*Journal of Child Language* (2022), **49**, 1119–1146
doi:10.1017/S0305000921000568

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

# Is there a bilingual disadvantage for word segmentation? A computational modeling approach

Laia FIBLA[1,3], Nuria SEBASTIAN-GALLES[2], and Alejandrina CRISTIA[3]

[1]School of Psychology, The University of East Anglia, UK, [2]Center for Brain and Cognition, Universitat Pompeu Fabra, Spain, and [3]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University, France
Address for correspondence: Laia Fibla, School of Psychology, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK. E-mail: laia.fibla.reixachs@gmail.com

(Received 27 March 2020; revised 11 February 2021; accepted 17 June 2021;
first published online 3 November 2021)

**Abstract**

Since there are no systematic pauses delimiting words in speech, the problem of word segmentation is formidable even for monolingual infants. We use computational modeling to assess whether word segmentation is substantially harder in a bilingual than a monolingual setting. Seven algorithms representing different cognitive approaches to segmentation are applied to transcriptions of naturalistic input to young children, carefully processed to generate perfectly matched monolingual and bilingual corpora. We vary the overlap in phonology and lexicon experienced by modeling exposure to languages that are more similar (Catalan and Spanish) or more different (English and Spanish). We find that the greatest variation in performance is due to different segmentation algorithms and the second greatest to language, with bilingualism having effects that are smaller than both algorithm and language effects. Implications of these computational results for experimental and modeling approaches to language acquisition are discussed.

**Keywords:** word segmentation; infancy; computational modeling

## Introduction

Unlike in written language, there are no spaces between words when we speak. In fact, there are no obvious and infallible cues that indicate word boundaries (e.g., Brent & Siskind, 2001). Yet we know infants must have found a solution to this difficult problem because they know the meaning of some words by 6 months (Tincoff & Jusczyk, 1999, 2012), and therefore they must have been able to learn at least those phonological sequences or word forms. What is more, some evidence suggests that infants do not wait to learn true words (i.e., form-meaning pairings), but instead start segmenting their input and memorizing high frequency sequences as early as 6 months, to the point that they accumulate a proto-lexicon of about 500 word forms by 11 months (Ngon, Martin, Dupoux, Cabrol, Dutat & Peperkamp, 2013). The question of how infants approach the problem of word segmentation has been the focus of intensive cross-disciplinary research in the last years, combining experimental studies on infants and adults, mostly on monolinguals, and computational modeling. This paper is the first to investigate bilingual word segmentation using computational modeling. In the rest of this Introduction, we briefly introduce the problem of word segmentation and review other previous interdisciplinary research, before turning to our unique contributions.

### *Laboratory evidence showing that infants engage in word segmentation*

A tradition of laboratory experiments suggests that infants can segment words from running speech. In these studies, infants are typically played a word in isolation, over and over, until they reach a certain listening time criterion. They are subsequently presented with two passages, one containing the word and another not, and their listening preference is measured. Some experiments invert the order of these experiences, familiarizing to passages and testing with words in isolation. Setting aside order effects, monolingual infants show a significant preference for test trials containing the familiarized word, over test trials containing a foil, when they are as young as 8 months (Jusczyk, Houston & Newsome, 1999). This paradigm and derivations of it based on artificial languages has been used in literally hundreds of experiments on monolingual infants, meta-analyzed by Bergmann and Cristia (2015) and Black and Bergmann (2017).

The literature on bilingual infants is considerably scarcer, and results are not always consistent (for recent reviews of bilingual early language acquisition, including word segmentation, see Höhle, Bijeljac-Babic & Nazzi, 2020; Sebastian-Galles & Santolin, 2020). Polka and Sundara (2003) reported that English–French 7.5-months-old Canadian bilinguals were able to segment bisyllabic words in both their languages. However, a near replication suggested they may only succeed in French (Polka, Orena, Sundara & Worrall, 2017). With extended exposure, however, 8- and 10-month-old Canadian bilinguals succeed marginally or significantly in both their languages (Orena & Polka, 2019). Studying bilingual Spanish–Catalan infants, Bosch, Figueras, Teixidó, and Ramon-Casas (2013) found successful word segmentation in both languages at 8 months of age, and further generalized this success to 6-month-olds. Ibáñez (2017) replicates the positive finding for 6.5-, 8-, and 9.5-month-old Spanish–Catalan bilingual infants, using disyllabic target words. A recent analysis suggests that bilingual infants exposed to more frequent language mixing (including within-sentence switching) are more successful in segmenting words out of both of their languages (Orena & Polka, 2019). Moreover, monolinguals presented with a "bilingual" input in an artificial language setting failed to segment words from those languages (Antovich & Graf Estes, 2018, 2020), suggesting that in natural language acquisition, infants may lose the capacity to keep track of two sets of cues if exposed to a predominantly consistent input; or perhaps that multilinguals gain the capacity to track multiple sets of cues when exposed to variable input (see particularly follow-up analyses on frequency of mixing, in Orena & Polka, 2019; and on proportion of bilingual speakers in the environment, in Antovich & Graf Estes, 2020). Although many of these studies report significant preferences, it should be noted that the direction of preference is sometimes toward familiarity (as in most previous word segmentation studies on monolinguals) and sometimes towards novelty (i.e., longer looking to the test stimuli with a foil than with the familiarized word).

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

Of course, such diversity in outcomes may be due to the small sample sizes that many infant studies have, which make results hard to generalize (Oakes, 2017), in addition to the difficulty of recruiting homogeneous bilingual infants. This is one of the reasons why, we believe, it is particularly useful to lay the groundwork using methods that can establish the theoretical foundations, such as computational modeling.

### Segmenting words from bilingual input

In this section, we briefly discuss conceptual arguments for and against the possibility that segmenting words from bilingual input poses additional challenges compared to the monolingual case.

As will be explained in more detail in the next section, determining where words begin and end (word segmentation) may necessitate learning some aspects of the phonological system (phonological cues), as well as the way in which word forms are combined (lexical cues). Infants growing up in an environment where two languages are spoken are therefore exposed to two different lexical and phonological systems. Assuming that the overall quantity of speech infants hear is the same, one can estimate that there is twice as much to learn with the same overall quantity of input. If the discovery of distributional regularities depends on frequency, bilingual environments should be more challenging; and, if word segmentation is indeed challenging for bilinguals, we should expect, for instance, a delay in vocabulary learning. However, when properly matched, bilinguals and monolinguals know similar numbers of words in their vocabularies (see Gonzalez-Barrero, Schott & Byers-Heinlein, 2020 for a recent article showing how different approaches quantifying vocabulary size in bilinguals can lead to different results). Another variable to take into account is that depending on the languages that children are exposed to, the overlap in the lexicon and phonology across the two languages may vary. Even if the systems overlap to a great extent, this may not necessarily represent an advantage: there is less to be learned, but also a greater potential for confusion between the languages, since the linguistic variants may be harder to tease apart.

In addition, switching across languages is so common in multilingual environments that human language systems are developed in order to deal with that (Sitaram, Chandu, Rallabandi & Black, 2019). Speakers fluent in several languages may switch back and forth between them across utterances, or even integrate morphemes, words, or phrases from different languages within the same utterance. There is some discussion as to how frequently this occurs in child-directed speech, but one estimate, based on direct observations of English–Spanish bilinguals talking to their infant in a baby lab in Maryland, USA, found about 15.7% of utterances contained a within-sentence language switch (Bail, Morini & Newman, 2015), and this may be higher in communities in which multilingualism is more widespread. Even if parents do not switch languages within a sentence, unless they purposefully try to withhold speaking one of the languages in front of the child, they will speak some utterances in one language, and other utterances in the other language (see, for instance, Orena, Byers-Heinlein & Polka, 2019 for parents' speaking in both languages according to reports versus observation). And even if they do not, the child will nonetheless be exposed to changes in the language being spoken when one speaker uses one language, and the next speaker uses a different language.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

Finally, phonetic implementation may help or hinder the process of distinguishing the two languages. That is to say, imagine that the two languages have totally different phonetic inventories and that speakers are always native. In this case, the child can use these phonetic cues to separate the languages. Nevertheless, sometimes speakers are non-native, and they will say the words of one language using the phonetic implementation of sounds in the other language.

Up to here, we have presented arguments for the problem being harder when faced with bilingual than monolingual input. However, as we already mentioned, trajectories in lexical development are comparable between children exposed to one or two languages. In fact, it is also possible to argue for the prediction that there is no difference in terms of word segmentation specifically. If segmenting words depends on learning the sequences of sounds and the sequences of words found within utterances, and if speakers in a community produce stable input (that is, enough sections of a single language where both phonotactics and distributional properties of words are predictable), bilingual input may not cause particular difficulties.

In this project, we specifically control the amount of exposure to the languages as well as the prevalence of switching languages between utterances, assuming native speakers (i.e., with a native accent). We leave for future work the study of within-sentence code-switching, as well as the presence of non-native speakers.

### *Modeling word segmentation*

Computational models are a good complement to experimental language acquisition research, particularly to test the feasibility of hypothesized strategies. That is to say, if we propose that infants may use strategy X to segment words, then strategy X applied to a corpus of child-directed speech should succeed in segmenting at least some words. Models also uniquely facilitate cross-cultural and cross-linguistic analysis controlling for other variables (Monaghan & Rowland, 2017). Computational models are informative for studying the learnability properties of various language corpora, controlling for orthogonal aspects such as arousal, attention, and other performance factors. When laboratory manipulations become unfeasible (due to the number of participants or stimuli that would be required), computational models may be the only way to approach a phenomenon.

Most work that attempts to model infant word segmentation uses textual, phonologized transcriptions representing speech as a sequence of phones or syllables (for acoustically-based models that do not assume phones or syllables, see Frank, Monaghan & Tsoukala, 2019; Ludusan, Seidl, Dupoux & Cristia, 2015; Roy & Pentland, 2002). Orthographic transcriptions of child input are converted into phonological representations to create a “gold standard”, and then a version where all utterance-internal word boundaries have been removed is fed into a segmentation algorithm. Since infants are thought to be able to detect utterance boundaries from very early on (e.g., Seidl & Cristià, 2008), utterance breaks are typically not removed. The algorithm then returns the same corpus with word boundaries added. This output and the original “gold” are compared, to determine to what extent the hypothesized words and their boundaries match those that were present in the original text.

