Diversity and learnability in early language acquisition : across languages and cultures

Georgia Loukatou

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

Diversité et capacité d'apprentissage pour l'acquisition précoce des langues : à travers des langues et des cultures

Soutenue par
Georgia LOUKATOU
Le 25 Septembre 2020

Ecole doctorale n° 158
Ecole Doctorale Cerveau Cognition Comportement

Spécialité
Sciences cognitives – option intelligence artificielle

Composition du jury :

Riccardo, FUSAROLI Professor, Aarhus University	*Président*
Lisa, PEARL Professor, University of California Irvine	*Rapporteur*
Jill, LANY Assistant Professor, University of Liverpool	*Rapporteur*
Caroline ROWLAND Director, MPI for Psycholinguistics	*Examinateur*
Katerina, PALASIS Maître de Conférences, Université Nice	*Examinateur*
Mohamed, CHETOUANI Professor, UPMC	*Examinateur*
Alejandrina, CRISTIA Professor, ENS	*Directeur de thèse*

ENS DEC DÉPARTEMENT D'ÉTUDES COGNITIVES

Diversity and learnability in early language acquisition: across languages and cultures

Georgia Loukatou

Doctoral thesis supervised by Alejandrina Cristia

Laboratoire de Sciences Cognitives et Psycholinguistique
ENS-EHESS-CNRS-PSL University

Abstract

In this dissertation, we take a closer look at the astonishing diversity of input children grow up hearing all around the globe, and we ask how this diversity matters to language acquisition. For this, we employ interdisciplinary methods. We consider the type of language and culture as two principal sources of diversity, and we investigate them in two distinct parts of the dissertation.

Previous studies on language learning have focused mainly on English, and there is much less information on how other languages are learned. However, languages vary a lot to each other. Across languages, our first goal is to describe the nature of children's input and to identify its diversifying characteristics. For example, children learning Chintang are exposed to a polysynthetic language, with a particularly rich morphological system. How does this input compare to input from a language with a simpler morphological system, such as Japanese? Our second goal is to comprehend the relation between this diversity and learnability. We examine learnability in the context of language segmentation, a fundamental learning task. We assess how informative input is, and whether learning is affected by the characteristics described above. We further ask whether some cognitive strategies are viable across cross-linguistic environments.

To answer these questions, we conduct extensive analyses of ambient input in largely diverse environments. First, we retrieve this input from databases of longitudinal recordings. One such database is AcqDiv, which contains longitudinal recordings of caregiver-child interactions across eight languages differing in morphosyntactic features. We estimate robust descriptive measures of input quality, such as its lexical and morphosyntactic diversity. Second, we implement artificial language and modeling experiments. These methods allow us to inspect the learnability properties and segmentability of different kinds of input speech. We argue that segmentation of words, but also of other meaningful units, such as morphemes, should be considered when learning language. Moreover, we investigate whether previously proposed learning strategies for word segmentation perform above chance and stably for the AcqDiv languages.

In the second part of this dissertation, we look at differences across speech registers, speakers and cultural norms. Previous studies on early language learning focused mainly on child directed speech, mother-child interactions and WEIRD cultures. However, this input is not the only one children get exposed to when learning language. Across cultures, our first goal is to describe the nature of children's input and to identify diversifying characteristics. For example, how does input of children

learning Sesotho, who are mostly addressed by other children and receive little child-directed input from adults, compare to that of French-learning children? How does input overheard by French children differ from input directed to them? Our second goal is to study the relation between this diversity and learnability.

To answer these questions, we use data from longitudinal recordings such as the LENA-Lyon, the Demuth and other CHILDES corpora. First, we quantify the relative contribution of speech registers and speakers in Sesotho, French and English-learning children's overall input, and further compare this input based on corpus statistics. Second, making use of well-established segmentation models, we provide key insights on the segmentability of French overheard and child-directed input. We assess how informative child directed input is compared to overheard input, and whether segmentability differences between the two can be explained by their characteristics.

Résumé

Dans ce manuscript, nous examinons de plus près l'étonnante diversité d'input que les enfants grandissent en entendant, et nous demandons en quoi cette diversité est importante pour l'acquisition du langage. Pour cela, nous utilisons des méthodes interdisciplinaires. Nous considérons le type de langue et de culture comme deux sources principales de diversité, et nous les étudierons dans deux parties distinctes de cette thèse.

Des études précédentes se sont principalement concentrées sur l'apprentissage de l'anglais, et il y a beaucoup moins d'informations sur comment les autres langues sont apprises. Cependant, les langues varient étonnamment les unes des autres. Pour des langues différentes, notre premier objectif est d'identifier les caractéristiques de diversification de l'input. Par exemple, les enfants apprenant le chintang sont exposés à un langage polysynthétique, avec un système morphologique particulièrement riche. Comment cet input se compare-t-il à l'input d'une langue plus simple, comme le japonais? Notre deuxième objectif est de comprendre la relation entre cette diversité et la capacité d'apprentissage. Nous examinons cette capacité pour la segmentation du langage. Nous évaluons à quel point l'apprentissage des langues est affecté par ses caractéristiques. Nous nous demandons en outre si certaines stratégies cognitives sont viables dans des environnements multilingues.

Pour répondre à ces questions, nous consultons l'input de bases de données d'enregistrements longitudinaux. L'une de ces bases est AcqDiv, avec des interactions soignant-enfant dans huit langues dont les caractéristiques morphosyntaxiques diffèrent. Nous mesurons la qualité de l'input, sa

diversité lexicale et morphosyntaxique. Deuxièmement, nous mettons en œuvre des expériences de langage artificiel et de modélisation. Nous soutenons que la segmentation des mots, mais aussi d'autres unités significatives, comme les morphèmes, devrait être considérée lors de l'apprentissage. De plus, nous étudions si des stratégies d'apprentissage statistique précédemment proposées fonctionnent de manière stable pour les langues AcqDiv.

Dans la deuxième partie de la thèse, nous examinons les différences de l'input entre les registres vocaux, les locuteurs et les normes culturelles. Les études antérieures sur l'apprentissage précoce des langues se sont concentrées sur la parole dirigée à l'enfant, les interactions mère-enfant et les cultures WEIRD. Cependant, cet input n'est pas le seul auquel les enfants sont exposés lorsqu'ils apprennent la langue. À travers les cultures, notre premier objectif est d'identifier les caractéristiques de diversification. Par exemple, comment l'input des enfants apprenant le sésotho, qui provient principalement d'autres enfants et pas de la part des adultes, se compare-t-il à celui des enfants apprenant le français? En quoi l'input entendu, mais pas directement adressé aux enfants français diffère-t-il de l'input qui leur est adressé? Notre deuxième objectif est de comprendre la relation entre cette diversité et la capacité d'apprentissage.

Nous répondons à ces questions par des données issues d'enregistrements longitudinaux tels que LENA-Lyon, Demuth et autres corpus CHILDES. Premièrement, nous quantifions la contribution des registres et des locuteurs à l'input total des enfants apprenant le sésotho et le français, en utilisant des statistiques de corpus. Deuxièmement, en utilisant des modèles de segmentation bien établis, nous fournissons des informations clés sur la segmentabilité de l'input français destiné ou pas aux enfants. Nous comparons des deux, et nouns enquêtons si les différences de segmentabilité entre les deux types d'input peuvent être expliquées par leurs caractéristiques.

