Developmental Science

Developmental Science (2016), pp 1–11 DOI: 10.1111/desc.12390

PAPER

Co-occurrence statistics as a language-dependent cue for speech segmentation

Amanda Saksida, Alan Langus and Marina Nespor

Language, Cognition and Development Laboratory, SISSA – International School for Advanced Studies, Trieste, Italy

Abstract

To what extent can language acquisition be explained in terms of different associative learning mechanisms? It has been hypothesized that distributional regularities in spoken languages are strong enough to elicit statistical learning about dependencies among speech units. Distributional regularities could be a useful cue for word learning even without rich language-specific knowledge. However, it is not clear how strong and reliable the distributional cues are that humans might use to segment speech. We investigate cross-linguistic viability of different statistical learning strategies by analyzing child-directed speech corpora from nine languages and by modeling possible statistics-based speech segmentations. We show that languages vary as to which statistical segmentation strategies are most successful. The variability of the results can be partially explained by systematic differences between languages, such as rhythmical differences. The results confirm previous findings that different statistical learning strategies are successful in different languages and suggest that infants may have to primarily rely on non-statistical cues when they begin their process of speech segmentation.

Research highlights

- Although infants can use distributional regularities to start segmenting words from fluent speech, co-occurrence statistics are not equally informative in all languages.
- A possible source of statistical variance between languages is linguistic rhythm.
- Infants may use the language-specific information about rhythm to narrow down possible associative strategies to segment speech.

Introduction

When presented with a continuous stream of syllables, many animals, among them humans (adults and infants), non-human primates, and rodents, show a striking ability: they are able to group sequences of syllables with higher statistical coherence and delimit them from sequences with lower statistical coherence (Hauser, Newport & Aslin, 2001; Saffran, Aslin & Newport, 1996a; Toro & Trobalón, 2005). While this ability is not limited to speech but can be used to delimit coherent sequences also in non-linguistic auditory, visual, or cross-modal stimuli (Fiser & Aslin, 2002; Fiser, Scholl & Aslin, 2007; Kirkham, Slemmer & Johnson, 2002; Saffran, Johnson, Aslin & Newport, 1999), humans may use this ability predominantly for segmenting words from continuous speech. Across languages, syllable pairs with higher statistical coherence tend to belong to the same word (Harris, 1955; Hayes & Clark, 1970). Language learners could therefore use this information to segment syllable sequences that are more likely to constitute words in a given language. Word segmentation using these general associative learning mechanisms should be particularly useful when language-specific knowledge is degraded, unfamiliar or absent altogether. In fact, some experiments have suggested that very young infants rely more on the statistical rather than on the language-specific cues (i.e. lexical stress) when they collide, whereas later on this preference may weaken (Thiessen & Saffran, 2003, 2007).

Address for correspondence: Amanda Saksida, Language, Cognition and Development Laboratory, SISSA – International School for Advanced Studies, Via Bonomea 265, Trieste, Italy; e-mail: amanda.saksida@sissa.it

Numerous computational models have tried to identify the most probable learning strategy that might account for the observed behavior. One possibility is that humans are able to chunk the input into coherent sequences because they sometimes occur in isolation and because they occur frequently enough; these strategies are explored by a family of lexical and chunking models that build their predictions on various aspects of frequency detection (Goldwater, Griffiths & Johnson, 2009; Hewlett & Cohen, 2011; Perruchet & Vinter, 1998). A surprisingly strong sensitivity to frequency information has also been found experimentally, in French infants (Ngon, Martin, Dupoux, Cabrol, Dutat *et al.*, 2013). Another option is that humans cluster the most coherent sequences into word candidates on the basis of both frequency and co-occurrence probability; a version of this model that used mutual information as a co-occurrence measure could partially explain how infants start to segment the most common type of words in English and Dutch (Swingley, 2005). And finally, one of the first and most prominent models that attempted to explain this human behavior suggested that infants and adults compute primarily transitional (conditional) probabilities among adjacent syllables (Saffran *et al.*, 1996a; Saffran, Newport & Aslin, 1996b). Indeed, if they cannot use frequency information, infants appear to rely on transitional probabilities between syllables (Aslin, Saffran & Newport, 1998).

A recent study that modeled human experimental data on speech segmentation with the above presented models suggested that human performance is most successfully represented by a version of the lexical chunking models, and almost equally successfully by a transitional probability model that uses an absolute transitional probability threshold to determine word candidates (Frank, Goldwater, Griffiths & Tenenbaum, 2010). This leaves a certain degree of uncertainty about the type of computations humans are using for segmenting speech when only statistical cues are present. The uncertainty remains equally present if we focus on the complementary question: How successful is each of these lines of models in segmenting natural speech? Natural speech is very different from the stimuli presented in most of the experiments: it contains numerous words of different lengths, and the same syllables can co-occur with many different syllables also within words, yielding overall lower statistical coherency. Both lexical chunking models and models based on co-occurrence probabilities have been used to segment natural speech corpora (Batchelder, 2002; Boruta & Peperkamp, 2011; Fourtassi, Orschinger, Dupoux & Johnson, 2013; Gambell & Yang, 2006; Gervain & Guevara Erra, 2012; Hewlett & Cohen, 2011; Jarosz & Johnson, 2013; Johnson & Demuth, 2010; Johnson, 2008; Monaghan, Chater & Christiansen, 2005; Yang, 2004). For each type of model, various languages have been tested, and the common conclusion is that there are substantial cross-linguistic differences, no matter which type of model is used (Fourtassi *et al.*, 2013; Jarosz & Johnson, 2013). To account for these cross-linguistic differences, some explanations have been proposed: the differences in statistical segmentations could be related either to the differences in morpho-syntactic features (Gervain & Guevara Erra, 2012; Onnis & Thiessen, 2013), or to a general segmentation ambiguity, linked to a trade-off between syllabic complexity and word length (Fourtassi *et al.*, 2013). However, none of the studies analyzes numerous languages using exactly the same methodology. Furthermore, not many languages are analyzed in total. Therefore, two questions remain largely unanswered: (1) Does the variability of the results stem from some systematic differences across languages? If so, what are the probable accounts? (2) What do these differences mean for a language learner, i.e. how does a potential language learner select the most efficient among all the possible segmentation strategies, and what is the relationship between language-dependent and universal associative cues for learning?