Algorithms are often categorized into a few classes as a function of the primary segmentation cues they rely on (for a similar classification and further discussion, see Brent, 1999; Daland & Pierrehumbert, 2011). Models can additionally be classified

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

along other dimensions, including: whether the basic units in the input are phones or syllables; whether hierarchical structure is explicitly represented; and how experience is integrated. Importantly, there is no agreement on how several strategies of word segmentation might be weighted across development (for instance, see the debate between Johnson & Jusczyk, 2001; and Thiessen & Saffran, 2003), and no evidence on whether this varies as a function of language exposure. In view of this uncertainty, we follow Cristia, Dupoux, Ratner, and Soderstrom (2019) in reasoning that it is more appropriate to model infant word segmentation using a wide array of strategies, allowing us to focus on results that are stable across models. Next, we introduce these strategies, discussing plausibility of each in turn.

That said, we want to stress that we do not mean to argue in favour of any one strategy, as what is crucial to the present effort is that diverse strategies are represented.

### Baseline models

The simplest (baseline) models assume that infants treat every phrase, or every basic unit (a phone or a syllable), as a word. This strategy is very simple because it requires no memory–in fact, the input does not need to be processed in any way (other than perceiving syllable edges, and perceiving phrase edges).

Some believe these baseline models are plausible representations of what infants do, at least very early on. According to some developmentalists, memorizing whole utterances may be the only strategy available until late in the first year (DePaolis, Vihman & Keren-Portnoy, 2014; Keren-Portnoy, Vihman & Fisher, 2018). Such a strategy would not get infants very far, since many words do not occur in isolation: they may learn proper names like this, but not body part words (both of which appear in 6-month-olds' lexicon; Seidl, Tincoff, Baker & Cristia, 2015; Tincoff & Jusczyk, 1999, 2012). This is, of course, the limitation of these models, in that they will never suffice to learn the words of a language. Nonetheless, these baseline models may capture very early and uninformed strategies.

### Sublexically-driven models

Most attention by infant experimentalists has been devoted to strategies that build on very local indices, rather than optimizing the resulting lexicon. In the computational literature, sublexically-driven models are those that have as a goal to segment the input, drawing primarily from local cues. Among these models, we recognize two large subclasses, one that builds on sound co-occurrences (phonotactics) and another that builds on syllable co-occurrences.

Languages are typically described as having a small inventory of units (sounds or syllables), which are combined to form words. It is typically the case that only some combination of the basic units is legal, while other combinations are rare or illegal, particularly at word edges. For example, in English, the sequence tl can occur word-medially as in Atlantic, but not word-initially.

Phonotactic cues are a good source of information to word edges because they indicate sequences of sounds that are frequent or infrequent at certain syllable or word positions (Jusczyk, Luce & Charles-Luce, 1994). Evidence from American English learners suggests they can use phonotactics to extract word forms (Mattys, Jusczyk, Luce & Morgan, 1999).

Infants could learn phonotactics in a few ways. Once they have learned a few words (e.g., those that occur in isolation), they can extrapolate from that small lexicon. Alternatively, viewing phrase boundaries as word boundaries can be a useful strategy

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

while the proto-lexicon is being formed. Indeed, if the learner notices that a given sound sequence is more common between utterances than within utterances, then this is a very good indication that this sequence is an illegal word onset or offset in the language. In fact, there are many different slight variants in the computational modeling literature depending on the size of the window (whether two phones or more are considered), the position of the window (everywhere, focusing on utterance edges, focusing on phones straddling the phrase boundary, or extracting them from a proto-lexicon), and other details of implementation. For the purposes of the present study, what remains crucial to retain is that these models tend to be extremely successful in English (e.g., Daland & Pierrehumbert, 2011), with mixed evidence for some other languages (Russian: Daland, 2009; Korean: Daland & Zuraw, 2013).

A second type of statistical information has received even more attention than sequencing of phones, and that is the sequencing of syllables. Conceptually, this strategy is intermediate between sublexical and lexical approaches, relying on the intuition that when a sequence of syllables co-occurs frequently, then these syllables probably form a word. Nonetheless, we discuss it here because this intuition is commonly implemented via an algorithm that aims at positing boundaries: when a sequence of two syllables is infrequent, then one posits a boundary between them.

This strategy has been studied extensively in infancy (e.g., Aslin, Saffran & Newport, 1998; Pelucchi, Hay & Saffran, 2009; a review in Lany & Saffran, 2010; Saffran, Aslin & Newport, 1996), often (but not always) using artificial languages where all other cues (including phrase breaks) have been neutralised (a meta-analysis in Black & Bergmann, 2017). Importantly, these cues are also employed by bilingual infants: who can keep track of two sets of statistics presented in interleaved utterances better than monolingual infants (Antovich & Graf Estes, 2018, 2020).

Some computational research building on this idea (e.g., Gervain & Erra, 2012) employed an implementation that stayed close to Saffran's description of the process: when a probability of a transition between two syllables is lower than that of the neighboring syllables, then a boundary is posited. We will refer to this implementation as being based on a relative threshold: segmentation is decided by comparing a transition probability to others close by, and thus the threshold changes for every triad of syllables. This relative thresholding makes it impossible to segment very short utterances: if there are only three syllables, and thus two potential boundaries within the utterance, then neither can be the "local minimum". Notice that this strategy is not a problem in artificial languages where there are no sentence breaks, or in materials that have been produced to always have more than three syllables in a given sentence. The problem only emerges in natural, spontaneous language, where utterances can be of any length (including 2 or 3 syllables long).

Saksida, Langus, and Nespor (2017) proposed that infants can also use another threshold, based on the average transitional probabilities over the whole corpus. In a study with multiple languages, they show that segmentation scores are much better using this absolute threshold (for a systematic threshold exploration, see also Gervain & Erra, 2012). Additionally, Saksida *et al.* (2017) document considerable cross-linguistic variation, although the precise reasons behind this variation are yet to be elucidated.

Some connectionist models may be classified here because the model posits boundaries at specific points of the stream, typically when a unit in the network representing a boundary activates beyond a certain threshold (including Aslin *et al.*, 1998; Christiansen, Allen & Seidenberg, 1998). However, it is important to point out

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

that these models do not explicitly attempt to capture and use phonotactics or transitional probabilities between syllables; instead, when these and other properties (e.g., lexical stress in Christiansen *et al.*, 1998's study of English) are informative, then the network can successfully predict word edges.

### *Lexically-driven models*

We call lexical cues those that primarily serve the goal of finding the lexicon of word forms and/or which are based on the assumption that speech corpora contain a concatenation of words. Some research suggests that, already by 6 months of age, American English learners use words that function like proper names as wedges, using them to segment utterances including them (Bortfeld, Morgan, Golinkoff & Rathbun, 2005; Mersad & Nazzi, 2012), rendering lexical algorithms plausible at least in principle.

Lexically-driven algorithms are those that have as a primary goal to learn a set of "minimal word-level recombinable units" which are optimal to generate their input corpus. An early implementation of this idea was the Bayesian Minimum Description Length Chunker by Brent and Cartwright (1996), who documented that such a strategy is quite successful at segmenting American English child-directed speech. Goldwater, Griffiths, and Johnson (2009) generalized Brent and Cartwright (1996)'s proposal by showing that their implementation was formally equivalent to a unigram adaptor grammar, which in itself is a special case of adaptor grammars.

These models assume a "grammar", meaning a set of rewrite rules that may be as simple as three rules: 1) "All utterances are composed of words"; 2) "All words are composed of sounds"; 3) "Sounds are one of [a-z]". This simple grammar can generate the present text (and, in fact, any text using Latin orthography). Typical adaptor grammars used to model unsupervised, infant-like word segmentation are hierarchical, probabilistic and non-parametric (Johnson, Griffiths & Goldwater, 2007). These models are called "hierarchical" because the rules make reference to a hierarchy: utterances > words > sounds. Notice that, as a result, rules are actually equivalent to trees, and thus words can be described as sub-trees in this context. The type of adaptor grammar currently employed is always probabilistic, meaning that each rule or tree may come to have an associated probability of reuse. They are called "non-parametric" when, instead of using a finite set of parameters (e.g., a fixed set of sub-trees and their probabilities), they both generate sub-trees and estimate their probability.

Additional work using adaptor grammars has suggested that they are fairly successful at segmenting child-directed speech in various languages, including German, Spanish, Italian, Farsi, Hungarian, and Japanese (Phillips & Pearl, 2014), while still showing cross-linguistic differences as you would expect across languages where the syllable structure has different levels of complexity (Fourtassi, Börschinger, Johnson & Dupoux, 2013; Johnson, 2008; Loukatou, Stoll, Blasi & Cristia, 2018).

Many models of word segmentation may also be classified as relying primarily on lexical cues, even though they have an implementation that may not explicitly have a long-term storage (or lexicon), and they may not explicitly model lexical cues. Notably, models that are presented as "chunkers"–including TRACX (French & Cottrell, 2014; French, Addyman & Mareschal, 2011) MDLChunker (Robinet, Lemaire & Gordon, 2011), and the Retention & Recognition model (Alhama & Zuidema, 2017) – may be conceptually classified here because the creation of chunks (or word candidates) is their primary goal. Please note these models vary widely in

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

their implementation: for instance, relying on a neural network (French & Cottrell, 2014) or the symbolic sequential processing of input comparing it against an overtly defined long-term storage system (Alhama & Zuidema, 2017).