Acknowledgements

I should start by acknowledging my great supervisor, Alex Cristia. Alex, I am forever grateful for your guidance and support throughout the course of this work, and for all your personal advice and care. Thank you for believing in me, you are the best supervisor one could ask for.

A further thanks is owed to my brilliant colleagues in the Laboratoire des Sciences Cognitives and Psycholinguistique, who provided such a friendly atmosphere to work, Ava Guez, Camila Scaff, Naomi Havron, Sho Tsuji, Si Berrebi, Alice Latimier, Gerda Ana Melnik, Camille Williams, Monica Barbir, Cecile Issard, Cecile Crimon, Hualin Xiao, Mireille Babineau, Leticia Schiavon Kolberg.

I want to thank the amazing staff, Radhia Achheb, Catherine Urban, Michel Dutat, Vireack Ui, Anne Caroline Fievet, Isabelle Brunet. I am particularly thankful to Mathieu Bernard and Julien Karadayi for all their patience and assistance with technical matters.

I have also benefited from discussions with Jill Lany and Lisa Pearl, my PhD tutors, on all the topics contained in this thesis and more. I would also like to thank the members of my committee, Jill Lany, Riccardo Fusaroli, Caroline Rowland, Katerina Palasis, Mohamed Chetouani. I genuinely enjoyed our discussion, and your commemts have been extremely valuable.

I also wish to thank all my co-authors. Thank you Sabine Stoll, Damian Blasi, Steven Moran and Marie-Therese Le Normand for allowing me to work with your amazing data. A special thanks goes to the families and children that were recorded.

This thesis could not have been written without the care of friends and family. I wish to thank my parents for their encouragement, and Wassim for the love and support that helped me through the most difficult times.

This research was funded by a studentship from the Ecole doctorale Cerveau, cognition, comportement.

Diversity and learnability in early language acquisition: across languages and cultures

General Introduction

1. Early Language Acquisition

We begin the dissertation with an overarching chapter, where we present the benchmarks of early language acquisition, and we introduce some general concepts that will be discussed later on, in Part 1 and 2.

Early language acquisition refers to children's acquisition of their native language(s). All typically developing children acquire the ambient language. A careful observer cannot miss their outstanding learning progress, their increasing ability to comprehend and understand new aspects of their language. We may say that describing acquisition benchmarks is a view of early language acquisition *from the outside*. When we start wondering how this acquisition takes place, what are the processes that allow is to happen, we may say that we view early language acquisition *from the inside*.

1.1 From the outside

Language acquisition takes place early on. By the age of 3, children are competent speakers, producing "novel sentences that involve complicated constructions, words that reference abstract ideas or absent entities, sound sequences that mark the distinctive contrast of the native language" (Gierut, 2007, p.1). Observable milestones have been identified for the acquisition of words and morphemes by English-learning children. We describe some of these below (*Figure 1*).

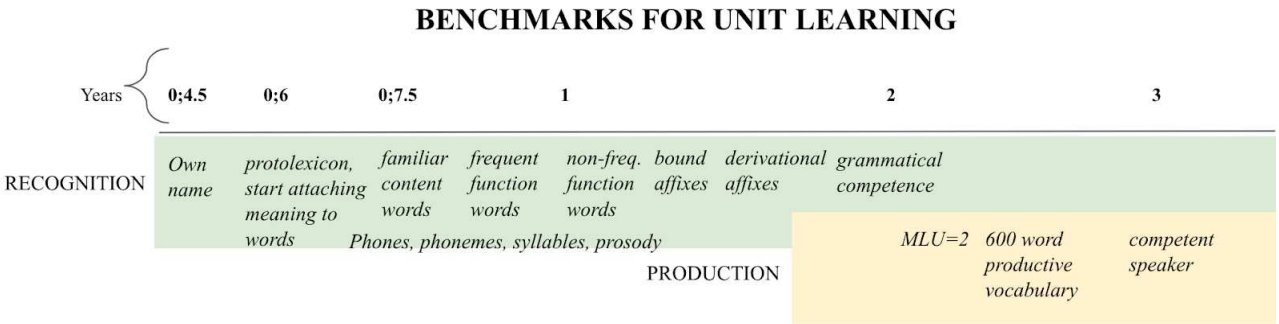

Figure 1. Indicative benchmarks proposed to describe English word and morpheme learning during language acquisition. Green refers to recognition and orange production.

By the end of the first year, children can already parse input speech of their native language (Pierrehumbert, 2003). They discover its phonemes (Swingley, 2009), phones (Pierrehumbert, 2003) and syllables (Bijeljac-Babic et al., 1993). They attune their innate sensitivity to acoustic variation to

the language of their environment (Kuhl et al., 1992). They can perceive the prosody of their language (Christophe et al., 2008) and its phonological phrases (Gout et al., 2004).

During their first year of life, children also start breaking speech into word-like units, in a task often called *word segmentation*. This task is challenging, and the fact that they accomplish it fast and effortless is astonishing for several reasons. First, speech is continuous, with no acoustic correlates to word boundaries (no 'white spaces' between words, as in some writing systems). We can notice that speech is continuous when we hear someone talking in an unknown language. Second, at the beginning of acquisition, children do not dispose of a word lexicon and do not know the regularities of their future language (Blanchard et al., 2009), so they do not have any knowledge specific to the language ready to help. Third, very few words occur in isolation in speech (Brent & Siskind, 2001), and, for those words that occur in isolation, there is no proposal on how they could be recognized (Gambell & Yang, 2006). Most words are never even heard in isolation (determiners for example), thus children should somehow *segment* them out of speech.

We can observe traces of segmentation by the age of 7 months, by examining children's recognition of words and functional items. For example, children recognize their own names by 4.5 months (Mandel et al., 1995), and segment familiar content words such as *cup* and *feet* from running speech by 7.5 months (Depaolis et al., 2014; Jusczyk & Aslin, 1995). Around their first year and prior to production, children also recognize frequent functional items (Christophe et al., 2008), such as determiners (by 11 months, (Shi, Werker, et al., 2006) and even less frequent function words (by 13 months, Shi, Cutler, et al., 2006). Prior to production, and maybe even prior to semantic knowledge, they recognize bound items such as affixes (Gomez & Gerken, 1999b; Mintz et al., 2002; Mintz, 2013). Briefly, children seem to build a protolexicon of candidate word-like units (often called "wordforms") by their first birthday (Bannard & Matthews, 2008; Ngon et al., 2013), and as they segment units acoustically similar to each other, they become more attentive to their acoustic details (Swingley & Aslin, 2002).

In this dissertation, we focus on a specific learning task: the segmentation task. However, since previous work on segmentation is limited, we also briefly refer to production and grammar acquisition, supposing that both result from a successful segmentation. In general, children comprehend words and functional items earlier than their age of production (Clark & Hecht, 1983). Following babbling, production of the first content words occurs at 12-20 months, and increases rapidly at 16-18 months (Diesendruck, 2007). This increase is often described as a 'vocabulary spurt' - but see Bloom (2004) challenging this notion. At around two years, children enter the two-word production stage (Sakai, 2005). By two and a half years, their productive vocabulary size is about 600

words, and at six years it often exceeds 10,000 words (Goodman et al., 2008). Morphological elements in production appear within the first year of talking (Clark, 2017).

Before their second birthday, English-learning children grasp the grammatical structure of the language, its categories, and can differentiate between nouns, adjectives, or even transitive and intransitive verbs (Booth & Waxman, 2009; Gelman & Taylor, 1984; Gómez & Lakusta, 2004; Höhle et al., 2004). By the age of 2 years, they master number, earlier for nouns than verbs, and, by the age of 3 years, they master most tenses (even though some, such as present perfect, are not fully mastered until 4-5) and derivational affixes.

