Modeling infant segmentation of two morphologically diverse languages

Georgia Rengina Loukatou[1] Sabine Stoll [2] Damian Blasi[2] Alejandrina Cristia[1]
(1) LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University, Paris, France
(2) University of Zurich, Zurich, Switzerland
georgia.loukatou@ens.fr

RÉSUMÉ

Les nourrissons doivent trouver des limites de mots dans le flux continu de la parole. De nombreuses études computationnelles étudient de tels mécanismes. Cependant, la majorité d'entre elles se sont concentrées sur l'anglais, une langue morphologiquement simple et qui rend la tâche de segmentation aisée. Les langues polysynthétiques - pour lesquelles chaque mot est composé de plusieurs morphèmes - peuvent présenter des difficultés supplémentaires lors de la segmentation. De plus, le mot est considéré comme la cible de la segmentation, mais il est possible que les nourrissons segmentent des morphèmes et non pas des mots. Notre étude se concentre sur deux langues ayant des structures morphologiques différentes, le chintang et le japonais. Trois algorithmes de segmentation conceptuellement variés sont évalués sur des représentations de mots et de morphèmes. L'évaluation de ces algorithmes nous mène à tirer plusieurs conclusions. Le modèle lexical est le plus performant, notamment lorsqu'on considère les morphèmes et non pas les mots. De plus, en faisant varier leur évaluation en fonction de la langue, le japonais nous apporte de meilleurs résultats.

ABSTRACT

A rich literature explores unsupervised segmentation algorithms infants could use to parse their input, mainly focusing on English, an analytic language where word, morpheme, and syllable boundaries often coincide. Synthetic languages, where words are multi-morphemic, may present unique difficulties for segmentation. Our study tests corpora of two languages selected to differ in the extent of complexity of their morphological structure, Chintang and Japanese. We use three conceptually diverse word segmentation algorithms and we evaluate them on both word- and morpheme-level representations. As predicted, results for the simpler Japanese are better than those for the more complex Chintang. However, the difference is small compared to the effect of the algorithm (with the lexical algorithm outperforming sub-lexical ones) and the level (scores were lower when evaluating on words versus morphemes). There are also important interactions between language, model, and evaluation level, which ought to be considered in future work.

MOTS-CLÉS : variation interlinguistique, apprentissage statistique, segmentation des mots, acquisition du langage.

KEYWORDS: cross-linguistic variation, statistical learning, word segmentation, language acquisition.

1 Introduction

Human infants are known to acquire a comprehension vocabulary of hundreds of words by two years of age (Hoff, 2013), and probably accumulate a protolexicon consisting solely of word forms, with no meaning attached, by the end of the first year (Ngon *et al.*, 2013). Infants' discovery of basic units in their input has been modeled as follows. A transcript of infant-directed speech is converted into phonological text, word boundaries are removed, and an algorithm is applied. Then, the word boundaries posited by the algorithm (and the resulting word tokens) are compared against the adult segmentation found in the original corpus.

By and large, two classes of algorithms have been heavily studied (Daland, 2009; Jarosz & Johnson, 2013). Algorithms in the *lexical* class are built to find the most economical system of minimal units needed to reproduce the input. Those in the *sub-lexical* class aim to find local cues allowing the learner to posit boundaries, for instance detectable via a dip in transitional probabilities. We will discuss both in more detail below, but for now it suffices to say that both are plausible given known infant experimental data (Mersad & Nazzi, 2012; Saffran *et al.*, 1996a).

The present study represents the first systematic attempt to apply both types of unsupervised word form discovery techniques to two morphologically diverse languages. In the next section, we summarize previous work with the lens of morphological distinctions.

1.1 Morphological variability predicts performance in previous modeling work

Most previous work modeling infant segmentation has focused on English (e.g. Lignos 2011; Venkataraman 2001; Christiansen & Curtin 2005; Phillips & Pearl 2015; Monaghan & Christiansen 2010); yet this is not an "average" language for segmentation. English words are mostly monomorphemic and monosyllabic, such that word, morpheme, and syllable boundaries usually coincide (DeKeyser, 2005).

Indeed, morphologically speaking, English can be classified as an analytic language, because most words have few or no morphemes other than the root. Synthetic languages are characterized by having richer inflectional morphology. They use morphemes such as prefixes, suffixes and infixes to convey certain features (e.g., gender) and/or the relation between words in a sentence (e.g., via case).

Languages can vary greatly in the degree of synthesis. Some languages, such as Hungarian and Tamil, are synthetic to a high degree, with a rich inflectional morphology in both nouns and verbs. Others are more intermediate, including many IndoEuropean languages, such as Italian, Spanish, and French, which have only a few suffixes in verbs and nouns. A distinction among synthetic languages that we will not study but is worth mentioning is that between agglutinative and fusional languages. In the former, morphemes are transparent and concatenated, whereas in fusional languages a single morpheme contains many features.