To answer the first question and possibly address the second one, we analyzed transcribed child-directed corpora from nine different languages (English, Polish, Dutch, Italian, Spanish, Hungarian, Estonian, Japanese, Tamil). Because of their relative simplicity and their prominence in infant studies, we limited ourselves to the models based on co-occurrence probabilities. There are broadly two types of such models. One model segments speech by posing a boundary where transitional or co-occurrence probabilities are locally lowest (relative thresholding) (Saffran *et al.*, 1996a). By definition, such a segmentation strategy implies that listeners only find the words that are longer than one syllable (Yang, 2004). In the alternative model, words can be segmented by extracting the syllable sequences with transitional probabilities higher than a certain absolute level of co-occurrence probability (absolute thresholding) (Gervain & Guevara Erra, 2012; Swingley, 2005). We used both versions, and in each of them we explored the information provided by different syllable-based distributional probabilities (forward and backward transitional probabilities, their combination, and mutual exclusivity). For each model and for each distributional cue, we computed the success rate in segmenting words in each language. To verify the hypothesis that different success rates in segmenting words using co-occurrence statistics may reflect cross-linguistic variation in morpho-syntactic and phonological properties, we compared some of these properties to the obtained results. These properties could

either define which statistical information is to be used, or could possibly serve as a cue to segment words that is more reliable than statistical cues.

Methods

Corpora

We analyzed the transcribed spoken corpora in nine different languages from the CHILDES database (MacWhinney, 2000): Estonian (Argus, 2004; Kohler, 2004; Korgesaar, 2011; Vija, 2004), Hungarian (Gervain & Guevara Erra, 2012), Japanese (Ishii, 1999; Oshima-Takane, MacWhinney, Sirai, Miyata & Naka, 1998; Ota, 2003), Tamil (Narasimhan, 2004), Italian (Antelmi, 2004; Antinucci & Parisi, 1973; Tonelli, 2004; Volterra, 1976), Spanish (Goga, 2006; Jackson-Maldonado & Thal, 1994; Vila, 1990), Dutch (Bol, 1995; Van Kampen, 1994; Wijnen, 1992), Polish (Smoczynska, 1985; Weist & Witkowska-Stadnik, 1986), and English (Korman, 1984; Swingley, 2005). We chose languages that belong to different linguistic families (Slavic, Romance, Germanic, Finno-Ugric, Dravidian, Japonic) and differed in a number of grammatical features, such as word order, morpho-syntactic complexity, and phonological features (Dryer & Haspelmath, 2011). The choice of the languages was determined by the availability of the child-directed corpora and the availability of native speakers who segmented the corpora into syllabic sequences. In each corpus, only the child-directed sentences spoken by adults were taken into account. Phonetic transcription was used wherever the spelling differed from the orthographic transcription. The syllabified corpora used for the analysis are available from the authors upon request.

In order to ensure that we were analyzing a comparable amount of data, we selected 3300 sentences for each language. Although the size of the corpora typically does not significantly change the co-occurrence statistics (Gambell & Yang, 2006), the relatively small sizes of the corpora could affect the results of the segmentation process. We compared the results in our study to the results using the larger amount of input data in the following languages in which larger corpora were available: Hungarian and Italian available from CHILDES database (Gervain & Guevara Erra, 2012; http://childes.psy.cmu.edu/derived/; 15,200 and 10,470 sentences each), and the syllabified corpora of Dutch (Swingley, 2005; van de Weijer, 1998; 10,700 sentences) and English (Korman, 1984; Swingley, 2005, 12,800 sentences), can be obtained by request from the author. The segmentation results with larger corpora are presented in the Supplementary Material.

Dependency measures

We analyzed adjacent dependencies (forward transitional probabilities (FTP), backward transitional probabilities (BTP), and mutual information (MI)) among the syllables in our corpora. The dependencies were computed as follows:

$$\mathrm{FTP(XY)} = \mathrm{frequency(XY)}/\mathrm{frequency(X)}$$

$$\mathrm{BTP(XY)} = \mathrm{frequency(XY)}/\mathrm{frequency(Y)}$$

$$\mathrm{MI(XY)} = \log_2(\mathrm{frequency(XY)}/(\mathrm{frequency(X)} * \mathrm{frequency(Y)}))$$

Adjacent transitional probability is the conditional probability statistic that measures how predictive adjacent elements are. It is the main statistical measure in various word segmentation models (Aslin *et al*., 1998; Frank *et al*., 2010; Tyler & Cutler, 2009). Both adults and infants can use both forward and backward transitional probabilities, and while the preference for using one of the two measures grows with the language experience, it is unclear whether infants rely more on one or the other measure or both (Onnis & Thiessen, 2013; Pelucchi, Hay & Saffran, 2009). We therefore also computed the logically possible combination of the two measures (FTP&BTP) (Gervain & Guevara Erra, 2012). For the combined FTP&BTP measure, both the FTP and the BTP values had to reach the segmentation criterion for a word boundary to be posited. Mutual information is a symmetrical measure, similar to transitional probabilities. It has been used to measure the strength of associations between words in written corpora and is now used in many corpora for extracting frequently co-occurring word pairs (Church & Hanks, 1990; Hayes & Clark, 1970; Mihalcea, Corley & Strapparava, 2006). Recently, mutual information has also been used to model the word segmentation process (Huang, 2012; Swingley, 2005). This measure is usually not normalized and its range varies in different corpora.

The languages we chose differ significantly in a number of quantitative features, such as average word and utterance length and syllabic diversity. In all languages, co-occurrence statistics are stronger within words and weaker at word boundaries: average values in all dependencies measures are lower in the across-word syllable pairs and higher in the within-word syllable pairs (Table 1). This confirms the findings of previous studies where TP drops at word boundaries were observed (Gervain & Guevara Erra, 2012; Harris, 1955; Hayes & Clark, 1970).

