in the middle of a bigram sequence if the probability of the sequence with a word boundary is higher than the probability without the boundary.

Other algorithms are also based on the idea that sequences with lower statistical coherence tend to span word breaks, but use backwards or forwards transitional probabilities (BTP and FTP respectively; in a sequence xy, BTP is the frequency of xy divided by the frequency of y; FTP by the frequency of x) or mutual information (MI). MI is defined as the log base 2 of the frequency of

algo	0	1/6	% mean	% min		% max	
AG	6/8	7/8	37	7	Rus	65	Ind
DiBS	**8/8**	**8/8**	30	25	Jap	41	Inu
FTPa	7/8	8/8	28	17	Inu	36	Ind
MIr	7/8	7/8	27	7	Inu	36	Ind
FTPr	7/8	7/8	25	11	Inu	30	Rus
PUD	6/8	6/8	22	7	Ind	34	Ses
BTPa	6/8	6/8	17	10	Ses	27	Ind
MIa	7/8	8/8	17	15	Jap	25	Inu
BTPr	6/8	5/8	14	9	Inu	22	Yuc
Base0	-	1/8	13	6	Tur	35	Inu
Base6	7/8	-	12	8	Tur	16	Inu

Table 2: Number of languages performing above baseline p=0 and p=1/6. Columns show the mean, the lowest and highest percentage of correctly segmented word tokens for each algorithm and the corresponding language. Languages are represented by the first three letters of their names. "PUD" stands for PUDDLE. "Base0" and "Base6" stand for baseline p=0 and p=1/6.

lang	% mean	% min		% max	
Inuktitut	17	7	MIr	41	DiBS
Chintang	25	9	BTPr	36	AG
Turkish	25	14	PUD	42	AG
Russian	22	7	AG	31	FTPa
Yucatec	27	16	MIa	48	AG
Sesotho	24	9	BTPr	39	AG
Indonesian	29	7	PUD	65	AG
Japanese	26	14	BTPa	43	AG

Table 3: Mean percentage of correctly segmented word tokens for each language. Languages are listed in rough order of morphological complexity (see Table 1). Columns show the mean, lowest and highest percentage of correctly segmented word tokens per language, and the corresponding algorithm. "PUD" stands for PUDDLE.

xy divided by the product of the frequency of x and that of y; the version in WordSeg draws from Saksida's implementation (Saksida et al., 2017). Whether to add a word boundary or not depends on a threshold, which can be based on a local comparison ($relative$, where one cuts if the TP or MI is lower than that for neighboring sequences); or a global comparison ($absolute$, where one cuts if the transition is lower than the average of all TP or MI over the sum of different phoneme bigrams). It should be noted that previous authors originally implemented TPs on syllables (Saksida et al., 2017; Gervain and Erra, 2012), but here the basic units are phonemes. Combining all of the above yields 6 versions, namely FTPr, FTPa, BTPr, BTPa, MIr and MIa.

Johnson and Goldwater (2009) elaborated on adaptor grammars (AG), which are ideal approximations to the segmentation problem. They assume that learners create a lexicon of minimal, recombinable units found in their experience. AG uses the Pitman-Yor process, a stochastic process of probability distribution which prefers the reuse of frequently occurring rules versus creating new ones to build a lexicon, then uses this lexicon to parse the input. This process is conceptually related to Zipf's Law (Zipf, 1935) and leads to realistic word frequency distributions.

Finally, Phonotactics from Utterances Determine Distributional Lexical Elements (PUDDLE) is an incremental alternative algorithm (Monaghan and Christiansen, 2010), where learners build a lexicon by entering every utterance that cannot be broken down further, and using such entries to find subparts in subsequent utterances.

WordSeg was used both for segmentation and evaluation. Each algorithm returns their input with spaces where the system hypothesizes a break.[1] Evaluation is done with reference to orthographic word boundaries. Scripts used for corpus preprocessing and segmentation as well as results and supplementary material are available at https://osf.io/6q5e3/.