Children tend to learn forms of words before their meanings (Jusczyk & Hohne, 1997; Swingley, 2007). Although early semantic learning will not be further discussed in this dissertation, we mention that this is another challenging task they need to tackle, given that there often are numerous hypotheses for a word's meaning. English-learning children start attaching candidate referents to wordforms at about 6 months, mostly for concrete items (Bergelson & Swingley, 2013; Diesendruck, 2007) and they keep refining the concepts based on input (He & Arunachalam, 2017).

1.2 From the inside

Language acquisition can be described as a product of mental processes, which receive as *input* information from the linguistic environment and produce as *output* the mental representation of the language, as well as the observable ability to comprehend and produce language (Hoff, 2006). Thus, the output of language acquisition is guided by the input, its availability and properties, as well as the computational system, its cognitive mechanisms and tools (Mintz et al., 2002). For the sake of this dissertation, we adopt this definition.

We can infer from this definition that children need to interact with the world (Jiang et al., 2020), and get exposed to surrounding language input (Morgan & Demuth, 1996), since language learning results from this exposure. Indeed, across acquisition theories (Chomsky, 1959; Christiansen & Chater, 2008; Elman, 1996; Gervain & Mehler, 2010; Michael Tomasello, 2001), there is a consensus that input is essential to acquisition, as it provides information on the language, its vocabulary and structure. However, before discussing more on language input, we provide below a brief description of the cognitive mechanisms and tools used in acquisition.

1.3 Learning mechanisms and cues

Several strategies have been proposed to account for word and morpheme segmentation and learning in language acquisition (see *Figure 2*). Contrary to adult language learning, children start learning language probably based on bottom-up (signal-derived) strategies (Mattys et al., 2005; Mersad & Nazzi, 2012; Pierrehumbert, 2003) - although they also seem to have prelexical access to some top-down information (e.g. word order, Gervain et al., 2008).

Traditionally, these strategies have been described to contain both language-general and language-specific cues. However, depending on different definitions of 'specificity' (whether they refer to the strategy or the content), this classification of cues is not straightforward (e.g. phonotactics, Blanchard et al., 2010; stress, Endress & Hauser, 2010). Many cues specific to speech are probabilistic, and could be framed as statistical cues. For instance, phonotactics can be framed as the probability of one sound following another within the speech stream. Even for stress, a traditionally language-specific cue, infants need to track the distributional frequencies of stressed syllables (Johnson & Jusczyk, 2001). For illustrational purposes, we present the cues in Figure 2 avoiding this distinction.

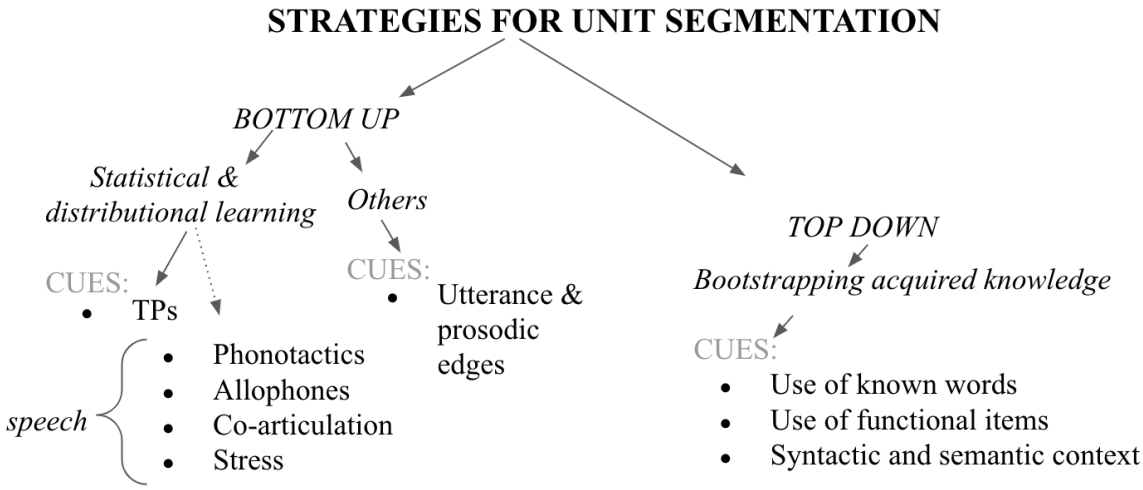

Figure 2: Indicative strategies proposed to account for word and morpheme learning in LA.

It has been suggested that transitional probabilities (TPs) probably provide initial information for the subsequent use of the speech cues (Junge, 2018; Swingley, 2005); see also Johnson & Jusczyk, (2001), Thiessen & Saffran (2003) for experimental and Vallabha et al. (2007) for modeling work. **In this dissertation, we test the segmentability of input based on bottom-up cues, using cues such as TPs, phonotactics and utterance boundaries. We describe these cues below.**

In laboratory experiments, children manage to extract *patterns* out of linguistic systems (Chambers et al., 2003; Gerken, 2006; Saffran & Thiessen, 2003), and through some kind of abstraction, they generalize them to new stimuli (Berko, 1958; Gomez & Gerken, 1999a; Tenenbaum et al., 2011). These patterns emerge from regularities in input (Conway et al., 2010; Lany & Gómez, 2008; Romberg & Saffran, 2010). This inductive computation can be understood in probabilistic terms, in order to capture the uncertainty of the learner; their beliefs are updated as they process the (often insufficient or ambiguous) input (Gopnik & Tenenbaum, 2007).

Statistical learning refers to this ability of extracting statistical patterns from a stream of perceptual experiences (Romberg & Saffran, 2010), and is considered a bottom-up strategy. Laboratory experiments have shown that children can compute basic statistics, such as element frequency, and their frequency of co-occurrence, based on statistical and distributional regularities. This information can prove useful across multiple linguistic levels; *lexical* (Estes et al., 2007; Romberg & Saffran, 2010), *semantic* (Vouloumanos & Werker, 2009; Yurovsky et al., 2014; Diesendruck, 2007; Gleitman, 1989), *phonological* (e.g. Maye et al., 2002) and *morphosyntactic* (e.g. Thompson & Newport, 2007) -also see Fourtassi & Dupoux (2014) for a model learning parallel linguistic representations based on statistical learning.

We are specifically interested in the lexical level. Sound sequences occurring within words are more likely to co-occur than sound sequences occurring incidentally across boundaries, and this can be measured with transitional probabilities (Aslin et al., 1998). Saffran et al. (1996) demonstrated that, when hearing an artificial language, English-learning 8-month old children can perceive transitional probabilities across syllables and use this information to segment words. A similar strategy has been observed in neonates (Bulf et al., 2011; Teinonen et al., 2009) and in tamarin monkeys (Hauser et al., 2001).

Even though the validity of statistical learning mechanisms was strongly supported by Pelucchi et al. (2009), who showed that English-learning 8-month olds could track transitional probabilities in naturally produced, grammatical stimuli in Italian, E. K. Johnson & Tyler (2010) failed to replicate an artificial language study when the words of the language had variable length. A meta-analysis of studies implementing transitional probabilities for segmentation showed a significant but small effect, whose presence depended on the type of speech used, real or synthetically produced (Black & Bergmann, 2017). Since the learning experience is necessarily simplified in the lab (Endress et al., 2009; E. K. Johnson & Seidl, 2009; Pierrehumbert, 2003), the ecological validity of mechanisms

should be checked by testing their potential usefulness on real-life input (Frank, Goldwater, et al., 2010; Swingley, 2005).