Non-adjacent dependencies were not measured because in our corpora the proportion of words containing more than two syllables is relatively low. Fur-

Table 1 *Quantitative features of the analyzed corpora. We use the standard corpus terminology, where type means a unique string in the corpus (regardless of its frequency), and token means any string (regardless of its uniqueness)*

	English	Polish	Dutch	Spanish	Italian	Hungarian	Estonian	Japanese	Tamil
Words	9196	16529	14475	14710	15777	12669	14590	9621	9226
Words/utterance	3.787	5.008	4.387	4.456	4.780	3.839	4.422	2.936	2.795
Syllables/word	1.159	1.740	1.275	1.670	1.830	1.700	1.690	2.330	2.340
Syllables types	993	1137	1049	764	788	1718	1289	371	1108
Bigrams types	7720	8936	8200	7474	7852	10807	10710	3956	7108
Bigrams tokens	11703	28738	18467	24571	28882	21492	24649	23029	21612
Word order	SVO	SVO	SVO	SVO	SVO	SOV	SVO	SOV	SOV
Proportion of syllabic intervals (%V)	40.1	41	42.3	43.8	45.2	48.2		53.2	53.3
Mean word-internal FTP	0.493	0.145	0.387	0.181	0.242	0.228	0.233	0.141	0.262
Mean word-straddling FTP	0.156	0.126	0.114	0.092	0.041	0.153	0.096	0.056	0.109
Mean word-internal BTP	0.477	0.259	0.321	0.233	0.133	0.323	0.186	0.101	0.222
Mean word-straddling BTP	0.158	0.080	0.126	0.072	0.093	0.108	0.112	0.088	0.141
Mean word-internal MI	−5.295	−10.691	−7.057	−10.194	−10.979	−7.454	−9.520	−12.881	−8.954
Mean word-straddling MI	−10.196	−12.554	−11.271	−12.674	−13.196	−11.057	−11.578	−14.063	−11.717

thermore, infants appear to disregard non-adjacent dependencies when the adjacent ones are high (Gomez, 2002; Gomez & Maye, 2005). Although smaller and larger constituents are sometimes considered as well (Batchelder, 2002; Bonatti, Peña, Nespor & Mehler, 2007; Brent & Cartwright, 1996; Monaghan *et al.*, 2005), syllable has been predominantly recognized as a minimal perceptual unit in speech and has therefore been used as a minimal input unit also in models of infant speech segmentation (Bertoncini, Floccia, Nazzi & Mehler, 1995; Gambell & Yang, 2006; Mehler, 1981; Saffran *et al.*, 1996a; Swingley, 2005). Furthermore, in the study of Gervain and Guevara-Erra, the segmentation algorithms based on co-occurrence probabilities performed significantly worse when computed over phonemes than over syllables both in Hungarian and Italian (Gervain & Guevara Erra, 2012). We therefore used adjacent dependencies among syllables in our study.

Segmentation algorithms

In each of the corpora we used two segmentation algorithms to determine possible word boundaries, both using only the dependency measures described above. In the first algorithm (Relative), the boundaries are set wherever the dependency measure value P (FTP, BTP, FTP&BTP, or MI) of a syllable pair (XY) is weaker than in the neighboring ones (ZX and YW), defined as:

$$P(ZX) > P(XY) < P(YW)$$

For the second model we used an algorithm that looks for drops in TP or MI values below a certain general threshold, setting the word boundary whenever the values are lower than the threshold and extracting the words that contain syllable pairs with co-occurrence probabilities higher than the threshold (Absolute algorithm). Although in principle many different thresholds can be used for delimiting words in the stream (Gervain & Guevara Erra, 2012; Swingley, 2005), it is highly improbable that any real language learner would repeatedly segment word candidates from the input using many different thresholds and then select the threshold that gives the best result. We therefore took the average values of syllable pairs – computed separately for each dependency measure in each language – as suitable absolute thresholds. To verify that the average FTP, BTP, FTB&BTP, and MI are valid candidates for thresholds, we compared the result obtained by using a threshold that gave the best results among 100 percentile thresholds (Gervain & Guevara Erra, 2012) to the result obtained using the average values as an absolute threshold in each language. We found no significant difference (GLM with within-language factors Threshold and Measure showed significant effect of Measure ($F(3) = 9.859$, $p = .000$) and no effect of Threshold ($F(1) = 0.116$, $p = .742$), with significant interaction between the two factors ($F(3) = 14.898$, $p = .000$), reflecting much bigger differences between the results with the two thresholds in the FTP&BTP measure). The results using 100 percentile thresholds in both algorithms using all three measures are presented in the Supplementary Material (Figure S1).

Evaluation measures

We evaluate each segmentation strategy using the conventional information retrieval measures (Baeza-Yates & Ribeiro-Neto, 1999): Precision, Recall, and their harmonic mean, *F*-score, as defined in:

$$\text{Precision} = \#\text{hits}/(\#\text{hits} + \#\text{false alarms})$$

$$\text{Recall} = \#\text{hits}/(\#\text{hits} + \#\text{misses})$$

$$F = (2 * \text{precision} * \text{recall})/(\text{precision} + \text{recall})$$

Whenever the proportion of hits is substantially lower than the proportion of falsely selected or missed words (when either precision or recall are lower than 35%), the *F*-score will be lower than 0.5.

Procedure

In each corpus, we first measured the dependencies (FTP, BTP, FTP&BTP, MI) among syllables within words and at word boundaries. In a second step, the information about the word boundaries was removed and only the utterance boundaries remained. Utterances in child-directed speech corpora in CHILDES are delimited either by pauses or by another person's utterances. In both cases a word boundary is indicated without any doubt. Furthermore, infants are sensitive to utterance boundaries and can make use of them (Jusczyk, 1999). We therefore kept the utterance boundaries as unambiguous word boundary information. Each of the segmentation algorithms produced a distinct set of word candidates in each language. These word candidates were compared to the actual words in the same corpus. The input for learning (measuring the dependencies) and modeling the word segmentation was the same because we wanted to directly compare the actual words and the word candidates produced by the models in each corpus. The scripts (in Python) that were used during the whole procedure are available upon the request from the authors.