3 Results

Results are shown in Tables 2 (reporting on algorithms) and 3 (reporting on languages). Next, we address our research questions.

1. **Do algorithms perform above chance level for all languages?** If chance is defined as the highest of the two baselines (p=0, 1/6), 1 algorithm performed above chance in all 8 languages (DiBS). However, if we relax this criterion, AG, FTPa, FTPr, MIr and MIa also performed above chance for nearly all languages. No algorithm performed below chance level for more than half of the languages.

2. **Is the rank ordering of algorithm performance similar across languages?** Figure 1 illustrates the correlation of performance order for algorithms across languages. Spearman correlations (median=.38) suggested that there is a similar rank ordering

[1] Because of time constraints, only the first 50000 utterances of the three largest corpora, Turkish, Russian and Indonesian, were segmented by AG. This would play a negligible role in results, since variation in corpus size beyond the first 5k utterances does not affect performance of this segmentation system (Bernard et al., 2018).

of algorithm performance across languages. Inuktitut and Russian were the only languages not following the general ordering.

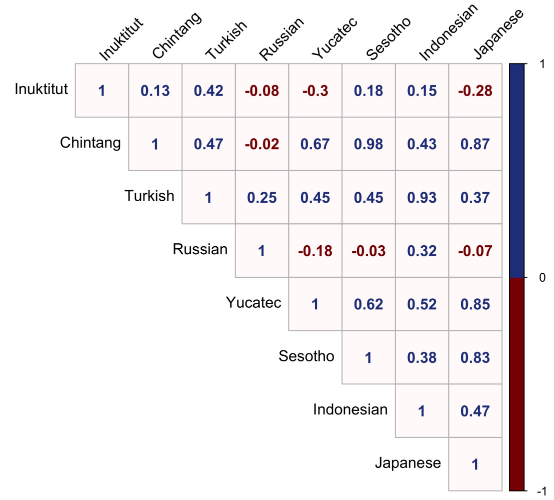
The models' detailed performance, measured in percentage of correctly segmented word tokens, can be found in the online supplementary material and in this paper's Appendix. An error analysis would be beyond the scope of this paper. However, three categories of incorrect cases have been measured and can be found online. This analysis documents cases of oversegmentation (words split up in their components), undersegmentation (two or more words segmented as one) and missegmentation (all other errors).

4 Discussion

First, no algorithm performed systematically below chance level in our study. However, we cannot say that they all performed above chance for all languages either. This is mainly due to the good results in baseline p=0, especially salient for morphologically complex languages such as Inuktitut. This is expected, since in this language a substantial number of sentences are composed by a single word (which morphologically encodes what in other languages would be expressed syntactically by using several words).

Second, there was some stability in the order of performance for algorithms across this set of diverse languages, suggesting that these unsupervised word discovery algorithms pick up on general linguistic properties that are stable across our sample, and not language-dependent cues that could potentially not work for some languages.

In this distinct performance ranking, some algorithms were systematically above chance and among the first in order of performance. These include DiBS and AG, combining both desiderata of cross-linguistic stability and high segmentation performance. DiBS, the one algorithm in our sample applying a phonotactics strategy, was robust across languages and not strongly affected by the differences found across these languages in morphology and phonological complexity (counter previous conclusions based on English versus Korean, Daland and Zuraw 2013). DiBS implements an optimal boundary setting based on the Bayes' theorem and co-occurrence statistics. Thus, our results support previous experimental findings that infants may use such tools to acquire language.