In previous laboratory experiments, children also made use of phonotactic regularities (Mattys & Jusczyk, 2001) and allophonic distributions (Gambell & Yang, 2005; Peter W. Jusczyk, Hohne, et al., 1999) to segment words. They may learn about phonotactics by looking at the distribution of utterance-final and utterance-initial phone sequences (Chambers et al., 2002, 2003) - also see Blanchard et al. (2009) and Hayes & Wilson (2008) for modelling phonotactic learning. Conditional probabilities approximate this learning (e.g. the probability that [d] comes after [p] is very small, so there may be a boundary). Allophonic distributions can also be informative on word boundaries. For example, allophone /t/ is aspirated at the beginning of words (e.g. 'table') but unaspirated at the end (e.g. 'hat').

Finally, phrase and utterance edges are acoustically salient to children early on (Gout et al., 2004; Shukla et al., 2011; Tyler & Cutler, 2009). Infants could assume that these edges are also word boundaries, and use them to segment edge-final and edge-initial words (Seidl & Johnson, 2006).

Last, we briefly mention some top-down segmentation strategies here. In this dissertation, we test the segmentability of input based on one top-down strategy, the use of known word-like items (Christophe et al., 1997). Other strategies include the use of semantic and syntactic content (Gillette et al., 1999; Gleitman, 1990). In early acquisition, very few content words are identifiable, such as the child's own name and the appellation for her parents. Functional items, which usually are frequent, short, and at the borders of prosodic units, are recognized early on (Christophe et al., 2008a; Shi et al., 1998) and have a similar bootstrapping role. In experimental settings, children already at 8 months use some of them to segment content words (Hallé et al., 2008; Mintz, 2013; Shi & Lepage, 2008; Shi et al., 2006).

Guide to Chapters

This dissertation is divided into 10 chapters and two parts. We began with an introductory chapter, where we presented the concept and benchmarks of early language acquisition. Chapters 2-5 are the first part of the dissertation, where we deal with diversity and learnability across linguistic systems. **Chapter 2 is the introduction of Part 1**. In Chapter 2 we mention previous studies looking at cross-linguistic diversity and learning. We discuss the need for future research and how it gets adressed in this dissertation. **Chapter 3 is a modeling study under review**, where we test the segmentability of input from two languages differing in richness of their morphological features, Chintang and Japanese. An innovative aspect of this study is the evaluation of segmentation performance in both words and morphemes. **Chapter 4 is a published modeling study,** where we test the cross-linguistic viability of segmentation strategies. We investigate whether they perform above chance and stably across eight typologically diverse languages. **Chapter 5 an ongoing artificial language study**. We ask whether human participants segment words and/or morphemes when exposed to an artificial language where, just like human languages, words are composed by morphemes. **Chapter 6 is the conclusion to Part 1**.

Chapter 7 is the introduction to Part 2. In this chapter, we mention previous studies looking at diversity and learning from different registers and speakers. **Chapter 8 is a corpus study under review** where speech heard by children is compared across two diverse cultures, in Lesotho and in France. The amount and quality of input is investigated across speech registers. We categorize input as directed to the target children, other children or adults. Input is also measured across speakers; the target children's mothers, other adults and other children. **Chapter 9 is a published modeling study**, where we test the segmentability of speech registers from input to French-learning children. We ask whether performance differences between registers can be explained by their specific characteristics. **Chapter 10 is the conclusion to Part 2**. Finally there is is a **general discussion**, where we discuss future lines of research and some personal insights about language learning in general.

Part 1

2. Diversity across languages

Having set down some basic definitions and provided a summary of previous relevant results from English learners, we start the first main part of the dissertation. In this introductory Chapter 2, we look at which cross-linguistic characteristics may diversify the input children grow up hearing, and how these characteristics can affect learnability. In 2.1, we talk about the diverse typological features of languages around the world. In 2.2, we discuss how different languages are learned, based on previous experimental and modeling evidence, and we emphasize the need for future research. In Chapters 3, 4 and 5 we address this need with three different studies. In Chapter 6, we provide conclusions based on our results.

2.1 Input across languages

Several questions can be asked if we want to truly understand language acquisition (Bertolo, 2001); What is learned during acquisition? What strategies are used to learn? When do we say that learning has been successful? However, we cannot begin to ask these questions, if we don't know the linguistic input children receive. This knowledge will help us restrain hypotheses for all other questions on language acquisition. What we know for sure is that input of children around the world is *extremely diverse*, as languages differ remarkably in their typological features (Norcliffe et al., 2015; Bickel, 2014; Evans & Levinson, 2009).

Already in 1985, Tomasello & Mannle suggested that research should "investigate more thoroughly the nature and effects of the total range of language models available to language learners" (p. 916). Several years later, Evans & Levinson (2009) emphasized that the cognitive science community is not yet aware of the diversity across languages. The realization that language acquisition theories should explain how *all* languages are learned, and how children cope with this variation, has recently led to increasing attention towards cross-linguistic work (Stoll & Bickel, 2013).

2.1.1 Variation in typological features of language input

Inflectional *morphology*, as will be discussed later on, is a major source of diversity (Penke, 2012). Traditional linguistics categorize languages as isolating/analytic, synthetic (which can be fusional and agglutinative) and polysynthetic. Analytic languages have very little affixation, almost no bound morphemes and no systematic word derivation process (see Mandarin Chinese, Vietnamese). Synthetic languages have higher morpheme-word ratios and richer morphological systems through agglutination and fusion; Fusional languages have a small set of bound morphemes and often some free morphemes, each morpheme marking several grammatical functions (see Greek, Spanish). Agglutinative languages have a large set of bound morphemes, and each morpheme has one grammatical function (see Hungarian, Swahili). Finally, polysynthetic languages have large sets of both agglutinative and fusional morphemes, and they can have more than one stem in a single word, e.g. by incorporating the subject and object nouns into a verb stem.

This categorization has been challenged by modern linguists, who consider morphological complexity as more of a continuum. In any case, it is evident that languages can differ maximally. For example, in English, verb stems only optionally co-occur with only a few, and not necessarily bound, affixes. In other languages, stems can combine with many affixes, and the combinations also depend on specific inflectional classes (e.g. German plural has seven shapes, distributed over sixteen inflectional classes), on allomorphy, exponence, or even the speech context and its participants (e.g. Chintang, Stoll et al., 2017). According to Stoll et al. (2017), in English, stems appear in the same form frequently and across contexts, whereas in Chintang, children hear 15.3 times more unique verb forms.

Other than morphology, languages exhibit enormous diversity in *phonology* (e.g. in their size of phonemic inventory, Evans & Levinson, 2009, phonemic properties, Pierrehumbert, 2003, word stress Jusczyk, Houston, et al., 1999), in *semantics* (e.g. lacking or having elaborate semantic distinctions, Evans & Levinson, 2009) and in *syntax* (e.g. in their rules of case government, recursion, Everett, 2005, word order and word classes, Hengeveld, 1992; Dixon & Aĭkhenvald, 2004).

All this diversity should find a place in our understanding of language acquisition. Actually, researchers could benefit from linguistic diversity, as it can "provide a natural laboratory of variation [...] 6000 natural experiments in evolving communicative systems" (Evans & Levinson, 2009, p.432). A way to do so is by testing how learning mechanisms work across languages, and how features varying across languages can have effects on learning.