Results

Overall differences

In all languages and in both algorithms, overall segmentation results are relatively high, spanning from 0.49 to 0.86 (Figure 1). (The table with Recall, Precision, and *F*-scores can be found in the Supplementary Material, Tables S1 and S2.) However, languages that perform best in one algorithm tend to have the lowest results in the other algorithm. Overall differences between languages could therefore only be assessed if the results from each algorithm are observed separately. Post-hoc comparisons of one-way analysis of variance between the results in the nine languages in Absolute algorithm ($F(8) = 10.110$, $p < .001$) show significant differences between Dutch and English on the one side and Estonian, Japanese, and Tamil on the other side. In the Relative algorithm, differences are smaller and significant only between Tamil on the one hand and English and Dutch, on the other hand ($F(8) = 3.130$, $p = .012$). The results are overall significantly higher with the Absolute algorithm than with the Relative algorithm. The differences between the results when different dependency measures are taken are somewhat smaller, the biggest being when FTP&BTP were measured. The general linear model (GLM) of variance between the *F*-scores in each algorithm and all four measures – with within-language factors Measure and Algorithm, and Language as a covariate – shows significant effect of Algorithm ($F(1) = 16.370$; $p = .005$), and of Measure ($F(3) = 4.418$; $p = .015$), with no interaction between the factors ($F(3) = 0.805$, $p = .505$) (Figure 2).

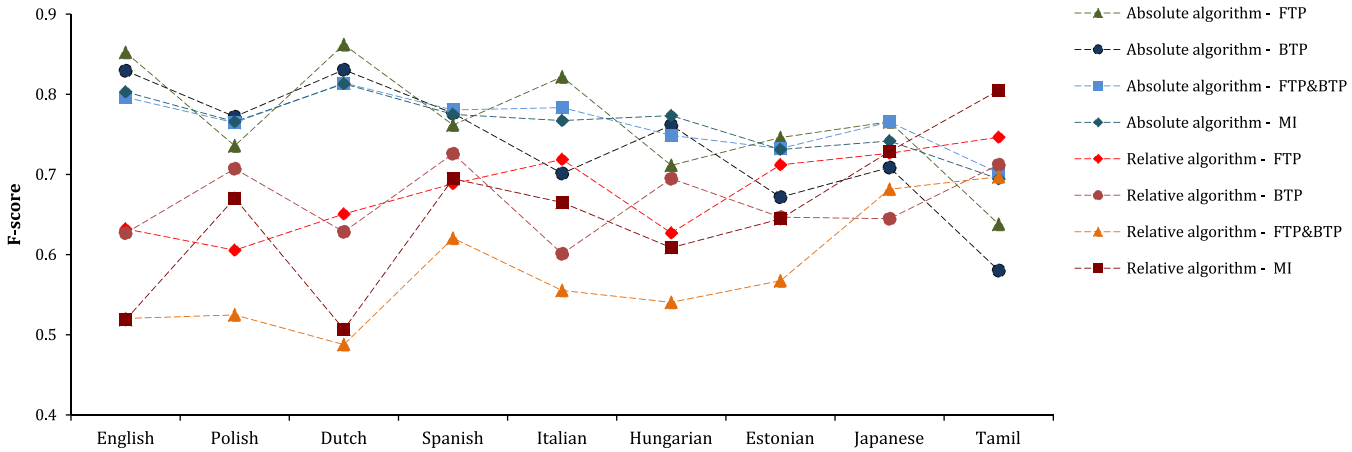

Figure 1 *The F-scores in each algorithm for all measures across languages. The languages are ordered according to their average proportions of vocalic intervals: English is a stress-timed language with the lowest proportion, while Tamil a mora-timed language with the highest proportion of vocalic intervals.*

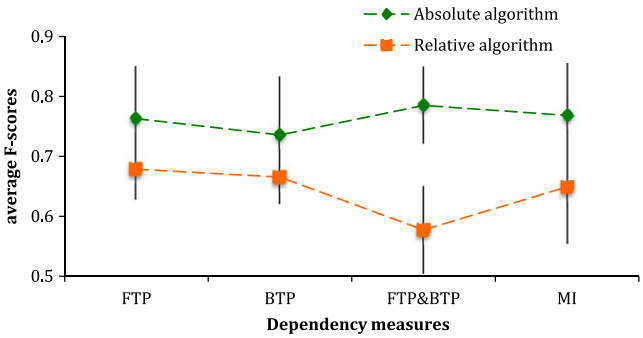

Figure 2 *Means of F-scores in the nine languages in the two algorithms with dependency measures used in the analysis. Error bars represent standard deviation. The differences between the results when using different measures are not significant within each algorithm, but there is an overall difference between the two algorithms.*

Cross-linguistic differences

Although the results are overall high, there are substantial differences among languages in each algorithm, and the goal of this article is to assess the possible sources of these differences. One possibility is that the cross-linguistic variation in segmentation success lies in morpho-syntactic differences between the languages analyzed: In head-initial languages with a default subject-verb-object (SVO) word order, such as English or Italian, the head of a phrase precedes its complements and therefore forward TPs could be more informative than backward TPs. Conversely, in head-final languages with a default subject-object-verb (SOV) word order, thus with the head of a phrase following its complements, BTPs could be more informative (Gervain & Guevara Erra, 2012; Onnis & Thiessen, 2013). However, our language samples cannot confirm this hypothesis. BTPs are more informative only in Polish and Spanish, which are head-initial languages with a default SVO order, and in Hungarian (which is a head-final language, as predicted by the hypothesis), whereas in other two head-final languages in our sample, Japanese and Tamil, FTPs and MI are more informative. If we add Word order as the between-language factor to the general linear model with the factors Measure and Algorithm, we find no interaction between Word order and the factor Measure. Furthermore, Word order is not a significant predictor of the results with any of the measures (see correlations in Table 2). In our sample of languages, word order therefore cannot serve as a predictor of which dependency measure is more informative.