Languages with a rich morphology are of particular interest in the context of segmentation. Since complex words are formed by the combination of many easily separable morphemes, lexical algorithms could break words up into the component morphemes (Batchelder, 2002). Additionally, highly synthetic languages usually have longer words, and may have longer utterances (in number of phonemes or syllables). Longer utterances mean more alternative parses can be posited, and thus more

uncertainty particularly (but not only) for lexical algorithms (Fourtassi *et al.*, 2013). Moreover, lexical algorithms often implement a drive for economy, whereby reuse of minimal units is preferred over postulation of additional lexical units. This could specifically lead to problems for languages where, by virtue of inflectional morphology, a lexeme has many surface forms, each used less frequently, and where it may be more economical to break up the word into roots and affixes, which can be more efficiently re-used.

Finally, corpora of such languages could contain fewer repetitions of each word token (since each lexeme can have different surface forms) and thus a higher proportion of hapaxes than analytic languages, which might affect performance in lexical algorithms where the probability of generating a word is partially a function of its frequency. All the above predict better performance for less than more synthetic languages, and for this difference to be more marked in the performance of lexical than sublexical algorithms.

Overall, previous results support our main predictions, with synthetic languages yielding lower segmentation performance than analytic languages. Fourtassi *et al.* (2013) applied a probabilistic lexicon-building algorithm on English and Japanese. English (analytic) yielded a Token F-score of 0.77 and Japanese (synthetic-agglutinative) 0.69. Qualitative inspection suggested to the authors that the algorithm broke apart morphological affixes, generating more oversegmentation errors for Japanese than English, which fits the reasoning laid out above well.

This effect was replicated by Boruta *et al.* (2011) with another lexicon-based model, documenting better results for English than French (synthetic-fusional), and for French than Japanese. The author reported a higher proportion of hapax words in Japanese than for English, and a lower likelihood of correct identification by the algorithm for hapaxes than words with more than one repetition. Finally, results were dismal for Sesotho, another highly synthetic language characterized by even more complex morphology than Japanese (Johnson, 2008).

Although the arguments above are most relevant to lexical algorithms, previous work using sub-lexical ones also confirms the hypothesized trend. A diphone-based segmentation model developed by Daland performed lower for the morphologically complex Russian than English. For their part, Saksida *et al.* (2017) used a set of segmentation models based on transitional probabilities on a range of corpora. English had a maximum score of 0.85, whereas Japanese, Tamil, and Hungarian, all synthetic, a maximum of 0.75.

As in the lexical literature, other work even suggests differences among synthetic languages. In Gervain & Erra (2012), better results were found for the less complex Italian than the more complex Hungarian. The authors commented that there may be more oversegmentation in the latter language, as some of the erroneously segmented words were real morphemes, which is interesting given that this unsupervised algorithm has not been designed to be sensitive to lexical and morphological composition.

2 The present study

The key question motivating this study is whether languages that vary in morphological complexity differ in segmentability. To answer it, we looked for languages that were morphologically diverse, but for which there were closely matched and comparable corpora. Previous authors have often argued that lower performance for highly synthetic languages arises from oversegmentation, mainly based on

Language	Verb agr.	Split erg.	Compactness	Syncr.	V syn.	N syn.
Japanese	none	low	cumulative	none	low	1
Chintang	some	medium	distributive	some	high	3

TABLE 1 – Differences between the two languages according to (Bickel *et al.*, 2013). Verb Agr(eement), Split Ergativity of Case (proportion of ergative case alignments), Compactness, (Prevalence of) Syncr(etism), V(erbal) Syn(thesis), N(ominal) Syn(thesis).

qualitative inspection of results (but see Johnson 2008). To assess this question more systematically, we inquired whether performance varied as a function of the level of linguistic representation on which segmentation is evaluated. [1]

Our goal was not to test an exhaustive list of languages, which was not feasible at present. Instead, we opted to compare two language corpora which are part of the same database and have been transcribed using the same guidelines. Additional desiderata included that the morphological difference of these languages should be large and computationally assessed, and that other linguistic parameters such as syllable structure should be in similar levels.

All of these considerations led us to the ACQDIV database of linguistically diverse languages (Schikowski *et al.*, 2015). Languages in this database have been sorted based on clustering algorithms and according to several linguistic, typological variables. It may be worth pointing out that we did not test our models on English here, as previous literature has extensively presented results on various English corpora (some if it has been summarized above), and English did not exist in the chosen database.