2.2 Language acquisition across languages

Levinson (2012) said that human cognition is tuned to diversity and flexible to environmental input. Tomasello (2003) claimed that children's learning strategies are adapted to extracting information in any speech environment they happen to grow up in. What we know for sure is that, as far as linguistic diversity is concerned, children learn each and every one of the 6000 languages (Grimes, 1992). However, the trajectory and outcome of acquisition might differ (Hoff, 2006), due to different experiences in the linguistic environment (Fenson et al., 1994; Jones & Rowland, 2017).

Variation in children's grammatical development has mainly been the focus of previous cross-linguistic studies (Demuth, 1998; Devescovi et al., 2005; Stoll et al., 2017) - but see Bleses et al. (2008); Bornstein et al. (2004). Even though lexical development has been less studied across languages, grammatical and lexical development seem to be correlated (Frank et al., in prep). Since there is little previous evidence on *segmentation* across languages, we consider lexical and grammatical development as evidence of successful segmentation.

Dan Slobin and colleagues, pioneers in building a list of principles for cross-linguistic language acquisition, developed a field manual for cross-linguistic studies (Slobin, 1967). Since then, several cross-linguistic studies have been published. However, less than 10% of the world's languages have decent descriptions, and language acquisition corpora are still limited: 0.1% of 6000 languages spoken today (Evans & Levinson, 2009; Jaeger & Norcliffe, 2009).

CHILDES, an open repository, contains acquisition corpora for only a handful of languages, the majority of which is IndoEuropean. This is problematic because some IndoEuropean features happen to be rare in other language families (Stoll & Bickel, 2013). Specifically on the acquisition of agglutinative and polysynthetic languages, very little work has been done (Kelly et al., 2014). Existing work has often focused on the acquisition of specific linguistic phenomena involving tense, number and voice, mostly in observational studies not readily comparable with each other.

2.2.1 Analysing previous learning outcomes

While the order in which children acquire meaningful units has been studied in detail for English (e.g. Brown, 1973), we have little evidence for diverse languages (Clark, 2017). In languages where uninflected stems are possible words, such as English, first inflected forms are produced around the two-word production stage (Penke, 2012). In languages where this is not the case, children produce

their first contrasting morphemes at only two-to-three months after starting to speak, and inflected forms are produced already in the one word stage (Hungarian, Macwhinney, 1976; Finnish, Toivainen 1990; Italian, Pizzuto & Caselli 1994).

According to Stoll et al. (2017), morphology of Chintang, a polysynthetic language with an elaborate verb class system, is learned early on. Even children below age two display impressive amounts of morphological variation, producing 22-40 different affix combinations. English-learning and Chintang-learning children become competent speakers at approximately the same time, despite huge differences in structure of the languages.

African Bantu languages have elaborate noun class systems with prefixes formed by optional consonants. Despite this, children correctly produce these morphemes by two-and-a-half years, and know about their shape several months before systematic production (Demuth, 1992, 1998). Connelly (1984) suggests that they are even 6-10 months in advance of their English speaking peers in terms of producing morphologically elaborate utterances. Demuth (1990) observes that they are more advanced than English learners in their use of various grammatical constructions. Both Demuth and Stoll et al. (2017) suggest that increased attention could actually compensate for the complexity of some languages.

Studies in other morphologically rich languages (such as Inuktitut, Tzeltal) report similar patterns (Allen & Crago, 1996; Crago & Allen, 1998; Pfeiler, 2003). Fortescue (1984) mentioned that in a single recording of a child learning West Greenlandic at twenty seven months, the child used 24 derivational and 40 inflectional affixes - a following study on 5 children confirmed this pattern (Fortescue & Olsen, 1992).

Xanthos et al. (2011) reported a positive correlation between the mean size of inflectional paradigms and the speed of morphological development in child speech. He concluded that “although early exposure to a variety of inflectional forms may seem to complicate the learning task for the child, it may help children exposed to a richly inflected input to focus more on different forms and on differences in meaning expressed by inflectional means than children exposed to a less richly inflected input” (p.19).

However, most of these studies are observational, and none looked specifically at segmentation. We mention next some results from baby experiments and modeling studies with respect to cross-linguistic segmentation.

There is experimental evidence that children, already by their first year, can process analytically words and morphemes in languages such as English (see previous discussion in Chapter 1). There is much less evidence for children learning languages with richer morphological characteristics. In a recently published study by Ladányi et al. (2020), 15-month old children learning Hungarian, an agglutinative language, could segment words of their language into stems and affixes, especially when an affix was frequently used in the language. This outcome is in line with other experimental studies documenting segmentation in early development for English-learning children.

As far as modeling is concerned, previous work has addressed the issue of learnability in diverse languages for word segmentation (e.g. Goldwater et al., 2009). A review of the literature on infant word segmentation model performance across languages can be found in the Appendix A. Below is a sample graph showing the performance accuracy of several models for different languages (*Figure 3*). It can be observed that the majority of studies are on English or IndoEuropean data. Most other languages have only been studied once. In general, performance for English is higher than for other languages. A brief description of all models is given below, and more details are provided in the Appendix B.

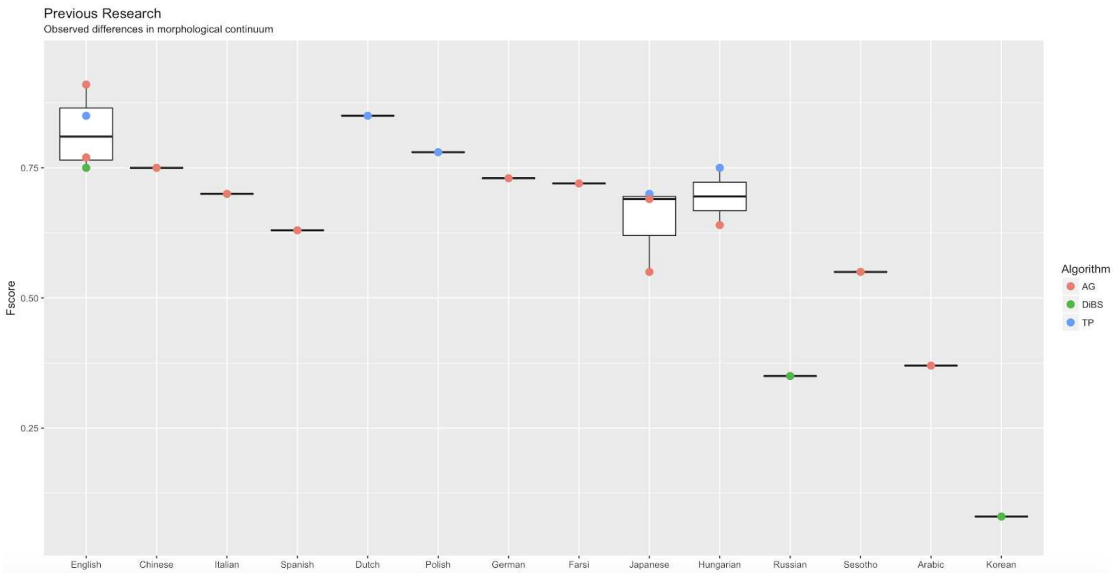

Figure 3. The x axis represents different languages previously modelled for word segmentation: English, Chinese, Italian, Spanish, Dutch, Polish, German, Farsi, Japanese, Hungarian, Russian, Sesotho, Arabic and Korean. The best results belong to English, Chinese, Dutch and Polish. The y axis represents F-scores as a measure of evaluation. When the F-score is 1 then the segmentation is perfect, and when the F-score is 0 the segmentation failed.