Another possibility is that segmentation differences stem from phonological differences and are linked to syllable complexity and word length (Fourtassi *et al.*, 2013). Both features are closely related to linguistic rhythm – a feature that is perceptually salient both for infant and adult language learners (Bolger, Trost & Schön, 2013; Ramus, Hauser, Miller, Morris & Mehler, 2000; Werker & Vouloumanos, 2000). Human languages vary significantly regarding their basic rhythm. Some have a more regular, machine gun-like rhythm, as in Spanish, and they are classified as syllable-timed. In others, the basic rhythm is less regular, yielding a Morse code-like rhythm, as in English, and they are classified as stress-timed. A third type of language, typically Japanese, appears to be isochronous sub-syllabically, and they are classified as mora-timed (Mehler & Nespor, 2004). Linguistic rhythm strongly correlates to average word length (words in stress-timed languages are shorter than in mora-timed languages) and to average syllabic complexity (most complex syllables are only possible in stress-timed languages). But recent studies have shown that linguistic rhythm can be most accurately quantified if the average proportion of vocalic intervals is measured (Nespor, Shukla & Mehler, 2011; Ramus, Nespor & Mehler, 1999). In stress-timed languages, the average proportion of vocalic intervals is the lowest, and in mora-timed languages it is the highest. We compared the

Table 2 *The bivariate correlations between the results in each segmentation model and the relevant features of the languages analyzed*

				Absolute algorithm				Relative algorithm			
Pearson's correlations		Word length (syll)	Vocalic proportions	FTP	BTP	FTP& BTP	MI	FTP	BTP	FTP& BTP	MI
Word order	rho	0.704	0.905	−0.643	−0.499	−0.584	−0.542	0.315	0.31	0.639	0.512
	sig	0.034	0.002	0.062	0.172	0.098	0.132	0.409	0.417	0.064	0.158
Word length	rho		0.888	−0.718	−0.824	−0.695	−0.856	0.691	0.314	0.874	0.929
(syllables)	sig		0.003	0.03	0.006	0.038	0.003	0.039	0.41	0.002	0
Vocalic	rho			−0.656	−0.811	−0.731	−0.818	0.737	0.193	0.842	0.73
proportions	sig			0.078	0.015	0.04	0.013	0.037	0.647	0.009	0.04

results in the languages analyzed to the average word length – measured in the same corpora – and to the average proportions of vocalic intervals – measured on independent adult-directed corpora (Nespor *et al.*, 2011; Ramus *et al.*, 1999). The measurements exist for all languages analyzed except Estonian, so the following analyses are carried out without Estonian data on rhythm.

The results from the two algorithms correlate both with word length and with rhythm measurements (Table 2). The segmentation with the Absolute algorithm in all dependency measures is negatively correlated to the proportion of vocalic intervals ($R^2 < -0.689$, $p < .078$) as well as to the average word length ($R^2 < -0.695$, $p < .038$). The word segmentation results with the Relative algorithm are, conversely, positively correlated to the proportion of vocalic intervals ($R^2 > 0.730$, $p < .040$ for results with FTP, FTP&BTP and MI measures, whereas for BTP, the correlation is non-significant) and to the average word length ($R^2 > 0.691$, $p < .039$ for results with FTP, FTP&BTP and MI measures and non-significantly for BTP). To assess the relative contribution of word order and average vocalic intervals as possible predictors for segmentation results, we ran multiple regression analysis. The results show that the models are overall less significant in predicting segmentation results when both word length and average vocalic intervals are included. However, for most of the results, word length is a slightly better predictor of the more successful segmentation strategy than the average vocalic proportions (Table 3). The variation in statistical segmentation could therefore be most effectively explained by the average word length, and also by a more general measure of linguistic rhythm.

Discussion

The first aim of this study was to examine how informative co-occurrence statistics is for segmenting words from the unsegmented input in various languages. The observations are the following: (1) Our results are relatively high (Yang, 2004); in all languages and in all measures the *F*-scores are higher than 0.49. The only additional information to the co-occurrence statistics in our model consisted of utterance boundaries: because they represent long pauses or changes of speaker, they are unambiguously informative about the word boundary. This information has significantly increased the proportion of correct word candidates (see Supplementary Material for the results without utterance boundaries). (2) Our results confirm previous findings that using an invariable (absolute) threshold is in general more efficient than finding locally minimal values of chosen segmentation measures (Frank *et al.*, 2010; Gervain & Guevara Erra, 2012). (3) Different co-occurrence measures (forward transitional probabilities, backward transitional probabilities, and mutual information) affect the results of word segmentation in the two algorithms, but the differences between the results in different measures are smaller than between different algorithms (Figure 2). The combined FTP&BTP measure gives worse results than the separate FTP and BTP measures, especially in the Relative algorithm. This leaves open the question of the usability of such combination (Gervain & Guevara Erra, 2012).

The second aim of our study was to account for the cross-linguistic differences that we found and that confirm previous observations about substantial differences among languages, even when the same or a similar segmentation algorithm is used (Batchelder, 2002; Four-

Table 3 *Multiple linear regression analyses of the relative contribution of Word length and Average vocalic intervals in predicting word segmentation results using Absolute and Relative algorithms with the four dependency measures. The predictor variables were entered simultaneously into the model*