Thus, to best complement previous work, we selected two morphologically diverse non-IndoEuropean languages present in the AcqDiv database, namely Japanese and Chintang (Bickel *et al.*, 2013). [2] Even though the languages belong to the same morphological category (synthetic agglutinative), they are very different in the degree of complexity.

Chintang is a polysynthetic language (i.e., having an extremely high ratio of morphemes per word), and thus has a more complex morphology than Japanese. It has higher verb and noun synthesis (number of categories expressed, word complexity – compare Paudyal 2015 for Chintang, and Kuno 1973; Tsujimura 2013 for Japanese), with up to 10 morphemes per word in Chintang. The languages also differ in that Chintang has distributive inflectional compactness (categories are expressed separately in distinct morphemes), whereas Japanese has cumulative compactness (grammatical categories are expressed cumulatively in fusioned affixes), which denotes the need for less morphemes than for Chintang, as we can see in Table 1. As per our desideratum above, the phonological complexity (phonemic inventory and syllabic structure) is similar across the two languages.

Most previous work used a single class of algorithms (but see (Ludusan *et al.*, 2017)). Although previous literature suggests that segmentation performance varies as a function of morphological type for both lexical and prelexical algorithms, our reasoning above predicts that effects of morphological complexity on segmentation should be stronger for lexical than prelexical algorithms. We therefore included algorithms of both types.

1. Note that it is not unreasonable to propose that infants segment morphemes, rather than words (Phillips & Pearl, 2014).
2. Chintang is a language of the Kiranti subgroup of the Sino-Tibetan language family spoken in Eastern Nepal by about 6,000 speakers.

3 Methods

The Chintang recordings took place 4h (cumulated during several sessions carried out within a week) per month, over 18 months, and involved 7 children aged between 6 months and 4 years and 4 months of age (Stoll *et al.*, 2016). For Japanese, recordings of 7 children aged 1 year 4 months, to 5 years 1 month, each lasting 40-70 minutes, were collected between once per week and once per month.

All child-directed and child-overheard speech had been carefully transcribed in a transparent orthography (as was the child's own speech, which was not analyzed here). We applied grapheme-to-phoneme rules to derive the phonological representation of utterances. After processing, the Chintang corpus contained 296,939 utterances, with an average of 2.7 words, 5.4 syllables, and 11.4 phones per utterance ; and the Japanese corpus 264,945 utterances with an average of 3 words, 5.7 syllables, 11.2 phones per utterance). All utterances where morpheme annotation was incomplete were removed, so the Japanese morpheme corpus after processing contained 85267 utterances with 2 morphemes, 3.2 syllables, 6.3 phones per utterance and the Chintang morpheme corpus contained 280319 utterances with 4.5 morphemes, 5.5 syllables, 11.4 phones per utterance.

To perform inferential statistics, each corpus was divided in ten equal subparts based on number of utterances. Within-sentence word boundaries were removed and fed to three models, which varied on the cognitive strategies applied, as follows. [3]

3.1 DiBS

For the sub-lexical Diphone Based Segmentation model (DiBS) (Daland, 2009; Daland & Pierrehumbert, 2011), segmentation decisions are based on the basis of diphone probabilities. A probability of word boundary ranging from 0 up to 1 is assigned to each diphone found within every utterance. For the present work, we are using one of the unsupervised version of DiBS called phrasal DiBS. [4] In this version, the algorithm treats phrase edges as a proxy for word edges given the unquestionable assumption that diphones spanning a phrase boundary are also spanning word breaks. It estimates two parameters from the corpus, to be fed into the formula 1.

$$p(\#|xy) = \frac{f(\# \wedge xy)}{f(xy)} \tag{1}$$

where $f(\# \wedge xy)$ is the number of [xy] sequences with a phrase boundary in the middle, and $f(xy)$ is the the number of [xy] sequences in any position (Daland, 2009). When this ratio is higher than a parameter called "probability of word boundary", then the system will make a hard decision that there is indeed a break. The probability of word boundary is calculated using Formula 2.

$$\frac{Nw - Nu}{Np - Nu} \tag{2}$$

where Nw stands for number of words, Nu number of utterances, and Np number of phones.

3. We actually used the defaults in the WordSeg package (Bernard *et al.*, submitted). For more information, visit https ://wordseg.readthedocs.io/en/latest/algorithms.html

4. Other versions of DiBS, not used for this paper, are Baseline DiBS (where the boundary in $f(\# \wedge xy$ is the true word boundary, and thus is fully supervised), and Lexical DiBS (which requires a small vocabulary to be provided by the programmer, and thus necessitates additional assumptions as to how the child learned $those$ words).