Before moving on, we discuss two of the many experiments that have been carried out using TRACX, namely Simulations 9 and 10 of French *et al.* (2011). Those Simulations studied the segmentation of what the authors describe as "bilingual microlanguages": in both Simulations, there are 27 words in each of two languages, resulting from the combination of three symbols occurring in initial position, three in medial, and three in final (e.g., in one language, [a b c] are initial, [d e f] medial, and [g h i] final, with adg, aeh, afi being 3 of 27 possible words in the language). In Simulation 9, symbols do not overlap across languages (i.e., language beta: [j k l] are initial, [m n o] medial, and [p q r] final), whereas in Simulation 10 they do (specifically language beta is [d e f] are initial, [a b c] medial, and [p q r] final). Results show that TRACX correctly classified the words into two separate groups, including for a subset of words that had been held out from the training. While these results are very interesting, the similarity between these "bilingual microlanguages" and input that bilingual infants may experience is not so much as to consider the question of bilingual segmentation closed.

### *Hybrid models*

Some models' structure allows them to benefit from both sublexical and lexical cues (including Monaghan & Christiansen, 2010; Perruchet & Peereman, 2004). For instance, PARSER (Perruchet & Peereman, 2004) generates chunks (sequences of sounds or syllables) randomly, selecting the most cohesive chunks, strengthening their memory trace with repetition and weakening it with forgetting. Forgetting is implemented not just through decay, but also through interference with items appearing in the input stream that have some, but not all, of the syllables present in the stored chunks. As a result, transitional probabilities present in the input directly affect the likelihood of a chunk being retained, and thus the performance of the model when segmenting an input stream. Another example is PUDDLE (Monaghan & Christiansen, 2010), introduced in more detail below, which breaks up the input stream if it recognizes in it a chunk in its long-term memory, except if the remainder of the sentence fails to respect the phonotactics that have been extracted from its long-term memory store.

## *The present study*

In this study, we aim to model bilingual word segmentation. Modeling is an interesting approach for our understanding of bilingual language acquisition for several reasons. The most salient one is that we can study a complex problem with an unparalleled degree of control over variables that could lead to confounds. Variables such as the amount of exposure to each of the languages, the prevalence of language switching, and the presence of accents can be adjusted by the experimenter in order to study effects on performance, while all else is held equal. It is especially useful to test whether universal mechanisms of language acquisition can be applied in contexts of great variation in the language environment, which is the case for bilingual children. Moreover, this methodology can easily be extended to several languages to measure specific phenomena cross-linguistically and as a function of specific language combinations. In this study we will look at Catalan, Spanish, and English to take into

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

account the role of language distance by creating one bilingual corpus featuring a similar pair of languages (Spanish + Catalan); and another containing a more distant pair of languages (Spanish + English).

Following much previous work, the input our computer simulations use is phonological transcripts generated from linguistic corpora representing natural, spontaneous speech occurring around monolingual infants. In their original format, the linguistic corpora used are orthographic transcriptions of natural interactions between monolingual infants and their caregivers at home. We then combine transcripts into a pair of generated monolingual corpora (by mixing transcripts within each language separately) and a matching bilingual corpus (by mixing transcripts across languages). The advantage of creating an artificial bilingual corpus from several monolingual transcriptions is that it protects our data from confounds (e.g., intrinsic differences in how different researchers define utterances across corpora; random variation in how talkative or lexically diverse the speech of different caregivers is).

We could further control the proportion of experience with each language by setting it at 50% so as to model a perfectly balanced bilingual (i.e., 50% Catalan + 50% Spanish; 50% Spanish + 50% English). Additionally, we varied the frequency of language switching while maintaining total proportions stable, by changing across the two languages every other utterance or every 100 utterances (see Figure 1 for an example). For example, in the Catalan and Spanish combination, the first type of switching would imply a Catalan utterance followed by a Spanish one, whereas in the second type of switching this would happen every 100 utterances (100 utterances in Catalan would be followed by 100 utterances in Spanish). We concatenated whole utterances from different languages, without mixing within utterances.

In general terms, two general hypotheses were postulated before the experiments were carried out:

### The confusion hypothesis

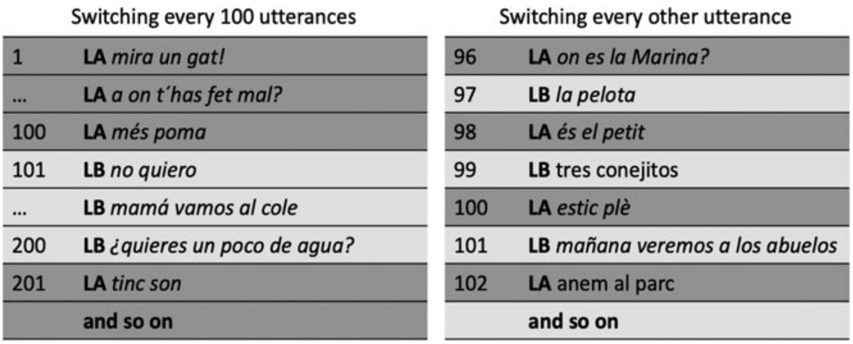
Bilinguals have less input of each language, thus reducing the chances of accumulating relevant data. Moreover, the models used here do not assume that there may be two distinct systems in the input, but just one. Thus, data are internally inconsistent. Based on this hypothesis, we predict that segmentation scores are lower for the bilingual corpus than both matching monolingual corpora.

### The robustness hypothesis

Since we only switch languages at utterance boundaries, information within those boundaries is always consistent. Thus, it may still be possible to generate a lexicon and a set of statistics that is appropriate for the bilingual input. Based on this hypothesis, we predict that the performance for the bilingual corpus overlaps with that of the matching monolingual corpora.

Before proceeding, we would like to make it clear that we do not claim to say that these algorithms represent faithfully what actual human infants do. Instead, the goal is to capture information in the input, and algorithms are one way to make sure that this information is plausibly useful. Therefore, the algorithms used here sample from the space of possibilities in terms of the type of algorithmic approach, with representatives of sublexical, lexical, and hybrid approaches.

However, we have not integrated every computational model that has ever been proposed, a topic to which we return in the Discussion. Additionally, we stress that we use multiple algorithms in order to capture a variety of strategies infants may be



Switching every 100 utterances

| | |
|---|---|
| 1 | **LA** *mira un gat!* |
| ... | **LA** *a on t´has fet mal?* |
| 100 | **LA** *més poma* |
| 101 | **LB** *no quiero* |
| ... | **LB** *mamá vamos al cole* |
| 200 | **LB** *¿quieres un poco de agua?* |
| 201 | **LA** *tinc son* |
| | **and so on** |

Switching every other utterance

| | |
|---|---|
| 96 | **LA** *on es la Marina?* |
| 97 | **LB** *la pelota* |
| 98 | **LA** *és el petit* |
| 99 | **LB** tres conejitos |
| 100 | **LA** *estic plè* |
| 101 | **LB** *mañana veremos a los abuelos* |
| 102 | **LA** anem al parc |
| | **and so on** |

**Figure 1.** Example of language switching every 100 utterances versus every other utterance for Spanish and Catalan. LA represents language A: in our study it could have been either Catalan or English (this example uses Catalan). LB represents language B: Spanish, in this case.

using; and our goal is not to compare the algorithms between them to say which is "better" or "worse"–for two important reasons. First, the comparison would be unfair because previous work that has more thoroughly explored these algorithms in English (e.g., Bernard, Thiolliere, Saksida, Loukatou, Larsen, Johnson, Fibla, Dupoux, Daland, Cao & Cristia, 2020) has found that their performance varies enormously as a function of different parameters that are used. Second, it is highly likely that the ranking in their performance varies as a function of language (and perhaps corpora) characteristics (Loukatou, Moran, Blasi, Stoll & Cristia, 2019). Therefore, any discussion of the algorithms' performance here is limited to providing a backdrop over which to interpret effects associated to bilingual exposure.

## Methods

This manuscript was produced using RMd (Baumer & Udwin, 2015) and Papaja (Aust & Barth, 2018) on R (R Core Team, 2019). The code, data, and other materials (other than code that is part of the WordSeg package, which is by Bernard & Cristia, 2018) are available from Fibla and Cristia (2019).

The processing and analysis of our data is performed with several steps. First, we select comparable monolingual corpora (Data). Then, orthographic transcriptions are converted into phonological form within each language (Phonologization). Next, we combine utterances across transcripts within language to generate the monolingual corpora, and across languages to generate the bilingual corpora (Concatenation). After removing word boundaries and representing the input in the appropriate unit, we separately apply each of the algorithms (Segmentation). Finally, the results of the algorithms are evaluated (Evaluation).

### *Data*

#### *Languages*

Catalan and Spanish are two rather similar Romance languages. The phonological system of Catalan is composed of seven vowels, /a ɛ e i ɔ o u/, and 25 consonants (in the central Catalan dialect used here). Catalan has many monosyllabic words that can end in a consonant cluster. The syllabic structure of Catalan is the following: there is an optional syllable onset that can include up to two consonants; an obligatory syllable nucleus, consisting of a vowel that can be optionally preceded and/or followed by a semivowel; and an optional syllable coda that can include one, two, or even three consonants.

The Spanish inventory consists of five vowel phonemes, /a e i o u/, and 21 consonants (in the Castilian dialect used here). The syllable structure is somewhat simpler: it consists of an optional syllable onset that can include one or two consonants; an obligatory syllable nucleus, consisting of a vowel that can optionally be preceded and/or followed by a semivowel; and an optional syllable coda, consisting of one or two consonants.

Finally, North American English (or Standard American English) is a West Germanic language. The variety of English we considered has 13 vowels and 25 consonants. English has generally complex phonotactics. For example, it allows complex consonant clusters permitting a syllable to start with up to three consonants (e.g., strict /strɪkt/) and complex codas, since a syllable can end with four consonants (e.g., sixths /sɪksθs/). English requires an obligatory syllable nucleus, consisting of a vowel or syllabic sonorant that can be optionally preceded and/or followed by a semivowel.