	FTP			BTP			FTP&BTP			MI		
Model	St. coeff. Beta	t	Sig.	St. coeff. Beta.	t	Sig.	St. coeff. Beta	t	Sig.	St. coeff. Beta	t	Sig.
Absolute algorithm												
(Constant)		3.619	0.015		5.347	0.003		7.93	0.001		12.072	0
Word length	−0.694	−1.039	0.346	−0.775	−1.713	0.147	−0.566	−0.92	0.4	−0.945	−2.567	0.05
Average vocalic intervals	−0.039	−0.059	0.955	−0.122	−0.271	0.797	−0.228	−0.37	0.727	0.021	0.057	0.956
Relative algorithm												
(Constant)		2.141	0.085		2.939	0.032		0.87	0.424		2.998	0.03
Word length	0.344	0.538	0.614	0.658	0.723	0.502	0.597	1.319	0.244	1.332	4.493	0.006
Average vocalic intervals	0.432	0.675	0.53	−0.392	−0.431	0.685	0.311	0.687	0.523	−0.453	−1.529	0.187

tassi *et al.*, 2013; Gervain & Guevara Erra, 2012; Jarosz & Johnson, 2013; Johnson & Demuth, 2010; Johnson, 2008; Yang, 2004). One of the proposals was that the main morpho-syntactic differences are the cause of the differences in co-occurrence statistics: forward transitional probabilities might be more informative in languages with head-initial phrase structure and the default SVO word order, whereas backward transitional probabilities are more informative in head-final languages with SOV word order (Gervain & Guevara Erra, 2012; Onnis & Thiessen, 2013). In our study, however, we could not find any systematic differences among the results in the languages when we divided them based on their basic word order.

The second proposal is that the differences among languages are relative to some of their intrinsic properties, such as average word length and syllabic complexity (Fourtassi *et al.*, 2013). Both features are strongly correlated to linguistic rhythm. We therefore ran multiple regressions between average word length, the average vocalic proportion, which is the most reliable indicator of linguistic rhythm (Nespor *et al.*, 2011; Ramus *et al.*, 1999), and the results in both algorithms. The results showed that vocalic proportions are significant predictors of the results in both algorithms, such that the Absolute algorithm will be most successful in stress-timed languages, and the Relative algorithm will be more successful in mora-timed languages. Although word length appears to be an even stronger predictor of the results, there are two points to be noted here: we took average word length directly from the corpora that we analyzed, whereas average vocalic proportions are taken from an independent analysis of adult-directed speech corpora. It is thus difficult to estimate a real relative contribution of the two. Second, even if word length is a better predictor of the most successful segmentation strategy, the question remains whether language learners could take advantage of the average word length as a cue. In fact, in order to compute average word length in their language, they would need to have at least an initial list of words already segmented.

If, conversely, rhythm is a good predictor of the type of segmentation strategy that will be successful in each language, what can that mean for a language learner? It is known that experience with language changes the strategies for segmenting words. Adult speakers are sensitive to different types of co-occurrence statistics in different languages (Onnis & Thiessen, 2013); the most prominent stress pattern in the native language alters segmentation preferences (Jusczyk, Cutler & Redanz, 1993; Jusczyk, Houston & Newsome, 1999); native language phonotactics can serve as a constraint for segmentation (Johnson, Jusczyk, Cutler & Norris, 2003; Johnson & Jusczyk, 2001; Mersad & Nazzi, 2011). But because infants are born without such rich knowledge about their native language, statistical information was proposed as a bootstrapping mechanism to start segmentation (Saffran *et al.*, 1996a; Swingley, 2005; Thiessen & Saffran, 2003). The results of our study indicate a strong correlation between distribution-based segmentation strategies and linguistic rhythm, which offers another hypothesis: rhythmical information both narrows down possible lexical structures and offers information about which type of statistics is more informative. The fact that humans are sensitive to rhythmical information from birth on (Ramus *et al.*, 2000) and that infants' segmentation is strongly influenced by the rhythmical properties of their native language (Mersad, Goyet & Nazzi, 2010; Nazzi, Iakimova, Bertoncini & Alcantara, 2006; Nishibayashi, Goyet & Nazzi, 2015) gives further support to this hypothesis. Furthermore, the possibility that rhythmical information determines which type of statistical segmentation is more informative is consistent with recent studies that have explored the interaction between domain-general and language-specific learning mechanisms during early language development: perceptual primitives and general learning mechanisms that play an important role in language development, such as frequency detection, sensitivity to edges, and the process of generalization, are constrained by universal properties of language (Endress & Hauser, 2009; Gervain, Nespor, Mazuka, Horie & Mehler, 2008; Hochmann, Endress & Mehler, 2010; Lidz, Waxman & Freedman, 2003). Word segmentation based on co-occurrence statistics could therefore be another general learning mechanism constrained by language-specific knowledge. There are, however, several questions that remain open in the present study. While this study is based on the input that infants receive during their language development, further experimental work is needed to establish whether infants indeed employ different segmentation strategies based on the rhythmical properties of their native language. Furthermore, we have only tested the two models that are based on probabilistic measures of distributional dependencies among adjacent syllables. While this is beyond the scope of the present study, the presented hypothesis, according to which a language-specific non-statistical property (such as linguistic rhythm) can predict segmentation success, could be tested with other segmentation models. This would shed additional light on the issue of the interaction between language-specific phonological properties and more general statistical properties of languages.

Acknowledgements

We thank Sašo Živanovič, Luca Filippin, Romain Brasselet, Marijana Sjekloća, Francesca Gandolfo, Linda Langus and Iga Nowak for helping with the data collection and analysis, and Jacques Mehler, Ansgar Endress, Luca Bonatti, Alejandrina Cristia, Abdellah Fourtassi, and Franck Ramus for useful comments on earlier drafts of this manuscript. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 269502 (PASCAL) to Jacques Mehler, from a SISSA Young Scientist Award to Alan Langus, and from a SISSA doctoral grant to Amanda Saksida.

References

Antelmi, D. (2004). The Antelmi corpus. *CHILDES Database*.

Antinucci, F., & Parisi, D. (1973). Early language acquisition: a model and some data. In C. Ferguson & D. Slobin (Eds.), *Studies in early language development* (pp. 607–619). New York: Holt.

Argus, R. (2004). The Argus corpus. *CHILDES Database*.

Aslin, R.N., Saffran, J.R., & Newport, E.L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, **9** (4), 321–324. doi:10.1111/1467-9280.00063

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*, Vol. **463**. New York: ACM Press.