Notice that while Formula 1 requires nothing but a phone inventory and the phrase boundary location, Formula 2 makes reference to the gold number of words, and thus can be said to be supervised. Daland argues convincingly that this can be viewed as a convenient shortcut, rather than a design flaw in the algorithm. Indeed, we can imagine children being born with a parameter akin to the minimal word length requirement (McCarthy & Prince, 1986), or deducing them from other aspects of the language (length of the shortest utterances, distance between stressed syllables, etc.)

In its original implementation, this model required that the input be coded in phonemes. To allow direct comparison with previous work, we did not alter this requirement.

3.2 TP

The Transitional Probabilities (TP) family builds on the assumption that the transitional probability between adjacent syllables is lower at word boundaries than at word middles (Gervain & Erra, 2012; Saffran *et al.*, 1996b; Saksida *et al.*, 2017). Forward TP (FTP) is defined as

$$FTP(AB) = \frac{f(AB)}{f(A)} \tag{3}$$

where $f(AB)$ is the frequency of occurrence of the syllabic sequence AB and $f(A)$ is the frequency of occurrence of the syllable A.

Backward TPs (BTP) instead divides the product by the times the second syllable appears. $P(B)$ is the probability of occurrence of the syllable B.

$$BTP(AB) = \frac{f(AB)}{f(B)} \tag{4}$$

Notice that these formulas simply provide an indirect estimate of how likely B is given A (or B, in Formula 4) and thus are not sufficient to posit a boundary. In fact, the decision on whether to posit a boundary between two syllables can be taken with at least two different methods. The Absolute TP method (TPa) uses the average of the TPs over the sum of bigrams for the whole corpus as a threshold. We can have either Backward TPa (BTPa) and Forward TPa (FTPa).

The Relative TP (TPr) cuts words when the TP value of a bigram is weaker than the TP of the neighboring ones, so we have Backward TPr (BTPr) and Forward TPr (FTPr). For example, if AB is a bigram in a sequence of XABY, then a boundary would be posited if Equation 5 is true.

$$TP(XA) > TP(AB) < TP(BY) \tag{5}$$

This model's input is coded into syllables, so we syllabified the corpora using the Maximal Onset Principle, according to which a syllable's onset should be extended as much as possible, as long as it stays phonotactically legal (Bartlett *et al.*, 2009).

All previous work has analyzed corpora unitized at the level of the syllable, and not of phonemes, as input to this family of algorithm. Although a TP version with phonemes as basic units, rather than syllables, is cognitively possible, we preferred the latter in order to compare our results against others' using TP as it had been used in previous studies. Reviewers indicated this results in a loss

of comparability across models, given that the other two models are based on phonemes. If a reader feels the same way at this point, we would like to underline here that our research goal was neither to provide an exhaustive mapping of all models nor even to carry out specific comparisons across the models. Our main research question, as stated above, concerns the effects of $morphological$ differences across languages. The use of varied models (defined as the conjunction of input and processing decisions) allows us to assess whether patterns are observed $despitemodelvariation$, or whether certain patterns may be only obvious for subsets of models.

3.3 AGu

The third model, a member of the Adaptor Grammar (AG) family, adopts a lexical approach (Goldwater *et al.*, 2009; Johnson & Demuth, 2010), meaning that it tries to find patterns of sequence of units that repeat in the input and uses that lexicon to parse the input. It is technically a great deal more complex than either of the models just discussed. For reasons of space, we provide a mainly verbal explanation, and refer readers to Johnson *et al.* (2007) for a technical introduction to adaptor grammars, including mathematical formulae and general properties.

In a nutshell, we use here a Pitman-Yor Adaptor Grammar, which is a generalized verion of probabilistic context-free grammars (PCGF) (Johnson *et al.*, 2007). In context-free grammars, corpora are generated as a function of the repeated application of a set of rewrite rules, which, in the process of parsing/generating a corpus, are selected independently and at random. Contrastingly, in PCGFs, each rule is assigned a probability, and the probability of a given parse is the product of the probabilities of the rules that may have been invoked to generate the input text.

Briefly, a given Adaptor Grammar is described as a function of (N,T,R,θ,a/b) N is a finite set of nonterminal symbols (in our grammar below, Sentence, Word, Phoneme), T is a finite set of terminal symbols (in our grammar below, the actual phonemes of the language), R is a set of rewrite rules, θ is a probability distribution over the different rules, and a/b are concentration parameters governing re-use versus generation (as explained below).

For this study, the input was represented using phonemes, and we used a basic version of adaptor grammar assuming no dependencies between words (unigram). For this reason, we will refer to our implementation of the model in what follows as AGu. This is the simplest adaptor grammar one can imagine, and it presupposes only two assumptions – (a) Sentences are composed of one or more reusable words, and (b) Words are composed of one or more basic terminal units (phonemes). These assumptions are encoded into rules enumerated below.