### Corpora

From the CHILDES database (MacWhinney, 2009), we selected corpora gathered during interactions between infants and their caregivers in a natural environment (such as the home) when children's ages ranged from ten months of age to 4 years of age. More specifically, the selected Catalan monolingual corpora (Bel, 1999; Llinàs-Grau & Coll-Alfonso, 2001; Miquel Serra, 1989) contained speech to 9 children (6 girls and 3 boys; age range 0;10–3;0). The Spanish corpora (López, 1997; López Ornat, Fernández, Gallo & Mariscal, 1994; Martínez, 1995; Miquel Serra, 1989; Vila, 1990) contained the transcriptions of speech to 6 children (3 girls and 3 boys; age range 0;10–3;0). The English corpus (Providence; Demuth, Culbertson & Alter, 2006) contains data from 6 children (3 girls and 3 boys; age range 0;11–4;0). Some transcripts were excluded because the transcript was shorter than 100 utterances (the minimum given our switching requirements, to be explained in the Concatenation section), or because we found sentences in the other languages.

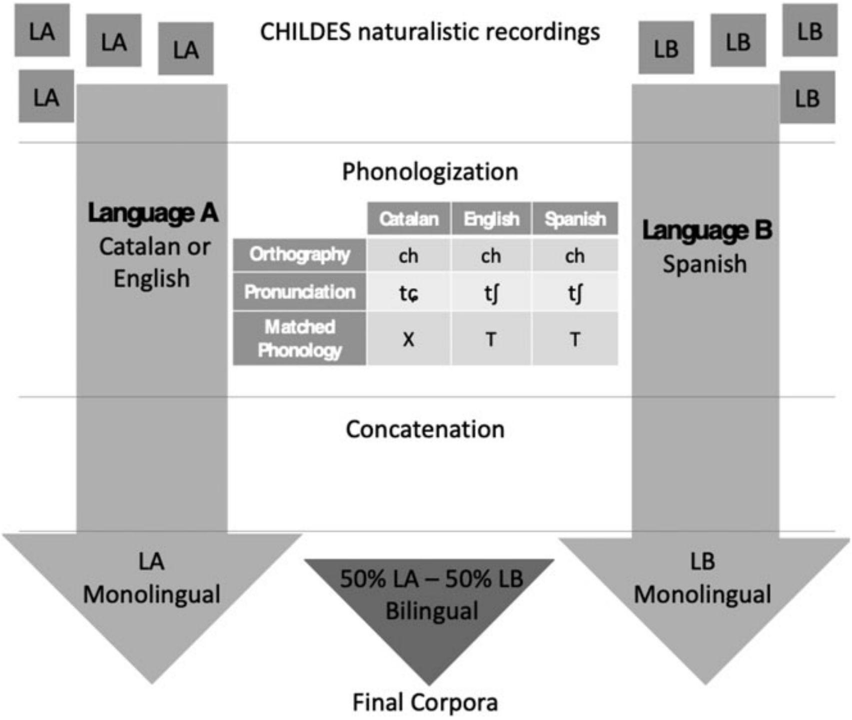
Whereas our pre-processing method (explained below) allows us to match the corpora perfectly in size, we could not control for other properties. In particular, lexical diversity and utterance length was not directly controlled across the original corpora, as this would have meant disturbing their naturalness, and we simply report on these characteristics. It must be borne in mind that this lack of matching may lead to differences across the monolingual corpora that are difficult to interpret (i.e., a difference between languages X and Y may reflect intrinsic differences in ease of segmentation, or a divergence in whether transcribers required few versus many indicators to decide that an utterance ended). However, the comparison between a pair of monolingual corpora and their corresponding bilingual corpus remains valid given the way the corpora are generated.

### Phonologization

Phonological representations were generated from orthographic transcriptions with methods chosen to ensure a good quality transcription. This step and the next one, concatenation, are represented in Figure 2.

For Catalan, we used the phonemizer package (Bernard, 2018) to call the text-to-speech tool eSpeak (Duddington, 2008), which contains a pronunciation

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press



| | Catalan | English | Spanish |
|---|---|---|---|
| Orthography | ch | ch | ch |
| Pronunciation | tɕ | tʃ | tʃ |
| Matched Phonology | X | T | T |

**Figure 2.** Phonologization and concatenation steps. LA represents language A: in our study, it could have been either Catalan or English. LB represents language B: Spanish, in this case.

dictionary complemented with orthography-to-phonology rewrite rules to deal with out of vocabulary items. Inspection revealed there were some systematic errors in the pronunciation, which were corrected using computational rewrite rules.

Given the transparency of Spanish orthography, the Spanish corpora were phonologized using computational orthography-to-phonology rewrite rules.

Some of the algorithms require syllables as the basic representation unit. For both Spanish and Catalan, syllabification was done through a script building on the Maximize Onset Principle (Phillips & Pearl, 2018), implemented in four steps as follows: 1) each sentence is parsed from right to left; 2) once a vowel nucleus is detected, the first consonant to the left of it is considered as a potential onset; 3) if this onset appeared as a possible onset in a list provided to the algorithm, then it is retained, and otherwise a syllable boundary is inserted; 4) the process continues with the consonant to the left of that one, until the maximal onset cluster is established.

For English, we chose the text-to-speech system FESTIVAL (Taylor, Black & Caley, 1998) in the phonemizer package, which was preferred over eSpeak because it provides syllable boundaries. This is particularly important given the fact that one of the baseline calculations uses syllables–hence, the advantage of needing them. FESTIVAL uses a dictionary where each word is looked up, and orthography-to-phonology rules for out-of-dictionary words.

Next, rewrite rules were applied to match the phonemes across corpora (i.e., all phonemes that are thought to be the "same" across Catalan and Spanish, or English and Spanish, are represented using the same character). For instance, if the voiceless postalveolar affricate sound /ʧ/ of the English word much /mʌʧ/ was tagged using T, the same tag was used for the same sound present in the Spanish word chocolate /ʧoko'late/ but not for the Catalan voiceless alveolo-palatal affricate sound /tç/ from the word fletxa /fletçə/ which was tagged using a different symbol (such as X).

### Concatenation

An artificial bilingual corpus was necessary so as to control for language exposure, frequency of language switching, and any potential confounds (differences in lexical complexity, sentence length, etc). Since language switching necessarily entailed switching to a different transcript, therefore disrupting the naturalness of the "conversation", we decided to apply the exact same switching procedure to the monolingual corpora.

The procedure was as follows.

Imagine that there are 4 transcripts: 2 Catalan (utterances A and B) and 2 Spanish (utterances p and q). In the 1-sentence switch condition, we switch after each sentence, such that the resulting corpora will be: AB (Catalan monolingual); pq (Spanish monolingual); and ApBq (Catalan–Spanish bilingual). As is clear in the example, this results in a bilingual corpus that is double the size of the monolingual corpora. To remove a confound (that would entail the assumption that a bilingual child hears twice as many utterances as a monolingual child), we divided the bilingual corpus into two, and considered only the first half. The characteristics of the three generated corpora are shown in Table 1.

To assess whether differences between algorithms and between conditions (monolingual, bilingual) were reliable, we divided each corpus into 10 equally-sized subparts. Although these 10 subparts are not completely independent (since perhaps the same transcript contributes to the end of one of the parts and the beginning of the other), they suffice to provide a minimum bound of variation. Any difference across conditions that is within this range of variation is not greater than the within-corpus variation, and thus cannot be seen as meaningful.

### Segmentation

Before describing the segmentation algorithms, we would like to reiterate that we do not want to convince readers that actual human infants are using one of these specific algorithms. Instead, our goal is to use computational algorithms to capture information that infants could plausibly capture as well. Therefore, algorithms presented here are representative of different approaches of infant processing. The plausibility of the approaches (at large, rather than specific implementations) was discussed in the Introduction. Bernard and Cristia (2018) released an open source package that facilitates this kind of informational study of word segmentation properties. We used this package for corpus preparation and segmentation. Fuller description of the algorithms and this package can be found in Bernard *et al.* (2020) We will report on results for 7 implemented algorithms (see Results section as well as Table 2 for a summary of the algorithms included in this work). Please note that each conceptual strategy could be implemented in a myriad ways, and here we only show some of these implementations. We do this to make sure our conclusions about bilingual exposure are robust to the algorithm; and not to argue for or against a given algorithm or implementation.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

**Table 1.** Properties of the Catalan "cat", Spanish "spa", English "eng", and bilingual corpora. Utts indicates the number of utterances, PSWU the percentage of utterances that were single words, WPU indicates the mean number of words per utterance, Tokens and Types refer to words, and MATTR is Moving Average Type to Token Ratio (window size of 10 tokens).

| | Utts | PSWU | WPU | Tokens | Types | MATTR |
|---|---|---|---|---|---|---|
| cat-cat | 29400 | 22 | 3.84 | 112836 | 5273 | 0.89 |
| spa-cat | 29400 | 20 | 4.02 | 118205 | 7117 | 0.937 |
| spa-spa | 29400 | 18 | 4.33 | 127286 | 6112 | 0.907 |
| eng-spa | 29400 | 22 | 3.87 | 113916 | 6425 | 0.953 |
| eng-eng | 29400 | 23 | 3.89 | 114374 | 3938 | 0.926 |

**Table 2.** Summary of the segmentation algorithms included in this work by Name: "utt" utterance baseline, "syll" syllable baseline, "ag" Adaptor Grammar, "dibs" Diphone Based Segmentation, "tprel" transitional probabilities with relative threshold, "tpabs" transitional probabilities with absolute threshold. Type indicates the class of algorithm. Unit indicates how the corpus was unitized: "n/a" not applicable,"syll" boundaries can only be posited between syllables, "phon" boundaries can be posited between phones.

| Name | Type | Unit | Description |
|---|---|---|---|
| utt | baseline | n/a | treat every utterance as a word |
| syll | baseline | Syll | treat every syllable as a word |
| ag | lexical | phon | find the word forms optimal to generate the corpus |
| dibs | sublexical | phon | break up sequences if low phonotactics |
| tprel | sublexical | syll | break up sequences if local TP minimum |
| tpabs | sublexical | syll | break up sequences if TP lower than global TP |
| Puddle | hybrid | phon | break up sequences if known subsequence found and what remains respects phonotactics, memorize otherwise |

### *Baselines*

Following Lignos (2012), we incorporate two simple segmentation strategies: consider every utterance as a word, and consider every syllable as a word.