Batchelder, E.O. (2002). Bootstrapping the lexicon: a computational model of infant speech segmentation. *Cognition*, **83** (2), 167–206. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11869723

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: rhythmical basis of speech representations in neonates. *Language and Speech*, **38** (4), 311–329. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8816088

Bol, G.W. (1995). Implicational scaling in child language acquisition: the order of production of Dutch verb constructions. In M. Verrips & F. Wijnen (Eds.), *Papers from the Dutch-German Colloquium on Language Acquisition* (pp. 1–13). *Amsterdam Series in Child Language*. Amsterdam: Institute for General Linguistics.

Bolger, D., Trost, W., & Schön, D. (2013). Rhythm implicitly affects temporal orienting of attention across modalities. *Acta Psychologica*, **142** (2), 238–244. doi:10.1016/j.actpsy.2012.11.012

Bonatti, L.L., Peña, M., Nespor, M., & Mehler, J. (2007). On consonants, vowels, chickens, and eggs. *Psychological Science*, **18** (10), 924–925. doi:10.1111/j.1467-9280.2007.02002.x

Boruta, L., & Peperkamp, S. (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* (pp. 1–9). Portland, OR: Association for Computational Linguistics.

Brent, M.R., & Cartwright, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, **61** (1–2), 93–125. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8990969

Church, K.W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16** (1), 22–29. doi:10.3115/981623.981633

Dryer, M., & Haspelmath, M.S. (Eds.) (2011). *The world atlas of language structures online*. Munich: Max Planck Digital Library. Retrieved from http://wals.info/

Endress, A.D., & Hauser, M.D. (2009). Syntax-induced pattern deafness. *Proceedings of the National Academy of Sciences of the United States of America*, **106** (49), 21001–21006. doi:10.1073/pnas.0908963106

Fiser, J., & Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences of the United States of America*, **99** (24), 15822–15826. doi:10.1073/pnas.232472899

Fiser, J., Scholl, B.J., & Aslin, R.N. (2007). Perceived object trajectories during occlusion constrain visual statistical learning. *Psychonomic Bulletin & Review*, **14** (1), 173–178. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17546749

Fourtassi, A., Orschinger, B.B., Dupoux, E., & Johnson, M. (2013). Whyisenglishsoeasytosegment? In *CMCL 2013 Cognitive Modeling and Computational Linguistics Proceedings of the Workshop* (pp. 1–10). The Association for Computational Linguistics.

Frank, M.C., Goldwater, S., Griffiths, T.L., & Tenenbaum, J.B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, **117** (2), 107–125. doi:10.1016/j.cognition.2010.07.005

Gambell, T., & Yang, C. (2006). Word segmentation: quick but not dirty. Unpublished Manuscript, Available at http://www.ling.upenn.edu/ycharles/papers/quick.pdf.

Gervain, J., & Guevara Erra, R. (2012). The statistical signature of morphosyntax: a study of Hungarian and Italian infant-directed speech. *Cognition*, **125** (2), 263–287. doi:10.1016/j.cognition.2012.06.010

Gervain, J., Nespor, M., Mazuka, R., Horie, R., & Mehler, J. (2008). Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, **57** (1), 56–74. doi:10.1016/j.cogpsych.2007.12.001

Goga, I. (2006). Educating attention in early development of imitation, language and object manipulation abilities in human infants. University Babes-Bolyai, Faculty of Psychology and Education Sciences, Cluj-Napoca.

Goldwater, S., Griffiths, T.L., & Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, **112** (1), 21–54. doi:10.1016/j.cognition.2009.03.008

Gomez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science*, **13** (5), 431–436. doi:10.1111/1467-9280.00476

Gomez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, **7** (2), 183–206.

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

Hauser, M.D., Newport, E.L., & Aslin, R.N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, **78** (3), B53–B64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11124355

Hayes, J.R., & Clark, H.H. (1970). Experiments in the segmentation of an artificial speech analog. In J.R. Hayes (Ed.), *Cognition and the development of language* (pp. 221–234). New York: Wiley.

Hewlett, D., & Cohen, P. (2011). Word segmentation as general chunking. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 39–47). Association for Computational Linguistics.

Hochmann, J.-R., Endress, A.D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, **115** (3), 444–457. doi:10.1016/j.cognition.2010.03.006

Huang, C.R. (2012). Words without boundaries: computational approaches to Chinese word segmentation. *Linguistics and Language Compass*, **6**, 494–505. doi:10.1002/lnc3.357

Ishii, T. (1999). *The JUN corpus.*

Jackson-Maldonado, D., & Thal, D. (1994). Lenguaje y cognición en los primeros años de vida: resultados preliminares. *Revista Psicología y Sociedad*, **25**, 21–27.

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, E.K., & Jusczyk, P.W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, **44** (4), 548–567. doi:10.1006/jmla.2000.2755

Johnson, E., Jusczyk, P.W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology*, **46** (1), 65–97. doi:10.1016/S0010-0285(02)00507-8

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. *Proceedings of ACL-08: HLT*, June (pp. 398–406).

Johnson, M., & Demuth, K. (2010). Unsupervised phonemic Chinese word segmentation using adaptor grammars. *Proceedings of the 23rd International Conference on Computational Linguistics*, August (pp. 528–536). Retrieved from http://portal.acm.org/citation.cfm?id=1873841

Jusczyk, P.W. (1999). Narrowing the distance to language: one step at a time. *Journal of Communication Disorders*, **32** (4), 207–222. doi:10.1016/S0021-9924(99)00014-3

Jusczyk, P.W., Cutler, A., & Redanz, N.J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, **64** (3), 675–687. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8339688

Jusczyk, P.W., Houston, D.M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, **39** (3–4), 159–207. doi:10.1006/cogp.1999.0716

Kirkham, N.Z., Slemmer, J.A., & Johnson, S.P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, **83** (2), B35–B42.

Kohler, K. (2004). The Kohler corpus. *CHILDES Database.*

Korgesaar, H. (2011). The Korgesaar corpus. *CHILDES Database.*

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, **5**, 44–55.