2.2.2 Linking input to learning outcomes

It has been suggested that factors such as *frequency* (Ambridge & Lieven, 2011; Endress et al., 2009), *salience* (Penke, 2012; Yung Song et al., 2009), and *transparency* (Callanan & Sabbagh, 2004; Diesendruck, 2007; Dressler et al., 2010; Penke, 2012) across linguistic levels and especially in morphology may affect learning (E. Clark, 2017; Goldfield, 1993; Stephany & Voeikova, 2009; Tatsumi et al., 2018).

For example, more occurrences of a word offer more opportunities to learn the word. Content word frequency in input correlates with frequency in children's vocabulary (Swingley & Humphrey, 2018), even though the effect is not linear and it is stronger for production than for comprehension (Goodman et al., 2008) - see also (Kurumada et al., 2013), showing that Zipfian, and not equal word frequency facilitates adult segmentation. In such morphologically rich languages, content word learning should be challenging, as verb and noun stems obligatorily combine with many, varying units under adjacency constraints. Each word form would only have few occurrences in speech.

Morphological transparency, in terms of affixation without altering a stem's phonological form, is also supposed to affect word learning. Some languages may have a large set of rules to be learned (Stoll et al., 2012) and some paradigms are less regular than others (Clark, 2017). The form of a morpheme may vary because of gender and number agreement (one element can have a single form, e.g. Hebrew, Hungarian, Turkish, or a large number of allomorphs, e.g. Russian, German, SerboCroatian, Icelandic) (Levy, 1983; Wittek & Tomasello, 2002). Additionally, affixes can consist of discontinuous morphological elements e.g. a case suffix and a preposition (Bybee, 1995; Clark, 2017; Penke, 2012).

It may be more straightforward to learn the structure of such a language if we reframe the word learning task as a task of learning meaningful units such as stems and affixes. Specifically for functional items (function words and affixes), when a functional item combines with different stems (high type frequency and productivity) and occurs frequently in the input (high token frequency), it may even trigger the segmentation of inflected words (Plunkett, 1993). Moreover, children seem to acquire inflectional markers earlier when they are salient, for example due to longer duration (Hsieh et al., 1999; Peters & Menn, 1993).

It is still unclear at what extent diversifying factors, such as frequency, salience and transparency, which have been previously suggested to have an impact on segmentation, may affect learning in real life. These factors are also related to the broader term of complexity. Rich languages are generally

considered as more *complex* than others based on their morphological, phonological or syntactic features (Shosted, 2006). One general and often implicit assumption in language acquisition is that such languages are more challenging to learn. For example, 8- and 12-month old infants need longer amounts of processing time to encode more 'complex' than simpler stimuli (Hunter et al., 1983). However, complexity can be defined in many different ways (Gierut, 2007; Miestamo et al., 2008), and there is no agreement on standard metrics (Kusters & Muysken, 2001; McWhorter, 2001).

When defined as an increase in amount of information, complexity might actually be beneficial for learning (e.g., Thiessen et al., 2005). This is especially true, when many sources of information converge (e.g., several cues point to the same units) - also see adult studies by Billman (2007) and Stadler (1992). Little evidence exists on child language learning, but children could identify units in complex languages - see studies by Gerken et al. (2005), Gómez (2002), Thiessen & Saffran (2009) and Teinonen et al. (2009). In this context, less complex would actually make learning more difficult, because it would provide only partial information about linguistic structure. Moreover, *too simple* stimuli may also elicit less interest from infants and thus affect learning (th*e goldilocks effect,* Kidd et al., 2012).

In sum, the relation between complexity and learnability is not linear, and it is still unclear whether a morphologically elaborate language should be considered more complex, thus more difficult to learn.

2.2.3 Comparing learning outcomes across languages

Whether an item counts as easy versus difficult is not easily measured, and even less so across languages, when basic linguistic units such as words are not necessarily comparable.

For example, a child is exposed to a Semitic language where consonants contain the main meaning of the word, while another child learning an IndoEuropean or Turkic language should take both consonants and vowels as identifying words (Clark, 2017). Similarly, the universal notion of 'word' has been challenged - many 'word' cases are not clear-cut, and depend on a language's morphosyntactic system. We can easily describe a word as a single, independent unit in isolating languages such as English, but in other languages, especially morphologically rich ones, learners recognize morphemes as sub-units within words. On the other side, according to Demuth (1988), children learning Sesotho, whose nouns belong to fourteen classes marked by prefixes and agreement, first focus on whole noun phrases, and then on words.

In sum, segmentation may concern different units than words for some languages. It can be straightforward for English words, but in the example case of another language, such as Sesotho, it would not be clear what level to segment first: phrases, words or morphemes.

2.3 Future research

The input children hear entails enormous diversity with respect to its typological features, as was described above. However, previous evidence on learnability is scarce, and not readily comparable. One way to investigate this is by testing scaled-up, diverse input through computational modeling, a useful means of quantifying the effect of diversity on acquisition (e.g. Jones & Rowland, 2017). More studies are needed in order to study cross-linguistic input in terms of segmentability; we should ask how informative this input is for detecting word boundaries across all these diverse languages. **We address the issue in Chapters 3 and 4, by modeling the segmentation of input heard by children in naturalistic settings across languages.**

Second, some features were hypothesized above to have an impact on learning, specifically those at the lower ends of frequency, transparency and salience (Cutler & Carter, 1987). In future studies, we need to take into account specific features that may affect segmentation, and ask whether they account or not for segmentability differences. **We address this issue in Chapter 3, by testing the predictive value of a large set of diversifying features** (and one complementary study can be found in the Appendix C)**. We will also come back to this issue later on, in Chapter 9.**

Third, evidence from acquisition of real, diverse languages, even though scarce, suggests that children tune to the structures of their languages early on (see also Bates & MacWhinney, 1987; Dressler et al., 2010). Since all typically developing children acquire language, and modeling can inform on the efficiency of particular strategies in learning, we should ideally be looking for general, universal learning mechanisms. For example, it has been suggested that statistical learning mechanisms may provide the child with enough data in order to tune to the specifics of their language (Christiansen & Chater, 2008; Morgan & Newport, 1981). However, most studies on segmentation strategies have been based on laboratory experiments, the majority of which involve English-learning children exposed to controlled input in their own or in an artificial language. **We address this issue in Chapter 4, by looking at the performance viability of several learning mechanisms, using longitudinal recordings across a diverse set of languages.**

Last, future studies investigating learnability across languages, should take into account the issue of comparability. Segmentation may concern different units than words for some languages, and some morphological properties may actually trigger segmentation within inflected words. **We address this in Chapters 3 and 5, by reframing the word segmentation task as a task of segmenting meaningful units such as morphemes, in a modeling and an artificial language experiment.**

3. Does morphological complexity affect word segmentation? Evidence from computational modeling*

Abstract: How can infants detect where words or morphemes start and end in the continuous stream of speech? Previous computational studies have investigated this question mainly for English, where morpheme and word boundaries often align. Yet many languages are morphologically more complex, which may present additional difficulties for segmentation.