1. Sentence $\rightarrow$ Word (Word)
2. Word* $\rightarrow$ Phoneme (Phoneme)
3. Phoneme $\rightarrow$ a
4. Phoneme $\rightarrow$ ...

The items on the left are non-terminals and those on the right may be lower-level non-terminals (as in rewrite rules 1-2) or terminals (as in the remaining rewrite rules). Items between parentheses are optional. Items with an asterisk can be generated *d*e novo, or they can be added as rewrite rules during parsing, and subsequently re-used throughout the corpus. Notice additionally that these rewrite rules are equivalent to creating trees – for instance, the word "see" may be parsed as the application of rule (2) followed by the rules expanding the Phoneme non-terminal into the terminal phonemes /s/, /i/.

As just mentioned θ is a vector corresponding to each and every rule in R, a number that represents the probability of expanding the non-terminal on the left hand side of the given rule into possible terminal(s) or non-terminal(s) on the right hand side. Let us imagine a parse where the system finds a sequence A, which is a non-terminal. The Gaussian distribution G_A corresponds to the set of trees associated with the non-terminal A. For example, /sit/ could in theory be parsed as /s i/ or /si/ – G_A would establish that there is a probability p that corresponds to the tree resulting in the parse /s i/ with and $1 - p$ to the tree resulting in the parse /si/.

Recall that the system can parse the input using just the list of original rules, in which case the generation of an utterance results from the repeated application of only rules 1-2 and the terminal phonemes in the utterance. It can also, however, create new rules to shortcut this process. In the implementation we use for this paper, we allowed the creation of rules as sequences of phonemes, which effectively means we allowed the creation of a lexicon or morphological inventory, which can be used to segment utterances into a sequence of words or morphemes (without using the 'Phoneme $\rightarrow$ terminal' rules). In the version we are using (Johnson *et al.*, 2007), probabilities for re-use versus regeneration are based on the Pitman-Yor Process, a stochastic process of probability distribution which pits the reuse of frequently occurring trees (or rules) versus creating new trees (or rules). [5]

Specifically, this process is governed by the "concentration" set of parameters a and b, which determine whether generated rules are costly and thus whether the system should be reusing rules or create many new rules. For the sake of comparability with previous work, we used the values that had been preferred for experiments on English, French and Japanese adult and child corpora at the time the package was first used (e.g., Fourtassi *et al.* 2013), namely a=0.0001 and b=10000. To put this in context, initially, the lexicon is empty (i.e., no shortcut rules have been created), so the first utterance will be parsed as a sequence of words, each composed by phonemes, with this whole probability distribution governed by G_A. For simplicity, assume the system posits a single word – w_1. The second time the same sequence of terminals is found, the system can extract this word from the adaptor's lexicon (and hence it is w_1), generate another rule with the same non-terminal, or generate the word from scratch using G_A. This whole process is repeated in several runs of 2000 times. Finally, Minimum Bayes Risk is used to find the most common sample segmentations (Johnson *et al.*, 2007).

4 Results

Recall that our main research question is whether there are differences in segmentability across languages as a function of morphological complexity, particularly as the more complex language (Chintang) may be oversegmented when the output is evaluated at the level of words (rather than morphemes) and when using a lexical algorithm (AGu, as opposed to the sublexical DiBS and TPs). To answer this question, we fit a regression model to token F-scores data declaring language, level, algorithm and their interactions as independent predictors, and corpus as repeated variable (given that the corpora had been cut into 10 subparts so as to be able to carry out inferential statistics). This regression accounted for most of the variance in the data, $R^2 = .90$ ($F(23, 216) = 98.5$, $p < .001$).

As predicted, the coefficient estimating the language effect between Chintang and Japanese is positive (0.069), suggesting higher scores for the latter. Interestingly, this effect is smaller than that of level (morpheme versus word, -0.11). The presence of all 2- and 3-way interactions, however, discourages

5. It is often described as the Chinese restaurant process, where new customers can be seated at a new table or an extant table, and in the latter case they will tend to be placed in tables with many customers than tables with a few.