### *Lexical algorithm*

We drew one algorithm from the class of lexical models, and more specifically from the Adaptor Grammar (AG) family (Goldwater *et al.*, 2009; Johnson *et al.*, 2007). The grammar we used is the simplest that can be represented in the context of word segmentation, and it contains two main rewrite rules (an Utterance is one or more Words; a Word is a sequence of Basic units), in addition to rules that write out basic units (e.g., Basic unit is a-z, for phonemes, or the list of syllables, when using syllable as unit).

This model is called "adaptor" because the system can learn specific instances of these rewrite rules. In our implementation, Word is the only unit that the model adapts from the input, meaning that in addition to this general rule there are also specific instances such as "dog is the sequence of phones d o g". This means that,

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

when provided with a sentence containing the sequence “dog”, the algorithm can try to model it as a sequence of phones by applying the general rules, or representing using the shortcut rule just described. We used WordSeg’s AG with its default parameters. As a result, 2000 sweeps of the segmentation over the corpus were done 8 times, pruning sweeps to avoid overfitting. Each pass induces a lexicon, a list of word types paired with the frequencies of occurrence in the corpus. Next, these different lexicons are used to generate likely segmentations of the corpora. Finally, Minimum Bayes Risk is used to adjudicate between alternative parses by finding the optimal set of rules.

### *Sublexical algorithms*

Two algorithms belong to the family of sub-lexical models. The first is the Diphone Based Segmentation, DiBS for short (Daland & Pierrehumbert, 2011). Among the three DiBS algorithms proposed by Daland (i.e., baseline-BiBS, phrasal-DiBS, lexical-DiBS), we selected the one that required the least amount of supervision: namely, phrasal-DiBS. Over the whole corpus, this version of DiBS calculates the general frequency of every phone bigram, as well as its frequency straddling a phrase boundary. The intuition behind this algorithm is: if most occurrences of this phone sequence straddle a boundary, then it is likely that these two phones cannot occur together; thus, when they do, this probably signals that the first phone belongs to one word, and the other phone belongs to another word. DiBS requires one more parameter to make the final decision of whether to segment or not. At present, this is one aspect of this algorithm that requires supervision, as this threshold is set by establishing the true likelihood of a word boundary (defined as the number of boundaries divided by the number of phones).

We sampled four implementations from the Transitional Probabilities (TP) family (Saksida *et al.*, 2017). Both rely on forward transitional probabilities on syllables. In addition to the TP based on a relative threshold that is most commonly discussed in the context of laboratory experiments, we also consider a TP algorithm based on an absolute threshold. We explain each using the example “baby cookie,” and focusing on the potential word boundary between the syllables “by” and “coo”. For both algorithms, the forward TP for “by + coo” is FTP (by + coo) = frequency (by + coo)/ frequency (by). Next, we must decide whether this FTP is low enough to warrant a boundary being placed between “by” and “coo”. For the relative case, we measure the FTP of each surrounding bigram (in this example, the FTP of “ba + by” and “coo + kie”), and we posit a boundary if both these FTPs are higher than that of the key bigram. For the absolute case, the threshold is the average FTP for all bigrams in the corpus.

### *Hybrid algorithm*

The last algorithm is hybrid, because it incorporates strategies from both sublexically-driven and lexically-driven models. It is ‘Phonotactics from Utterances Determine Distributional Lexical Elements’: PUDDLE, for short (Monaghan & Christiansen, 2010). PUDDLE builds a word form lexicon while using local cues to decide whether a given boundary is posited or not.

There are two long-term storage components: 1) a lexicon, and 2) a list of beginning and ending phoneme pairs, which are generated from the lexicon.

The model creates the lexicon as follows, considering one utterance at a time.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

Taking the first utterance, PUDDLE searches through it, from left to right, testing whether there is a subsequence in the utterance that is identical to any of the stored lexical items. Naturally, this will be false for the very first utterance in the corpus. Every time no subsequence match is found, then the whole utterance is entered into the lexicon, with an activation level of 1 (meaning it has been seen once). The algorithm then considers the next utterance. Let us imagine that the next utterance contains the first utterance–for instance, if the first utterance was "doggy" and the second "look at the doggy". In this case, there is a subsequence match. When this occurs, the system considers positing a boundary around that match–in this case: "look at the + doggy". This break will be posited only if the resulting subparts respect the phonotactics the system has extracted. To check this, the system extracts initial and final diphones of all items in the lexicon–in our example, "do" and "gy". The parse does not respect the learned phonotactics: because "lo" and "the", which are the initial and final diphones of "look at the", have not yet been observed. So the system again enters the whole utterance into the lexicon. Now, however, "lo" is a previously-observed diphone, which will eventually allow the system to potentially break up utterances starting with this diphone. Imagine that the next utterance is "doggy". This does match an item in the lexicon, and thus the activation of "doggy" is increased by one. The process continues precisely in this manner, with items being entered into the lexicon, and/or with their activation being increased, and initial and final diphones being used to filter out candidate parses.

As is clear from this description, PUDDLE shares features with both lexical and sublexical models. While PUDDLE does not overtly optimise minimal reusable units the way that AG systems do, it does store a lexicon and uses their frequency when parsing. Like sublexical models, PUDDLE takes into account coherence in statistical information of subword units. However, the specific function is quite different. For instance, whereas DiBS looks at the continuous frequency distribution of the phone preceding and following a potential break so as to divide a bigram, PUDDLE assesses whether the initial and final phone bigrams in a given parse have been observed.

### Evaluation

Previous work has used a variety of metrics. In the context of word segmentation, many have used a 'word versus partword' forced choice task (e.g., French *et al.*, 2011). While this is useful to compare against infant experiments, our goal here is not to argue for or against a given algorithm, but to describe how easily a corpus can be segmented given an algorithm. In this context, it seems less informative to use a handful of words and constructed partwords than to actually evaluate the whole corpus for segmentation accuracy.

Therefore, we adopt the Natural Language Processing/Speech Technology standard and use token recall and token precision (e.g., Ludusan, Versteegh, Jansen, Gravier, Cao, Johnson & Dupoux, 2014). This is also the approach adopted by previous work that attempts to compare the overall segmentability of different registers (child-versus adult-directed speech, Cristia *et al.*, 2019; Ludusan, Mazuka, Bernard, Cristia & Dupoux, 2017), and different languages (Caines, Altmann-Richer & Buttery, 2019; Loukatou, Stoll, Blasi & Cristia, 2018; Loukatou *et al.*, 2019), or simply evaluate proposed algorithms (e.g., Daland & Pierrehumbert, 2011; Goldwater *et al.*, 2009; Phillips & Pearl, 2014). These scores are calculated by comparing the output string, which contains hypothesized word breaks an algorithm supplies, against the original sentence containing word breaks. Token recall represents the proportion of words in the input that were recovered (i.e., number of words found divided by total number of words in the input). Token precision is the number of correct word tokens found out by the algorithm, divided by the number of word tokens hypothesized by the algorithm (i.e., what proportion of the algorithm words were real words). We multiply each by 100, in order to refer to it in percentage points rather than proportions. The scale in both scores goes from 0 = catastrophic segmentation, to 100 = perfect segmentation.

For instance, imagine that the original sentence was "a dog eats another dog" (5 word tokens), and the system returns "a dog eats a no therdog" (6 word tokens). There are 3 correct tokens (a, dog, eats), leading to a precision of 3/ 6 * 100 = 40% and a recall of 3/ 5 * 100 = 50%.

Based on token precision and recall, we can derive an F-score as the harmonic mean of the other two: 2*(precision*recall/(precision+recall)). F-score goes from 0 to 100%, with higher scores indicating better segmentation. For instance, in the example above, it is 2 * (40 * 50 / (40 + 50)) = 44.44%. This represents overall performance concisely, but does not capture potential precision-recall trade-offs.

Except for PUDDLE, none of the algorithms has a "learning curve". That is to say, provided with a corpus, they can process the whole corpus and return it containing hypothesized breaks, with no difference in performance for the beginning of the corpus compared to the end of the corpus. As a result, we can evaluate them on the whole corpus. PUDDLE, however, does learn over the course of exposure. To represent its performance at the end of the learning period, we run the algorithm five times. The first time, we present it with the corpus exactly as it is, and calculate precision and recall in the final 20% of utterances. We then take this final section, and move it to the beginning of the corpus, and then repeat the learning and evaluation, now evaluating on what originally was the 60–80% section of the corpus. This process is repeated three more times, resulting in an evaluation of the whole corpus, with the level of performance achieved when 80% of the corpus has been seen. This means PUDDLE has seen less of the corpus than the other algorithms. However, our goal here is not to compare the algorithms to each other, but rather we use multiple algorithms to assess the impact of algorithm choice on performance, and to check how different algorithms fare with bilingual as opposed to monolingual input. Therefore, we have prioritized comparability with previous work using WordSeg, which has always performed evaluation in this way, rather than changing the evaluation to occur based on holding out 20% of the dataset. We underline that this is not a problem because effects on performance by this extra experience are minimal (in the online materials at Fibla & Cristia, 2019, Section C shows that there are barely any differences when doing so versus the evaluation used here).

## Results

A first analysis checked for an effect of switching frequency, i.e., whether we changed transcript every sentence or every 100 utterances. A paired t-test on token F-scores revealed no difference between the two, both when considering all corpora together [t(34) = -1.26, p = 0.22] and when testing specifically the bilingual corpora [t(13) = -0.60, p = 0.56]. Henceforth, we focus on corpora put together by switching every sentence (see Table 3 for Token Precision and Recall). Section A of Fibla and Cristia (2019) provides more information on this.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

Figure 3 shows performance for the different algorithms in terms of F-score (for figures separating precision and recall and algorithms, please refer to section D of Fibla & Cristia, 2019). The error bars indicate 2 standard deviations across the ten subparts. If there is no overlap in these bars across two observations, we will say that they differ “significantly”. This word is thus not used in the sense of null hypothesis testing, but rather to indicate that the difference is greater than what might be seen due to a conservative measure of within-corpus noise. That is to say, if two such intervals overlap, we can be certain that the difference in performance is not meaningful. The opposite, however, does not follow: since our measure of noise is conservative, it could still happen that that two intervals do not overlap here, but are actually not meaningful with other noise estimations.