	Inuktitut	Chintang	Turkish	Russian	Yucatec	Sesotho	Indonesian	Japanese
Inuktitut	1	0.13	0.42	-0.08	-0.3	0.18	0.15	-0.28
Chintang		1	0.47	-0.02	0.67	0.98	0.43	0.87
Turkish			1	0.25	0.45	0.45	0.93	0.37
Russian				1	-0.18	-0.03	0.32	-0.07
Yucatec					1	0.62	0.52	0.85
Sesotho						1	0.38	0.83
Indonesian							1	0.47
Japanese								1

Figure 1: Correlation matrix of the rank ordering in algorithms' performance across languages.

Our study is the first to explore segmentation differences across both multiple algorithms and multiple languages. We therefore are in a position to compare segmentation performance differences across these two. We found that differences in average performance across algorithms (min=14 for BTPr, max= 37 for AG, 23% points) were larger than differences in performance across languages (min=17 for Inuktitut, max=24 for Indonesian, 7% points). This indicates that variation across languages was comparatively small.

Also, average percentage of correctly segmented words for the more morphologically complex languages (Chintang, Inuktitut and Turkish) was 19%, only 3% lower than average percentage for the simpler languages in our sample (Japanese, Sesotho and Indonesian). This is striking evidence that in this set of diverse languages, intrinsic differences in language structure may not be large enough to create particular difficulties in segmentation.

To sum up, this study provides evidence that, if infants do anything similar to one or more of the algorithms proposed in previous natural language processing research and investigated here, then they would be well-equipped to get a head start in segmenting word-like units regardless of what their native language is. Experimental evidence suggests slight variation in the timing of acquisition of different linguistic features, as a function of factors such as the transparency of forms, and the complexity of paradigms (e.g., Slobin 1985). Given the small differences found across our unsupervised word segmentation algorithms, such variation might come from something else, such as meaning acquisition, which would require algorithms different from the ones we explored here.

Before closing, we would like to acknowledge some limitations of this work. Defining words can be obscure (Daland, 2009) and there is no cross-linguistically valid general definition of 'word' (Haspelmath, 2011). Consequently, it would make sense to also evaluate unsupervised segmentation algorithms using morpheme edges and at other definitions of wordhood (Bickel and Zúñiga, 2018). For this, we would need appropriately annotated data sets, which are currently missing. What is worse, not every language lends itself to simple definitions: Some languages in ACQDIV lack morpheme segmentation simply because this is not feasible in that language.

In this paper, we focus on correctly segmented words. An error analysis would not be easily interpretable, because not all corpora have morpheme annotations. For example, when documenting oversegmentation errors, we would not be able to distinguish between reasonable cases where words are split up into meaningful, morpheme-like components, and other cases. Similarly, in an undersegmentation analysis, we would not be able to focus on collocations. Future work is invited to study in more detail such errors in the algorithms' performance.

Finally, computational models can be informative proofs of principle, but nothing assures us they truly represent what infants are doing. To this end, laboratory experiments (Johnson and Jusczyk, 2001) and the study of natural variation (Slobin, 1985) are irreplaceable, even if challenging to perform, particularly at a large scale and sampling from many different cultures.

Acknowledgments

AC acknowledges funding from the Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. This research was enabled by the European Union's Seventh Framework Program (FP7/2007–2013) under grant agreement 615988 (PI Sabine Stoll).

References

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

Shanley Allen. Unpublished. Allen Inuktitut Child Language Corpus.

Mathieu Bernard, Roland Thiolliere, Amanda Saksida, Georgia R. 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*.

Balthasar Bickel and Fernando Zúñiga. 2018. The 'word'in polysynthetic languages: Phonological and syntactic challenges. In Michael Fortescue, Marianne Mithun, and Nicholas Evans, editors, *The Oxford Handbook of Polysynthesis*, 158-186. Oxford University Press.

Michael R Brent and Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

Anne Christophe, Jacques Mehler, and Núria Sebastián-Gallés. 2001. Perception of prosodic boundary correlates by newborn infants. *Infancy*, 2(3):385–394.

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

Robert Daland and Janet B Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Robert Daland and Kie Zuraw. 2013. Does Korean defeat phonotactic word segmentation? In *Association for Computational Linguistics*, pages 873–877.