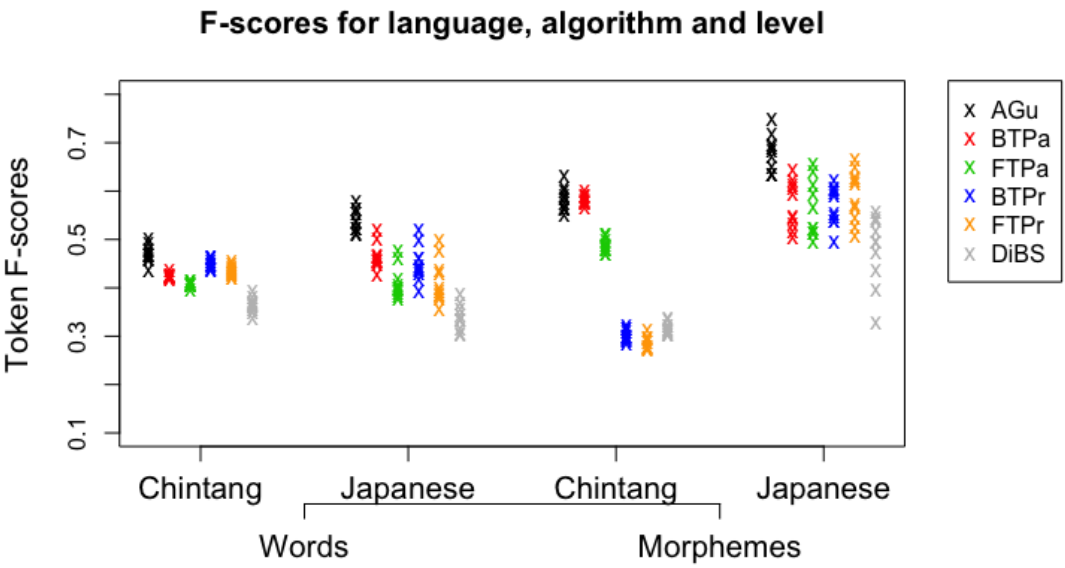

FIGURE 1 – Token F-scores across language and representational level. Models are marked by color. Each "x" represents one of the ten subparts of each corpus.

a simple reading of these main effects.

As clear in Figure 1 there were strong interactions between all three factors (language, level, model).

5 Discussion

When combined with results from the previous literature, we confirm the generalization that complex languages (such as the two studied here) lead to lower segmentation scores than morphologically impoverished languages (such as English). Indeed, for all algorithms, we obtained lower scores for Chintang and Japanese than English results previously documented. We focus on the word level for this comparison, since previous work has systematically evaluated performance on this level, and not the morpheme level. For AGu, we retrieve an average Token F-score of .54 for Japanese and .47 for Chintang, whose performance is close to that for Sesotho (Johnson, 2008), another morphologically complex language with high degrees of synthesis. Both are much lower than the .77 Fourtassi *et al.* (2013) previously documented for English.

Our maximum Token F-score for TP was .66, below the .85 recorded for English and even the .75 recorded for agglutinative languages by Saksida *et al.* (2017). Finally, the score of .4 for DiBS is similar to those recorded for Russian and much lower than the score registered for English (Daland, 2009).

How about the comparison between the two agglutinative languages included here ? Our regression documented an advantage for the less complex Japanese over the more complex Chintang. Although

results thus far confirm the prediction that the *degree* of synthesis affects segmentation, several aspects of our results strongly suggest that the answer is not simple. The language effect here is small and not the same for all models. In other words, even though part of the small effect could be attributed to the fact that both languages are agglutinative, results clearly indicate that morphological complexity is not the sole determinant for word segmentation, and invite a consideration within each algorithm instead.

Our strongest predictions pertained AGu, whose results matched our predictions well, with higher performance for Japanese than Chintang when evaluating on words. This language difference was reduced when evaluating on morphemes. Also consistent with the proposal that AGu, and probably lexical algorithms in general, are ideal to recover recombinable units is the observation that performance was overall higher for morpheme-based than word-based evaluation.

Readers may also notice that AGu achieves the highest scores, providing further evidence to previous observations that lexical models tend to outperform sub-lexical ones (Ludusan *et al.*, 2017). Although this is a desirable feature, we point out that the fact that it is more affected by morphological variation may make it implausible as a strategy for infants learning any and all languages.

As we had predicted by virtue of it being a purely phonotactic-based model, DiBS is less affected by language or level differences. Most token F-scores are similar to the morphologically complex Russian (.35), although markedly lower than English (Daland, 2009). However, Token F-scores obtained with DiBS vary markedly for Japanese morphemes, some of them reaching 0.5. The best explanation for these differences probably requires a recourse to phonological differences across the languages (Daland & Zuraw, 2013), which is orthogonal to our key question.

The most complex patterns are found for prelexical TP. Morphological complexity did not have a systematic effect on performance, as predicted, although we did find an interaction between subtypes, levels, and languages that is not simple to explain briefly.

Scores were higher for morphemes than words for all TPa. This fits in observations by Gervain & Erra 2012, who found that absolute-threshold TP tends to oversegment. When evaluated in morphemes, the greater number of boundaries is not a problem. Contrastingly, better performance for words than morphemes for relative subtypes is not unexpected given that in this class a boundary can only be posited in relatively long strings of syllables, and thus it will tend to undersegment when evaluated in terms of morphemes. TPa may meet both desiderata of high performance and cross-linguistic validity.