The most salient result on Figure 3 is probably that there are marked differences in performance across the algorithms, with a maximum F-score difference of 52.45. This has been observed in much previous work (e.g., Cristia *et al.*, 2019; Larsen, Cristia & Dupoux, 2017; Ludusan *et al.*, 2017), but it is interesting to bear this in mind when inspecting the size of the effect of other factors, so we describe their behaviour in some detail. The range of performance is nearly bounded by our two baseline algorithms, with the strategy of considering every utterance a word doing poorest (median F-score 8.91), and the strategy of considering each syllable as a word doing quite well (median F-score 48.52). Although we had expected the syllable-as-word strategy to do well in English child-directed speech, we were surprised by the Spanish and Catalan child-directed speech, as these are two languages in which linguistic descriptions do not suggest a necessary and strong alignment between syllable boundaries and word boundaries. Our results additionally suggest that, for three languages and two bilingual settings, sublexical approaches do allow a learner to find some words, but they also segment out items that are not words (low precision), and other words (low recall), resulting in intermediate F scores. An exception concerns the strategy using transitional probabilities between syllables with an absolute threshold (median F-score 54.78), which actually achieves much higher F-scores than the other sublexical approaches (DiBS median F-score 34.95; TP-rel median F-score 34.07). The hybrid algorithm achieved a level of performance similar to that of the sublexical algorihtms (median F-score 42.33). Better results are found with AG (median F-score 61.36), which is similar in performance to the syllable baseline, albeit with lower cross-language variability.

Next, we focus on the comparison among monolingual corpora, to determine whether the three monolingual corpora were equally easy or hard to segment. For ease of expression, we refer to the corpora on the basis of the language, but we remind readers that they could also vary on other characteristics by chance (e.g., the corpus producers’ definition of when utterances end).

Differences across monolingual corpora were sizable. Between Catalan and Spanish, the median difference (i.e., subtracting the Spanish score from the Catalan score within each algorithm separately) was 5.40% in precision, and 3.30% in recall, for an overall 4.34% difference in F-scores. Between English and Spanish, the median difference (i.e., subtracting the Spanish score from the English score within each algorithm separately) across algorithms was 5.87% in precision, and 1.76% in recall, for an overall 2.65% difference in F-scores. Overall, higher scores were obtained for Catalan and English than Spanish, and often significantly so. The only exception was DiBS, where Spanish scores are higher than those for Catalan (but not those for English) and TP-rel, where the trend is reversed, with higher scores for Spanish than Catalan.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

**Table 3.** Token Precision and Recall for the 7 algorithms, in the 5 language conditions. The acronyms stand for “eng” English, “spa” Spanish, “cat” Catalan, “utt” utterance baseline, “syll” syllable baseline, “ag” Adaptor Grammar, “dibs” Diphone Based Segmentation, “tpabs” transitional probabilities with absolute threshold, “tprel” transitional probabilities with relative threshold. Only the performance switching every utterance is shown, with overall length matched across the monolingual and the bilingual conditions.

| | Precision | | | | | Recall | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | eng_eng | eng_spa | spa_spa | spa_cat | cat_cat | eng_eng | eng_spa | spa_spa | spa_cat | cat_cat |
| utt | 23 | 6 | 22 | 6 | 18 | 4 | 20 | 5 | 22 | 6 |
| syll | 68 | 82 | 46 | 66 | 32 | 52 | 36 | 56 | 40 | 61 |
| ag | 67 | 53 | 63 | 59 | 61 | 61 | 63 | 63 | 67 | 65 |
| dibs | 47 | 31 | 40 | 24 | 46 | 32 | 43 | 29 | 40 | 28 |
| tpabs | 71 | 60 | 54 | 56 | 45 | 48 | 47 | 55 | 57 | 60 |
| tprel | 43 | 25 | 42 | 28 | 41 | 30 | 40 | 29 | 44 | 31 |
| puddle | 54 | 65 | 36 | 51 | 28 | 44 | 32 | 47 | 40 | 53 |

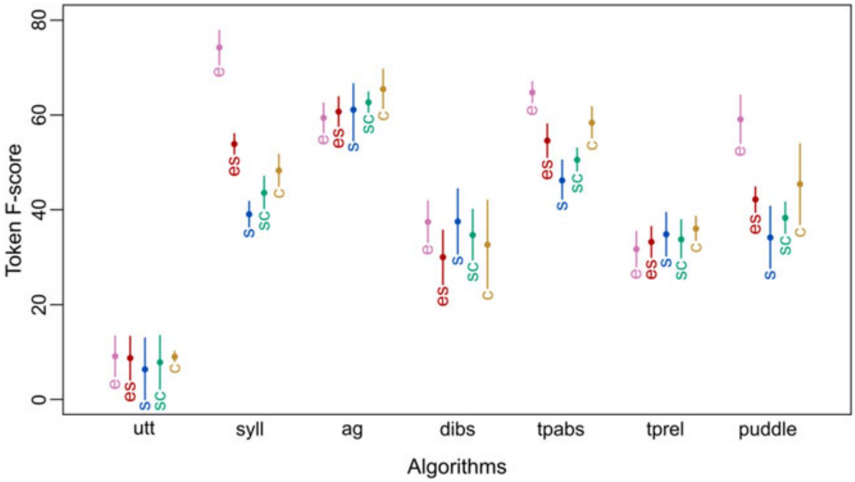


**Figure 3.** Token F-scores per algorithm (see Table 1 for acronym explanation) and corpus: pink “e” for English, brown “es” for English–Spanish, blue “s” for Spanish, green “sc” for Spanish–Catalan, gold “c” for Catalan. Error bars indicate 2 standard deviations over 10 subparts of the relevant corpus.

Finally, we turn to our key research question. We had predicted that, if there is a bilingual disadvantage, scores for the bilingual corpus should be below both of the matching monolingual corpora. In fact, this nearly never happens, as each bilingual corpus yields scores that are in-between those of the corresponding monolingual corpora. There is only one exception to this general pattern (see Fibla & Cristia, 2019, section D, for analyses separating precision and recall, as well as separating algorithms): DiBS for the English–Spanish bilingual corpus was lower than those for both English and Spanish monolingual, but with overlapping errors. In fact, if we calculate difference scores between two monolingual corpora and their corresponding bilingual corpus, the median difference across algorithms for the Catalan–Spanish case (i.e., subtracting the Spanish–Catalan score from the Catalan score and the Spanish score within each algorithm separately) was 1.56% in precision, 0.85% in recall, and 1.07% in F-scores. For the English–Spanish case, the median differences (i.e., subtracting the Spanish–English score from the English and the Spanish scores within each algorithm separately) were 1.00% in precision, 1.41% in recall, and 0.42% in F-scores. Notice incidentally that these differences are smaller than both algorithm and corpora differences discussed previously.

## Discussion

We set out to establish if word segmentation was necessarily degraded by bilingual exposure compared to monolingual matched exposure, or whether several segmentation models were robust to language switching. Following the confusion hypothesis, we had expected degradation to show up as scores for the bilingual corpus that were below scores for both of the corresponding monolingual corpora. Based on two bilingual corpora, we can clearly declare that this prediction does not bear out: it is nearly never the case that the bilingual corpora scores are below both

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

of the monolingual ones, with the median difference being only about 0.68% in F-scores. Importantly, these results ensued regardless of the degree of similarity between the languages combined in a bilingual corpus.

Further inspection of results revealed that the most important source of variance was algorithm, with the second most important source being language (or corpus) differences. Regarding the algorithm effect, we remind readers that the purpose of the present study was not to argue for or against a given algorithm, as it is an open question whether infants use some or all of them. Instead, we sampled from all relevant types to make sure that our conclusions were stable along this dimension, over which we have no real control. Similarly, differences across the monolingual corpora may reflect language differences, but could also be due to corpora differences (in e.g., the definition of utterance boundaries). What is crucial is the following: over the backdrop of algorithmic, cross-linguistic, and corpora variability, we can confidently state that the bilingual effect is smaller than all of these effects.

### *How generalizable may results be?*

In a nutshell, we found that bilingual exposure did not have a massive negative effect on word segmentation. Some readers may wonder, what would results be if we had assumed that infants knew they were learning more than one language, and were able to discriminate between them?

To answer this question, we can first refer to the current state-of-the-art knowledge on whether bilingual infants discriminate the two languages they are exposed to. Some of this research shows that infants (prenatally) exposed to bilingual input do sometimes discriminate their two languages. For example, Byers-Heinlein, Burns, and Werker (2010) showed that newborns prenatally exposed to two very different languages (English and Tagalog) have no difficulties in differentiating the two languages. Although to the best of our knowledge no research has investigated language differentiation in newborns prenatally exposed to Spanish and English, it is likely that these two languages can be differentiated at birth too. However, it is possible that some language pairs may not be discriminable by birth. No study has tested Spanish–Catalan discrimination in newborns (prenatally) exposed to both of these languages, but previous data suggests they may not be discriminated at birth (see Carbajal, Fér & Dupoux, 2016 for relevant modeling results). Moreover, bilingual language exposure does not affect language discrimination abilities, since bilingual young infants show language differentiation capacities that are similar to those reported in monolingually-exposed infants across ages and language pairs that are more versus less similar (e.g., Byers-Heinlein *et al.*, 2010: English–Tagalog newborns; Molnar, Gervain & Carreiras, 2014: Basque–Spanish 3.5 month-olds; Zacharachi & Sebastian-Galles, 2021: Spanish–Catalan 4–5 month-olds). That said, some evidence suggests that monolinguals and bilinguals may differentiate languages using different mechanisms (Bosch & Sebastian-Galles, 1997; Garcia, Guerrero-Mosquera, Colomer & Sebastian-Galles, 2018: Spanish–Catalan 4–5 month-olds). In sum, the two language pairs we used in our study are likely to represent two quite different starting points in bilingual language learning, one in which discrimination is possible in human newborns, and the other in which (at least based on current research) it is not affected. Thus, a first response to this question is that our approach is reasonable because some bilingual infants do not discriminate their two languages, at least early on.