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition*, **89** (3), 295–303. doi:10.1016/S0010-0277(03)00116-1

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd edn.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, **4**, 247–260.

Mehler, J., & Nespor, M. (2004). Linguistic rhythm and the acquisition of language. In A. Belletti & L. Rizzi (Eds.), *Structures and beyond: The cartography of syntactic structures* (pp. 213–221). Oxford: Oxford University Press.

Mersad, K., Goyet, L., & Nazzi, T. (2010). Cross-linguistic differences in early word form segmentation: a rhythmic-based account. *Journal of Portuguese Linguistics*, **9** (10), 37–65.

Mersad, K., & Nazzi, T. (2011). Transitional probabilities and positional frequency phonotactics in a hierarchical model of speech segmentation. *Memory and Cognition*, **39** (6), 1085–1093. doi:10.3758/s13421-011-0074-3

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, **1**, 775–780. doi:10.1.1.65.3690

Monaghan, P., Chater, N., & Christiansen, M.H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, **96** (2), 143–182. doi:10.1016/j.cognition.2004.09.001

Narasimhan, R. (2004). The Tamil coprus. *CHILDES Database.*

Nazzi, T., Iakimova, G., Bertoncini, J., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, **54** (3), 283–299. doi:10.1016/j.jml.2005.10.004

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. In M. van Oostendorp, C.J. Ewen, E.V. Hume & K. Rice (Eds.), *The Blackwell companion to phonology* (**Vol. 5**, pp. 1147–1159). Oxford: Wiley-Blackwell Publishing.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., et al. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, **16** (1), 24–34. doi:10.1111/j.1467-7687.2012.01189.x

Nishibayashi, L.-L., Goyet, L., & Nazzi, T. (2015). Early speech segmentation in French-learning infants: monosyl-

labic words versus embedded syllables. *Language and Speech*, **58** (3), 334–350. doi:10.1177/0023830914551375

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. *Cognition*, **126** (2), 268–284. doi:10.1016/j.cognition.2012.10.008

Oshima-Takane, Y., MacWhinney, B., Sirai, H., Miyata, S., & Naka, N. (1998). *CHILDES for Japanese*. The JCHAT Project Nagoya, Chukyo University.

Ota, M. (2003). *The development of prosodic structure in early words: Continuity, divergence and change*. Amsterdam: John Benjamins.

Pelucchi, B., Hay, J.F., & Saffran, J.R. (2009). Learning in reverse: 8-month-old infants track backward transitional probabilities. *Cognition*, **113** (2), 244–247. doi:10.1016/j.cognition.2009.07.011.Learning

Perruchet, P., & Vinter, A. (1998). PARSER: a model for word segmentation. *Journal of Memory and Language*, **39** (2), 246–263. doi:10.1006/jmla.1998.2576

Ramus, F., Hauser, M.D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, **288** (5464), 349–351. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10764650

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, **75** (1), 265–292. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10908711

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996a). Statistical learning by 8-month-old infants. *Science*, **274** (5294), 1926–1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, **70** (1), 27–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10193055

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.

Smoczynska, M. (1985). The acquisition of Polish. In D.I. Slobin (Ed.), *The crosslinguistic study of language acquisition* (pp. 595–686). Hillsdale, NJ: Lawrence Erlbaum Associates.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, **50** (1), 86–132. doi:10.1016/j.cogpsych.2004.06.001

Thiessen, E.D., & Saffran, J.R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, **39** (4), 706–716. doi:10.1037/0012-1649.39.4.706

Thiessen, E.D., & Saffran, J.R. (2007). Learning to learn: infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, **3** (1), 73–100. doi:10.1207/s15473341lld0301_3

Tonelli, L. (2004). The Tonelli corpus. *CHILDES Database*.

Toro, J.M., & Trobalón, J.B. (2005). Statistical computations over a speech stream in a rodent. *Perception & Psychophysics*, **67** (5), 867–875.

Tyler, M.D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, **126** (1), 367–376. doi:10.1121/1.3129127

Van de Weijer, J. (1998). Language input for word discovery. Unpublished doctoral dissertation, MPI Series in Psycholinguistics 9.

Van Kampen, N.J. (1994). The learnability of the left branch condition. *Linguistics in the Netherlands*, **199**, 83–94.

Vija, M. (2004). The Vija corpus. *CHILDES Database*.

Vila, I. (1990). *Adquisición y desarrollo del lenguaje*. Barcelona: Graó.

Volterra, V. (1976). A few remarks on the use of the past participle in child language. *Journal of Italian Linguistics*, **2**, 149–157.

Weist, R., & Witkowska-Stadnik, K. (1986). Basic relations in child language and the word order myth. *International Journal of Psychology*, **21**, 363–381.

Werker, J.F., & Vouloumanos, A. (2000). Who's got rhythm ? *Science*, **288** (5464), 280–281.

Wijnen, F. (1992). Incidental word and sound errors in young speakers. *Journal of Memory and Language*, **31** (6), 374–255. doi:doi: 10.1016/0749-596X(92)90037-X

Yang, C. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences*, **8** (10), 451–456. doi:10.1016/j.tics.2004.08.006

Received: 19 September 2013
Accepted: 6 November 2015

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Supplementary Material

Figure S1. Segmentation results with different absolute thresholds. Each point represents an F-score of the segmentation algorithm using one of the 100-percentile values.

Table S1. Recall, Precision, and F-scores in Absolute algorithm for forward (FTP), backward TPs (BTP), a combination (FTP&BTP), and mutual information (MI).

Table S2. Recall, Precision, and F-scores in Relative algorithm for forward (FTP), backward TPs (BTP), a combination (FTP&BTP), and mutual information (MI).

Table S3. Results without utterance boundaries.

Table S4. Correlations between the results and the proportions of vocalic intervals

Table S5. Segmentation with larger corpora

Table S6. Word segmentation results in all languages for both algorithms when over-segmented chunks are included.