5.1 Limitations and future directions

Clearly, our work barely scratches the surface in terms of segmentation differences and similarities across languages. We would look forward to further research incorporating more languages to investigate the impact of linguistic traits (both morpholexical, as studied here, and phonological, as studied elsewhere Daland & Zuraw 2013; Fourtassi *et al.* 2013).

One obvious roadblock facing the generalization of the approach used here to other morphologically varied languages is the sheer paucity of data, since there are overall few corpora of child-directed speech in typologically diverse languages. In ongoing work (Bernard *et al.*, submitted), we are exploring the stability of results as a function of corpus size, to assess what is the minimum size of corpus which would lead to generalizable results. Those analyses suggest that about 5,000 word tokens may suffice – but that analysis was only based on English, and thus further methodological

research is needed to confirm and extrapolate to typologically diverse languages.

But if that approximation were confirmed, then the next roadblock is whether the data that have been (a) morphologically parsed and (b) rendered comparable by e.g. using similar definitions of what a sentence, a word, and a morpheme are. This, we believe, will be an even more challenging obstacle. One of the reasons why we focused on only two languages was because of the way in which they had been carefully curated to be as comparable as possible. Our approach could be generalized to other corpora in AcqDiv which are large enough, although we believe substantial effort would be necessary to generalize it even further to corpora repositories such as CHILDES (MacWhinney, 2014), where, despite clear guidelines seeking to standardize input format, morphological and sentence parsing are ultimately left to the discretion of each corpus' curator.

As additional languages are studied, we hope future researchers retain our strategy of employing a range of plausible models. For instance, as noted above, we focused here on general patterns that were reproduced across different models. However, it would be interesting to "tweak" the models in various ways. One obvious line for exploration would involve changing the input from syllables to phonemes or vice versa, since each of the models used either one or the other, and current results in infant research suggest infants have access to both levels of representation (e.g., Bertoncini & Mehler 1981; Seidl *et al.* 2009).

A more interesting path would be to change the concentration parameters a and b in the adaptor grammar, which, as explained above, govern the reuse versus generation of new lexical items. It is likely that these parameters affect performance, and some previous cross-linguistic work has indeed varied them (Phillips & Pearl, 2014). Similarly, one could build more complex grammars, with various levels of collocation to model the fact that words/morphemes are not independent of previous words – and more generally, that there could be types of words/morphemes typically following each other (e.g., "the" will be followed by a noun or a noun phrase but rarely a verb in English). One consideration this line of research will face is how the child plausibly "decides" which parameters to use and which sets of rules to start with. That is, all of the implementations of the TP family are entirely unsupervised, and DiBS only relies on one supervised parameter which could be replaced in the future with some learning process. Both sublexical families are also extremely simple in terms of their processes and internal architecture. This contrasts even with the AGu system we used, and thus it remains for future work to assess to what extent the whole architecture can be derived in an unsupervised fashion.

5.2 Final conclusions

Both languages studied here yielded lower segmentation scores than those reported in previous work applying the same algorithms to a morphologically simpler language (English). Moreover, a regression suggested lower performance for the more complex of our two languages. However, this regression also suggested complex patterns of interaction depending on the specific model and the level evaluated (i.e., whether word or morpheme boundaries were considered). Future work on additional languages and models would be desirable.

Références

BARTLETT S., KONDRAK G. & CHERRY C. (2009). On the syllabification of phonemes. In *Proceedings of human language technologies : The 2009 annual conference of the north american chapter of the association for computational linguistics*, p. 308–316 : Association for Computational Linguistics.

BATCHELDER E. O. (2002). Bootstrapping the lexicon : A computational model of infant speech segmentation. *Cognition*, **83**(2), 167–206.

BERNARD M., THIOLLIERE R., SAKSIDA A., LOUKATOU G., LARSEN E., JOHNSON M., FIBLA L., DUPOUX E., DALAND R. & CRISTIA X. N. C. A. (submitted). Wordseg : Standardizing unsupervised word form segmentation from text.

BERTONCINI J. & MEHLER J. (1981). Syllables as units in infant speech perception. *Infant behavior and development*, **4**, 247–260.

BICKEL B., GRENOBLE L. A., PETERSON D. A. & TIMBERLAKE A. (2013). *Language typology and historical contingency : In honor of Johanna Nichols*, volume 104. John Benjamins Publishing Company.

BORUTA L., PEPERKAMP S., CRABBÉ B. & DUPOUX E. (2011). Testing the robustness of online word segmentation : Effects of linguistic diversity and phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, p. 1–9 : Association for Computational Linguistics.