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

As a second step, we approach the question logically: we found an unstable and minute difference between bilingual and monolingual exposure. Being able to discriminate the languages would have made the task, if anything, more similar to that of the monolinguals. But given that the difference between bilingual and monolingual artificial learners we observe is already negligible, any modeling effort that further attempts to assume language separation is bound to find exactly the same we already found. The opposite argument can also be made: if successful word segmentation crucially depended on initial language segregation, then a model that does not segregate should fail and show big differences between monolingual and bilingual model learners. Since we do not observe significant differences between monolingual and bilingual artificial learners (neither based on the language pair they are learning), perfect language segregation might not be essential for good segmentation. Thus, the results of our research provide important insights into the rather unexplored issue of how bilinguals start to build their two lexicons: if human infants exploit information similarly to our model learners, successful word segmentation does not depend crucially on initial language segregation.

Similarly, some readers may worry about the fact that our text-based representation does away with prosody, which may cue infants as to which language is being used in each sentence. Our reasoning was similar: had we "told" the models that each sentence had one or the other rhythm, then it would have been trivial if the bilingual model had performed comparably to the monolingual ones.

Our two design choices (not "telling" our model learners that there were two languages, and not providing them with prosody) made the problem more difficult for our models than what a human bilingual infant may expect. Indeed, bilingual infants may sometimes have additional cues, not only prosodic differences between the languages, but also contextual cues, whereby each language is used by a different set of talkers, at a different time or place, and/or when engaging in different routines. Our results show that, even without these additional cues, the informational cost of bilingual exposure is smaller than cross-linguistic or cross-corpora differences in word segmentation.

## *Limitations and future directions*

We see three main directions that we expect future research will engage in. First, one aspect of our results that we have not discussed in detail pertains to the fact that there were no differences between switching at every sentence, rather than every 100 utterances. This may relate to the fact that many of our algorithms capture information, rather than modeling in precise terms the process of segmentation actual infants may have. Specifically, we expect that models that rely on very local word repetition (for instance, by instantiating strong memory and salience constraints) will be more sensitive to the structure of conversations. In addition to exploring other algorithms, future studies could create corpora with other conditions not explored here, notably switching within utterances (including within words). We expect both sublexical and lexical approaches to suffer to a greater extent when switches can occur in these domains, but this is something that may be difficult to study in as controlled a manner as we have used here. A more tractable extension involves incorporating accented speakers, via phonological rewrite rules.

Second, some readers may feel that our choice of text-based modeling renders these studies implausible. Other computational modelers have argued that, since laboratory

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

experiments show infants can access a given cue, then a model based on the cue is indirectly shown to be cognitively relevant (a line of argument pursued, for instance, by Daland & Pierrehumbert, 2011; and Phillips & Pearl, 2014), and furthermore, recent research has added some more direct evidence for the relevance of text-based computational models through correlations with infants' early lexicon (Larsen *et al.*, 2017) and experimental results (Ngon *et al.*, 2013). While more such evidence would be welcome, computational models constitute an implementation of hypothesized algorithms, and can thus be used to provide proofs of concept.

Nonetheless, one may still argue that orthographic transcriptions do not capture the richness of the input that children hear, and thus a future direction may involve exploring monolingual and bilingual segmentation when finer grained details are considered. We think this is interesting in theory, as, for instance, it would get around the choice of whether phones or syllables are used as input. However, we think this will not work in practice, since there are several stumbling blocks standing in the way of using word segmentation models from the raw acoustic signal. The first and most salient one is that such models are greatly under-developed, and perform extremely poorly even with high quality recordings (Ludusan *et al.*, 2014). Second, for the study of bilingual acquisition in particular, controlling the input will be a lot more difficult. We would need to control for variables such as the number of speakers and the background noise, since current-day acoustic representations cannot abstract from such details the way humans do. An alternative may be to continue to use phonological transcriptions, which allow us to control for low-level variables, while implementing aspects that are of interest. For instance, one could simulate accented speech by incorporating sound mergers in certain talkers but not others.

Third, the most pressing avenue for research in modeling word segmentation involves studying more diverse languages (in the wake of Loukatou *et al.*, 2019). Current evidence suggests sizable differences across languages, and this although only a tiny fraction of the world's languages have been investigated. Similarly, it is unclear to what extent our results on bilingual language acquisition would generalize to languages that are more different. After all, all three languages studied here are Indo-European and therefore share many typological features. It would therefore be interesting to explore combinations that are likely to trick current word segmentation algorithms. For instance, we believe a combination such as Mandarin Chinese and Sesotho could be confusing because morphemes tend to be monosyllabic in the former but polysyllabic in the latter, and all syllables are simple in both languages. This should lead to a high level of ambiguity, with any polysyllabic sequence being ambiguous between a series of Mandarin words or a single Sesotho word. Combined with the rich, multilingual transcriptions of child-surrounding speech available from CHILDES (MacWhinney, 2009), the existence of an open-source, multi-platform, and easy to use word segmentation system (Wordseg; Bernard & Cristia, 2018) should greatly facilitate such extensions.

## Conclusions

This was the first paper to investigate bilingual word segmentation using computational modeling, an important tool to complement human infant research. Several simulations employed naturalistic corpora of child input to generate well-matched monolingual and bilingual corpora, which we then segmented with diverse cognitively-inspired algorithms. These simulations show that the bilingual cost in infant word

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

segmentation is smaller than the difference found across languages, which was in its stead smaller than that found across algorithms. We invite extensions exploring more diverse languages and language combinations, and additionally modeling other aspects of bilingual experience: in particular, within-utterance switching.

**Acknowledgements.** We are grateful to the families who agreed to be recorded and for their data to be shared, and the researchers who collected, transcribed, and shared the corpora – without them, this work would not be possible. We thank also the WordSeg team, particularly Mathieu Bernard and Manel Khentout, for implementing additional functions in the WordSeg package at our request. AC acknowledges the Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); and the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award. NSG acknowledges the European Commission Seventh Framework Programme (FP7/2007-2013): ERG grant agreement number 323961 (UNDER CONTROL) and the Catalan Government (SGR 2017–268).

**Competing interests.** The authors declare none.

## References

**Alhama, R. G., & Zuidema, W.** (2017). Segmentation as Retention and Recognition: the R&R model. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1531–1536). Austin, TX: Cognitive Science Society.

**Antovich, D. M., & Graf Estes, K.** (2018). Learning across languages: Bilingual experience supports dual language statistical word segmentation. *Developmental Science*, *21*(2), e12548. https://doi.org/10.1111/desc.12548

**Antovich, D. M., & Graf Estes, K.** (2020). One language or two? Navigating cross-language conflict in statistical word segmentation. *Developmental Science*, e12960. https://doi.org/10.1111/desc.12960

**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. https://doi.org/10.1111/1467-9280.00063

**Aust, F., & Barth, M.** (2018). papaja: Create APA manuscripts with R Markdown. Retrieved from https://github.com/crsh/papaja

**Bail, A., Morini, G., & Newman, R. S.** (2015). Look at the gato! Code-switching in speech to toddlers. *Journal of Child Language*, *42*(5), 1073–1101. https://doi.org/10.1017/S0305000914000695

**Baumer, B., & Udwin, D.** (2015). R markdown. Wiley Interdisciplinary Reviews: *Computational Statistics*, *7*(3), 167–177. https://doi.org/10.1002/wics.1348

**Bel, A.** (1999). *Teoria lingüística i adquisició del llenguatge. Anàlisi comparada dels trets morfològics en català i en castellà* (PhD thesis). Universitat Autònoma de Barcelona.

**Bergmann, C., & Cristia, A.** (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *9*(6), 901–917. https://doi.org/10.1111/desc.12341

**Bernard, M.** (2018). Phonemizer. Retrieved from https://github.com/bootphon/phonemizer

**Bernard, M., & Cristia, A.** (2018). Wordseg (Version 0.7). Retrieved from https://github.com/bootphon/wordseg/tree/master/wordseg.

**Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X.-N., & Cristia, A.** (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, (52), 264–278. https://doi.org/10.3758/s13428-019-01223-3

**Black, A., & Bergmann, C.** (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 124–129). Austin, TX: Cognitive Science Society.

**Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K.** (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

**Bosch, L., Figueras, M., Teixidó, M., & Ramon-Casas, M.** (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: Evidence from infants acquiring syllable-timed languages. *Frontiers in Psychology*, *4*, 106. https://doi.org/10.3389/fpsyg.2013.00106

https://doi.org/10.1017/S0305000921000568 Published online by Cambridge University Press

**Bosch, L., & Sebastian-Galles, N.** (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69. https://doi.org/10.1016/S0010-0277(97)00040-1

**Brent, M. R.** (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*(8). https://doi.org/10.1016/S1364-6613(99)01350-9

**Brent, M. R., & Cartwright, T. A.** (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1), 93–125. https://doi.org/10.1016/S0010-0277(96)00719-6

**Brent, M. R., & Siskind, J. M.** (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. https://doi.org/10.1016/S0010-0277(01)00122-6

**Byers-Heinlein, K., Burns, T. C., & Werker, J. F.** (2010). The roots of bilingualism in newborns. *Psychological Science*, *21*(3), 343–348. https://doi.org/10.1177/0956797609360758

**Caines, A., Altmann-Richer, E., & Buttery, P.** (2019). The cross-linguistic performance of word segmentation models over time. *Journal of Child Language*, *46*(6), 1169–1201. https://doi.org/10.1017/S0305000919000485

**Carbajal, M. J., Fér, R., & Dupoux, E.** (2016). Modeling language discrimination in infants using i-vector representations. In *Proceedings of cognitive science*.