Katherine Demuth. 2015. Demuth Sesotho Corpus.

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.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 42–54.

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.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Elizabeth K Johnson and Peter W Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44(4):548–567.

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.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–325. Association for Computational Linguistics.

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.

Chia-ying Lee, Timothy J O'donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

Constantine Lignos. 2012. Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics*, volume 30, pages 13–15.

Georgia R. Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2018. Modeling infant segmentation of two morphologically diverse languages. In *Proceedings of TALN*, pages 47–60.

Susanne Miyata. 1992. Wh-questions of the third kind; The strange use of wa-question in Japanese children. *Bulletin of Aichi Shukutoku Junior College*, 31:151–155.

Susanne Miyata. 2012a. CHILDES nihongoban: Nihongoyoo CHILDES manyuaru 2012. [Japanese CHILDES: The 2012 CHILDES manual for Japanese.

Susanne Miyata. 2012b. Nihongo MLU (heikin hatsuwachō) no gaidorain: Jiritsugo MLU oyobi keitaiso MLU no keisanhō [Guideline for Japanese MLU: How to compute MLUw and MLUm]. Kenkō Iryō Kagaku 2, 1–15.

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 and Michael Cysouw. 2018. *The Unicode cookbook for linguists: Managing writing systems using orthography profiles (Translation and Multilingual Natural Language Processing 10)*. Berlin: Language Science Press.

Steven Moran, Robert Schikowski, Danica 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)*, Paris, France. European Language Resources Association (ELRA).

Yuriko Oshima-Takane, Brian MacWhinney, Hidetoshi Shirai, Susanne Miyata, and Norio Naka. 1995. CHILDES manual for Japanese. *Montreal: McGill University*.

Barbara Pfeiler. Unpublished. Pfeiler Yucatec Child Language Corpus.

Lawrence Phillips and Lisa Pearl. 2014a. Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the Computational and Cognitive models of Language Acquisition and Language Processing Workshop*.

Lawrence Phillips and Lisa Pearl. 2014b. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of the 36th annual conference of the Cognitive Science Society*, pages 2775–2780, Quebec City, CA. Cognitive Science Society.

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.

Philippa Shoemark, Sharon Goldwater, James Kirby, and Rik Sarkar. 2016. Towards robust cross-linguistic comparisons of phonological networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–120.

Mohinish Shukla, Katherine S White, and Richard N Aslin. 2011. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, 108(15):6038–6043.

Dan Isaac Slobin. 1985. *The crosslinguistic study of language acquisition: Theoretical Issues*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Sabine Stoll. 2001. *The acquisition of Russian aspect*. Ph.D. thesis, University of California, Berkeley.

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. Audiovisional longitudinal corpus on the acquisition of Russian by 5 children.

George Kingsley Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin, Oxford, England.

Appendix

The models' performance, measured in percentage of correctly segmented word tokens, can be found in Table 4.

algo	Inu	Chi	Tur	Rus	Yuc	Ses	Ind	Jap
AG	20	36	42	7	48	39	65	43
DiBS	41	29	33	26	28	28	30	25
FTPa	17	30	30	31	22	30	36	29
MIr	7	29	29	30	33	25	36	30
FTPr	11	28	27	30	25	25	28	29
PUD	8	33	14	19	31	34	7	33
BTPa	14	12	19	23	20	10	27	14
MIa	25	16	15	21	16	17	16	15
BTPr	9	9	17	15	22	9	17	16
Base0	35	9	6	12	8	11	9	12
Base6	16	11	8	12	11	12	11	13

Table 4: Percentage of correctly segmented word tokens for each language and algorithm. Languages are listed in rough order of morphological complexity (see Table 1). "PUD" stands for PUDDLE. "Base0" and "Base6" stand for baseline p=0 and p=1/6. Languages are represented by the first three letters of their names.