Our study employed corpora of two languages that differ in the complexity of their morphological structure, Chintang (Sino-Tibetan) and Japanese. While Japanese displays moderate complexity, Chintang exhibits high levels of verbal and nominal synthesis. We employed two baselines and three conceptually diverse word segmentation algorithms, two of which rely purely on sublexical information using distributional cues, and one that builds a lexicon. The algorithms' performance was evaluated on both word- and morpheme-level representations of the corpora.

As predicted, both languages scored lower than previously documented results on English. Segmentation results for Japanese were better than those for the morphologically more complex Chintang. The language effect could not be explained by potential confounds, such as segmentation ambiguity, or by proximal causes, such as word length and lexical diversity. Better performance was observed when evaluating segmentation on morphemes rather than words in Chintang. Algorithms exhibited diverse performance patterns, interacting with both language and level.

Overall, our results indicate that languages varying in morphological complexity (assessed by the number of morphosyntactic features expressed synthetically), could vary in segmentability. Morphological complexity, however, is not the sole determinant; algorithm type and evaluation level can also contribute to predicting segmentation scores

***Loukatou, G., Stoll, S., Blasi, D. & Cristia, A. Does morphological complexity affect word segmentation? Evidence from computational modeling (under review).**

Does morphological complexity affect word segmentation? Evidence from computational modeling

1 Introduction

Typically-developing children acquire language effortlessly and implicitly in the first years of their life. They process linguistic material provided by their caregivers and others around them using robust learning mechanisms that do not require meta-linguistic awareness. Infants begin learning the building blocks of language, i.e., words or morphemes, from very early on, achieving a comprehension vocabulary of hundreds of words by two years of age (Bates et al., 1994). More precisely, during the first year of life, infants might build up a proto-lexicon storing candidate phonological forms (wordforms), which they first have identified based on the available frequency distributions in the input, before actually attaching meaning to these wordforms (Ngon et al., 2013). To break up the speech stream, infants can use prosodic cues (Shukla et al., 2011), co-articulation (Norris et al., 1997), constraints on stranded material (E. K. Johnson & Jusczyk, 2001), and even language-specific information they have learned in the past (words: Bortfeld et al., 2005; Mersad and Nazzi, 2012; syllable sequences: Black and Bergmann, 2017, and phonotactic patterns: Daland and Zuraw, 2013). Here, we report on a series of computational experiments that seek to shed light on the specific processes that young language learners could potentially be using when segmenting the incoming speech signal into word-like forms, or more generally smaller recombinable units.

Young language learners have to learn their language from scratch. In order to mimic this absence of knowledge, models used in previous computational experiments are often unsupervised, meaning that they do not have access to any kind of feedback (i.e., external information on whether they are doing well or poorly). By and large, three classes of algorithms have been used: lexical, sublexical, and baseline. Algorithms in the *lexical* class are often built to find the most economical system of minimal units needed to reproduce the input. They do so usually by creating a lexicon of chunks that are frequently encountered in speech. These salient segments could approximate infants'

first familiar word-like constructions. Algorithms in the *sublexical* class aim to find local cues allowing the learner to posit boundaries, detectable for instance by considering phoneme occurrences at utterance edges or via transitional probabilities. Infant experimental work suggests both classes are cognitively plausible (Mattys et al., 1999; Mersad & Nazzi, 2012; Saffran, Aslin et al., 1996). Finally, previous literature has sometimes used word segmentation *baselines* to evaluate the performance of algorithms (Çöltekin, 2011; Lignos, 2012; Venkataraman, 2001). Baselines represent the simplest strategies possible; for example, treating each basic minimal unit (phoneme or syllable) as words, or treating whole utterances as words.

1.1 Cross-linguistic performance

It has been proposed that language acquisition may not be a homogeneous process, identical in children regardless of the language they are acquiring, but instead that the acquisition process may vary across typologically diverse languages as a function of their grammatical structures (Slobin, 1985). However, the proportion of languages whose acquisition is represented in the literature is low (e.g. Stoll, 2015; Stoll and Lieven, 2014), and the majority of papers on first language development are on English (Slobin, 2014). This sampling bias is problematic because English is not an "average" language, particularly in terms of the properties that may influence segmentation. There is no case marking in nouns and only rudimentary morphological marking in verbal conjugation. Most English words have few or no morphemes other than the root (Aikhenvald, 2007) and as a consequence, word, morpheme, and syllable boundaries usually coincide (DeKeyser, 2005). In fact, the maximum number of morphemes per word in English is 3, which is on the lower end of the typological range (degree of synthesis, Bickel and Nichols, 2007, 2013b).

Languages vary greatly in their overall morphological complexity (Miestamo, 2008; Nichols, 2009; Sampson et al., 2009). A considerable fraction of languages are characterized by rich inflectional morphology and often feature multi-morphemic words. For example, Turkish has a rich concatenative inflectional morphology (Bickel &

Nichols, 2013a, 2013b; Ketrez & Aksu-Koç, 2009). Others are extremely complex such as the polysynthetic languages Tzeltal (spoken in Mexico; Brown, 1998) and Chintang (a Sino-Tibetan language spoken in the Himalayas of Eastern Nepal; Stoll et al., 2017) or Eskimo-Aleut languages such as Inuktitut (Allen, 1996; Bickel & Nichols, 2013b). Such languages use morphemes (prefixes, suffixes, circumfixes, and infixes) to code morphosyntactic features (e.g. gender, person, aspect, tense, or polarity), and/or the relation between words in a sentence (e.g., case or agreement). So far there is no common agreement on how to measure morphological complexity cross-linguistically, but it is undisputed that complexity is a gradient notion.

One of the main questions that arises is whether languages with a larger degree of morphological synthesis are more challenging to segment than languages with a lower degree of morphological synthesis, such as English. A number of computational modeling studies have investigated word segmentation in various languages (Batchelder, 2002; Blanchard et al., 2010; Caines et al., 2019; Daland, 2009; Fleck, 2008; Fourtassi et al., 2013; Kastner & Adriaans, 2017; Pearl & Phillips, 2018; Saksida et al., 2017). These results seem to suggest that languages with richer morphological profiles might be more difficult to segment than those with simpler morphology. In the following, we review this evidence in detail, grouping studies in terms of the type of segmentation strategy.

Starting with studies using a lexical approach, Batchelder (2002) compared the accuracy of a lexical segmentation algorithm (BootLex) on English, Japanese and Spanish corpora, and found that the algorithm performed best on English. Most other lexical work has employed versions of Adaptor Grammars (AG), which build a lexicon based on a hierarchical grammar provided by the user (Goldwater et al., 2009; M. Johnson, 2008). It finds patterns of frequent phone sequences in the input corpus, creates a lexicon based on these patterns at specified levels, and then uses the lexicon to segment the input. Using versions of this system, Boruta et al. (2011) documented better results for English than French, and better results for French than for Japanese, which roughly corresponds to the order of morphological complexity (see Fourtassi

et al., 2013 for convergent results). M. Johnson (2008) found better results for English than Sesotho, which is morphologically much more complex than English. It should be mentioned, however, that the data reported by Phillips and Pearl (e.g., Pearl and Phillips, 2018; Phillips and Pearl, 2014a) differed from other lexicon-based results. These authors varied the hierarchical grammar, inspecting both unigram and bigram models. Unigram models are those where the only levels are those of words and phonemes (i.e., sentences as sequences of words, words as sequences of phonemes), and the only level at which a lexicon is stored is the word level. Bigram models can also be defined, by stating that sentences are sequences of phrases, and phrases sequences of words, with the further possibility that the system will memorize common phrases. For the unigram version of the algorithm, English was at the bottom of the performance ranking. However, when a bigram grammar was used, English performed better than Farsi, Hungarian, and Japanese (Phillips & Pearl, 2014a). However, most work using lexical algorithms finds cross-linguistic differences in word segmentation performance that could be explained on the basis of complexity differences.