CHRISTIANSEN M. H. & CURTIN S. (2005). Integrating multiple cues in language acquisition : A computational study of early infant speech segmentation. *Connectionist models in cognitive psychology*, p. 347–372.

DALAND R. (2009). *Word segmentation, word recognition, and word learning : A computational model of first language acquisition*. PhD thesis, Northwestern University.

DALAND R. & PIERREHUMBERT J. B. (2011). Learning diphone-based segmentation. *Cognitive science*, **35**(1), 119–155.

DALAND R. & ZURAW K. (2013). Does korean defeat phonotactic word segmentation ? In *ACL (2)*, p. 873–877.

DEKEYSER R. M. (2005). What makes learning second-language grammar difficult ? a review of issues. *Language learning*, **55**(S1), 1–25.

FOURTASSI A., BÖRSCHINGER B., JOHNSON M. & DUPOUX E. (2013). Whyisenglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, p. 1–10.

GERVAIN J. & ERRA R. G. (2012). The statistical signature of morphosyntax : A study of hungarian and italian infant-directed speech. *Cognition*, **125**(2), 263–287.

GOLDWATER S., GRIFFITHS T. L. & JOHNSON M. (2009). A bayesian framework for word segmentation : Exploring the effects of context. *Cognition*, **112**(1), 21–54.

HOFF E. (2013). *Language development*. Cengage Learning.

JAROSZ G. & JOHNSON J. A. (2013). The richness of distributional cues to word boundaries in speech to young children. *Language Learning and Development*, **9**(2), 175–210.

JOHNSON M. (2008). Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and gy*, p. 20–27 : Association for Computational Linguistics.

JOHNSON M. & DEMUTH K. (2010). Unsupervised phonemic chinese word segmentation using adaptor grammars. In *Proceedings of the 23rd international conference on computational linguistics*, p. 528–536 : Association for Computational Linguistics.

JOHNSON M., GRIFFITHS T. L. & GOLDWATER S. (2007). Adaptor grammars : A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems*, p. 641–648.

KUNO S. (1973). *The structure of the Japanese language*, volume 3. MIT press Cambridge, MA.

LIGNOS C. (2011). Modeling infant word segmentation. In *Proceedings of the fifteenth conference on computational natural language learning*, p. 29–38 : Association for Computational Linguistics.

LUDUSAN B., MAZUKA R., BERNARD M., CRISTIA A. & DUPOUX E. (2017). The role of prosody and speech register in word segmentation : A computational modelling perspective. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, p. 178–183.

MACWHINNEY B. (2014). *The CHILDES project : Tools for analyzing talk, Volume II : The database*. Psychology Press.

MCCARTHY J. J. & PRINCE A. S. (1986). *Prosodic morphology*. Wiley Online Library.

MERSAD K. & NAZZI T. (2012). When mommy comes to the rescue of statistics : Infants combine top-down and bottom-up cues to segment speech. *Language Learning and Development*, **8**(3), 303–315.

MONAGHAN P. & CHRISTIANSEN M. H. (2010). Words in puddles of sound : Modelling psycholinguistic effects in speech segmentation. *Journal of child language*, **37**(3), 545–564.

NGON C., MARTIN A., DUPOUX E., CABROL D., DUTAT M. & PEPERKAMP S. (2013). (non) words,(non) words,(non) words : evidence for a protolexicon during the first year of life. *Developmental Science*, **16**(1), 24–34.

PAUDYAL N. P. (2015). *Aspects of Chintang syntax*. PhD thesis, University of Zurich, Philosophische Fakultat.

PHILLIPS L. & PEARL L. (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*, volume 36.

PHILLIPS L. & PEARL L. (2015). The utility of cognitive plausibility in language acquisition modeling : Evidence from word segmentation. *Cognitive science*, **39**(8), 1824–1854.

SAFFRAN J. R., ASLIN R. N. & NEWPORT E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, p. 1926–1928.

SAFFRAN J. R., NEWPORT E. L. & ASLIN R. N. (1996b). Word segmentation : The role of distributional cues. *Journal of memory and language*, **35**(4), 606–621.

SAKSIDA A., LANGUS A. & NESPOR M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science*, **20**(3).

SCHIKOWSKI R., PAUDYAL N. & BICKEL B. (2015). Flexible valency in chintang. *Valency Classes : a Comparative Handbook*.

SEIDL A., CRISTIÀ A., BERNARD A. & ONISHI K. H. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, **5**(3), 191–202.

STOLL S., MAZARA J. & BICKEL B. (2016). The acquisition of polysynthetic verb forms in chintang.

TSUJIMURA N. (2013). *An introduction to Japanese linguistics*. John Wiley & Sons.

VENKATARAMAN A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, **27**(3), 351–372.