**Christiansen, M. H., Allen, J., & Seidenberg, M. S.** (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268. https://doi.org/10.1080/016909698386528

**Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M.** (2019). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, *3*, 13–22. https://doi.org/10.1162/opmi_a_00022

**Daland, R.** (2009). *Word segmentation, word recognition, and word learning: A computational model of first language acquisition* (PhD). 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 *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 873–877). Retrieved from https://aclanthology.org/P13-2151.pdf

**Demuth, K., Culbertson, J., & Alter, J.** (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, *49*(2), 137–173. https://doi.org/10.1177/00238309060490020201

**DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T.** (2014). When do infants begin recognizing familiar words in sentences? *Journal of Child Language*, *41*(01), 226–239. https://doi.org/10.1017/S0305000912000566

**Duddington, J.** (2008). ESpeak. Retrieved from http://espeak.sourceforge.net

**Fibla, L., & Cristia, A.** (2019). Supplementary materials to: Is there a bilingual disadvantage for word segmentation? A computational modeling approach. https://doi.org/10.17605/OSF.IO/QGK9M

**Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E.** (2013). Whyisenglishsoeasytosegment. In *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (cmcl)* (pp. 1–10). Retrieved from https://aclanthology.org/W13-2601.pdf

**Frank, S. L., Monaghan, P., & Tsoukala, C.** (2019). Neural network models of language acquisition and processing. In P. Hagoort (Ed.), *Human language: From genes and brains to behavior* (pp. 277–291). Cambridge, MA: MIT Press.

**French, B., & Cottrell, G.** (2014). TRACX 2.0: A memory-based, biologically-plausible model of sequence segmentation and chunk extraction. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36). Retrieved from https://escholarship.org/uc/item/0zv7096m

**French, R. M., Addyman, C., & Mareschal, D.** (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614. https://doi.org/10.1037/a0025255

**Garcia, L. N., Guerrero-Mosquera, C., Colomer, M., & Sebastian-Galles, N.** (2018). Evoked and oscillatory eeg activity differentiates language discrimination in young monolingual and bilingual infants. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-20824-0

**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. https://doi.org/10.1016/j.cognition.2012.06.010

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

**Gonzalez-Barrero, A. M., Schott, E., & Byers-Heinlein, K.** (2020). Bilingual adjusted vocabulary: A developmentally-informed bilingual vocabulary measure. *PsyArXiv*. https://doi.org/10.31234/osf.io/x7s4u.

**Höhle, B., Bijeljac-Babic, R., & Nazzi, T.** (2020). Variability and stability in early language acquisition: Comparing monolingual and bilingual infants' speech perception and word recognition. *Bilingualism: Language and Cognition*, *23*(1), 56–71. https://doi.org/10.1017/S1366728919000348

**Ibáñez, M. T.** (2017). *Constraints on early word segmentation and mapping* (PhD thesis). Universitat de Barcelona.

**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*, 548–567. https://doi.org/10.1006/jmla.2000.2755

**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 Phonology* (pp. 20–27). Retrieved from https://aclanthology.org/W08-0704.pdf

**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* (pp. 641–648).

**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. https://doi.org/10.1006/cogp.1999.0716

**Jusczyk, P. W., Luce, P. A., & Charles-Luce, J.** (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*(5), 630–645. https://doi.org/10.1006/jmla.1994.1030

**Keren-Portnoy, T., Vihman, M., & Fisher, R. L.** (2018). Do infants learn from isolated words? An ecological study. *Language Learning and Development*, 1–17. https://doi.org/10.1080/15475441.2018.1503542

**Lany, J., & Saffran, J. R.** (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284–291. https://doi.org/10.1177/0956797609358570

**Larsen, E., Cristia, A., & Dupoux, E.** (2017). Relating unsupervised word segmentation to reported vocabulary acquisition, In *INTERSPEECH*, pp. 2198–2202. 2017. http://dx.doi.org/10.21437/Interspeech.2017-937

**Lignos, C.** (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the West Coast Conference on Formal Linguistics* (Vol. 30, pp. 13–15). Retrieved from http://www.lingref.com/cpp/wccfl/30/paper2821.pdf

**Llinàs-Grau, M., & Coll-Alfonso, M.** (2001). Telic verbs in early Catalan. *Probus*, *13*(1), 69–79. https://doi.org/10.1515/prbs.13.1.69

**López, A. I. O.** (1997). Categorías funcionales y adquisición de la primera lengua: Un análisis contrastivo. *Revista Española de Lingüística*, *27*(2), 425–446.

**López Ornat, S., Fernández, A., Gallo, P., & Mariscal, S.** (1994). *La adquisición de la lengua española.* Madrid: Siglo XXI.

**Loukatou, G., Stoll, S., Blasi, D., & Cristia, A.** (2018). Modeling infant segmentation of two morphologically diverse languages. In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN* (pp. 47–60). Retrieved from https://aclanthology.org/2018.jeptalnrecital-long.4

**Loukatou, G. R., Moran, S., Blasi, D., Stoll, S., & Cristia, A.** (2019). Is word segmentation child's play in all languages? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3931–3937.

**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)* (pp. 178–183). https://doi.org/10.18653/v1/P17-2028

**Ludusan, B., Seidl, A., Dupoux, E., & Cristia, A.** (2015). Motif discovery in infant-and adult-directed speech. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning* (pp. 93–102). Retrieved from https://aclanthology.org/W15-2413.pdf

**Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., & Dupoux, E.** (2014). Bridging the gap between speech technology and natural language processing: An evaluation toolbox for term discovery systems. *Association for Computational Linguistics*, 560–576.

**MacWhinney, B.** (2009). *The CHILDES project part 2: The database*. New York: Psychology Press.

**Martínez, C. A.** (1995). *La adquisición de las categorías gramaticales en español* (PhD thesis). Universidad Autónoma de Madrid.

**Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L.** (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494. https://doi.org/10.1006/cogp.1999.0721

**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. https://doi.org/10.1080/15475441.2011.609106

**Miquel Serra, R. S.** (1989). *Diferencias individuales en la adquisición del léxico inicial* (PhD thesis). Universitat de Barcelona.

**Molnar, M., Gervain, J., & Carreiras, M.** (2014). Within-rhythm class native language discrimination abilities of Basque-Spanish monolingual and bilingual infants at 3.5 months of age. *Infancy*, *19*(3), 326–337. https://doi.org/10.1111/infa.12041

**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. https://doi.org/10.1017/S0305000909990511

**Monaghan, P., & Rowland, C. F.** (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, *67*(S1), 14–39. https://doi.org/10.1111/lang.12221

**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. https://doi.org/10.1111/j.1467-7687.2012.01189.x

**Oakes, L. M.** (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

**Orena, A. J., Byers-Heinlein, K., & Polka, L.** (2019). What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science*, e12901. https://doi.org/10.1111/desc.12901

**Orena, A. J., & Polka, L.** (2019). Monolingual and bilingual infants' word segmentation abilities in an inter-mixed dual-language task. *Infancy*, *24*(5), 718–737. https://doi.org/10.1111/infa.12296

**Pelucchi, B., Hay, J. F., & Saffran, J. R.** (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685. https://doi.org/10.1111/j.1467-8624.2009.01290.x

**Perruchet, P., & Peereman, R.** (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics*, *17*(2–3), 97–119. https://doi.org/10.1016/S0911-6044(03)00059-9

**Phillips, L., & Pearl, L.** (2014). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)* (pp. 9–13). Retrieved from https://aclanthology.org/W14-0503.pdf

**Phillips, L., & Pearl, L.** (2018). Syllabic bayesian word segmenter. Retrieved from https://github.com/lawphill/phillips-pearl2014.

**Polka, L., Orena, A. J., Sundara, M., & Worrall, J.** (2017). Segmenting words from fluent speech during infancy–challenges and opportunities in a bilingual context. *Developmental Science*, *20*(1), e12419. https://doi.org/10.1111/desc.12419

**Polka, L., & Sundara, M.** (2003). Word segmentation in monolingual and bilingual infant learners of English and French. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 1021–1024).

**R Core Team.** (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

**Robinet, V., Lemaire, B., & Gordon, M. B.** (2011). MDLChunker: A MDL-based cognitive model of inductive learning. *Cognitive Science*, *35*(7), 1352–1389. https://doi.org/10.1111/j.1551-6709.2011.01188.x

**Roy, D. K., & Pentland, A. P.** (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113–146. https://doi.org/10.1207/s15516709cog2601_4

**Saffran, J. R., Aslin, R. N., & Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

**Saksida, A., Langus, A., & Nespor, M.** (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, *20*(3), e12390. https://doi.org/10.1111/desc.12390

**Sebastian-Galles, N., & Santolin, C.** (2020). Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology*, *2*, 47–68. https://doi.org/10.1146/annurev-devpsych-013119-023724

**Seidl, A., & Cristià, A.** (2008). Developmental changes in the weighting of prosodic cues. *Developmental Science*, *11*(4), 596–606. https://doi.org/10.1111/j.1467-7687.2008.00704.x

**Seidl, A., Tincoff, R., Baker, C., & Cristia, A.** (2015). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental Science*, *18*(1), 155–164. https://doi.org/10.1111/desc.12182

**Sitaram, S., Chandu, K. R., Rallabandi, S. K., & Black, A. W.** (2019). A survey of code-switched speech and language processing. arXiv Preprint arXiv:1904.00784.

**Taylor, P., Black, A. W., & Caley, R.** (1998). The architecture of the FESTIVAL speech synthesis system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, 147–151.

**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. https://doi.org/10.1037/0012-1649.39.4.706

**Tincoff, R., & Jusczyk, P. W.** (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175. https://doi.org/10.1111/1467-9280.00127

**Tincoff, R., & Jusczyk, P. W.** (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, *17*(4), 432–444. https://doi.org/10.1111/j.1532-7078.2011.00084.x

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

**Zacharaki, K., & Sebastian-Galles, N.** (2021). The ontogeny of early language discrimination: Beyond rhythm. *Cognition*, *213*, 104628. https://doi.org/10.1016/j.cognition.2021.104628

**Cite this article:** Fibla L, Sebastian-Galles N, Cristia A (2022). Is there a bilingual disadvantage for word segmentation? A computational modeling approach. *Journal of Child Language* **49**, 1119–1146. https://doi.org/10.1017/S0305000921000568