In general, work using sublexical algorithms also fits this description; differences in performance can usually be explained by complexity differences. Saksida et al. (2017) used a set of segmentation algorithms, all of them based on transitional probabilities, on a range of cross-linguistic corpora. Higher scores were found for English and Dutch than for Japanese, Polish and Hungarian. Gervain and Erra (2012) received better results for the less complex Italian than the more complex Hungarian in some cases, although this language performance was reversed when backward transitional probabilities were used.

Finally, segmentation baselines have rarely been used to compare performance across languages. Pearl and Phillips (2018) implemented a “random oracle” baseline, which had prior knowledge of the true probability of a word boundary after each unit in the corpus (e.g., 0.76 for English). Boundaries were then randomly inserted based on this probability. Performance differences across languages were observed, with the English corpus scoring higher than German and both scoring higher than Spanish, Italian, Farsi, Hungarian, and Japanese. In sum, our reading of the literature suggests

that lower segmentation performance is found for corpora of more morphologically complex languages than simpler ones across all three families of algorithms (lexical, sublexical, and baseline).

Caines et al. (2019) deserve a special mention, because they used many different algorithms across many languages in CHILDES, a repository of child-centered transcriptions (MacWhinney, 2014). Although they don't specifically discuss typological features, they attempt to relate lexico-phonological features such as word length and lexical diversity to segmentation performance. As with the work just summarized, results were evaluated only in the word level.

1.2 Goal of segmentation

The current standard for modeling studies is to evaluate segmentation algorithms on the word level (Daland, 2009). Several reasons made us wonder whether evaluation on the word level alone is optimal.

To begin with, there are at least three notions of "word": orthographic word, grammatical word, and prosodic word. According to Haspelmath (2011), orthographic spaces are to some extent guided by language structure, even though spelling can be purely conventional in some cases. Grammatical words are units defined by morphosyntactic criteria, such as cohesiveness, fixed internal order, and conventional meaning. Finally, phonological words are units defined by phonological criteria, such as segmental and prosodic features like stress (Dixon & Aikhenvald, 2002). Words are not the only meaningful, recombinable units that may be found in running speech. On the contrary, morphemes can be defined as the minimal meaningful units. Moreover, morphemes and words are not homogeneous classes. For example, functional elements make up a class that cuts across words and morphemes, containing both function words (words expressing grammatical or structural relationship with other words in the sentence) and affixes. Although these definitions seem easy in the abstract, there is no single, valid, and standard definition across languages of any of these levels (Bickel & Nichols, 2007; Bickel & Zúñiga, 2017).

Is there evidence that some or other of these units are psychologically valid for infants? Carefully reading experimental evidence, we found some of it suggests that infants can segment phonological words (E. K. Johnson & Jusczyk, 2001) as well as morphemes (Marquis & Shi, 2015; Mintz, 2013) out of running speech. Furthermore, they can segment functional elements early on (Hallé et al., 2008; Höhle & Weissenborn, 2003; Marquis & Shi, 2015; Mintz, 2013; Shi, Cutler et al., 2006; Shi & Gauthier, 2005; Shi & Lepage, 2008; Shi, Marquis et al., 2006; Shi et al., 1999). Functional elements could be used as cues to further bootstrap word segmentation, because of their distinctive properties (Kim & Sundara, 2015; Shi, Werker et al., 2006; Willits et al., 2014) and contribute to robust learning, especially for languages with rich morphological systems.

On the modeling side, most previous work has evaluated segmentation on orthographic words. Since child-centered corpora are rarely annotated at the level of morphemes, previous computational work has not quantitatively evaluated performance on the morpheme level (cf. M. Johnson, 2008). However, there have been qualitative reports considering morphemes in addition to words (Gervain & Erra, 2012; M. Johnson, 2008). Specifically, it has been argued that some algorithms tend to over-segment words (i.e., words would be split up during segmentation). Previous authors have argued that lower segmentation performance for some morphologically complex languages would arise from oversegmentation when evaluating on words (Gervain & Erra, 2012; M. Johnson, 2008). Gervain and Erra (2012) commented that there may be more oversegmentation in Hungarian, as some of the segmented material formed real morphemes, which is interesting given that this unsupervised algorithm is not informed about lexical and morphological composition. Similarly, Fourtassi et al. (2013) segmented an English and a Japanese corpus using a probabilistic lexicon-building algorithm. Qualitative inspection showed that the algorithms broke off morphological affixes, with more oversegmentation cases for Japanese than English. These observed oversegmentation errors suggest that algorithms might segment out morphemes, or at least functional elements, including affixes, in addition to or instead

of some notion of words.

In a nutshell, infant segmentation may target words as well as morphemes (Kim, 2015; Marquis & Shi, 2015), and therefore, if we want to model this segmentation process, evaluating results at the level of both words and morphemes might be a more informative approach.

2 The present study

This study investigates whether languages varying in morphological complexity differ in segmentability (Section 1.1). This question is addressed by assessing segmentability of two morphologically diverse languages, one of which exhibits an extreme degree of morphological complexity. In addition, we ask whether specific language features such as type-token ratio, word length, and utterance length might account for performance differences.

Moreover, this study examines whether algorithms segment out morphemes instead of words (Section 1.2). We inquire whether performance varies as a function of the level of linguistic representation on which segmentation is evaluated, comparing the algorithms' performance based on either orthographic words or morphemes. We further report percentage of under- and oversegmentation on each level.

Regarding the use of orthographic words, it was preferable over other definitions of wordhood for three reasons. First, it allows comparison with previous computational work. Second, it was already available in the corpora we were using. Third, it is unclear that it is much worse or much better than alternative definitions. Phonological and morphosyntactic criteria for word segmentation are also problematic and cannot decide controversial cases: Phonological words may not be consistent within and across languages (Schiering et al., 2010) and they often fail to coincide with morphosyntactic words (Dixon & Aikhenvald, 2002). For this, and as previous segmentation studies did, we used the existing orthographic word boundaries for our word level.

2.1 Languages

In this paper, corpora from two morphologically diverse languages are studied, Japanese (Japonic) and Chintang (Sino-Tibetan). Both languages were chosen on the basis of their typological characteristics and are part of the ACQDIV database, which contains longitudinal corpora of language acquisition for 10 maximally diverse languages (Moran et al., 2016; Stoll & Bickel, 2013). To create this database, a new approach of sampling languages was introduced, called the Maximum-Diversity approach. More than 10 major typological variables that characterize inflectional marking (grammatical case, exponence, possessor agreement, inflectional compactness, syncretism, verb position, verb agreement, split ergativity of agreement markers, split ergativity of case, flexivity, verbal synthesis, nominal synthesis) were considered (Stoll & Bickel, 2013). Languages were sampled from the two largest typological databases, WALS (Dryer & Haspelmath, 2013) and AUTOTYP (Bickel et al., 2017; Nichols et al., 2013), resulting in 5 clusters of maximally diverse languages.

Chintang and Japanese are in two different clusters. The main feature of interest in the present paper is their difference in the degree of synthesis, which allowed us to study the effect of morphological complexity. The degree of morphological synthesis was measured by looking for the maximally inflected verb and noun form, and determining the number of grammatical and lexical categories (morphosyntactic features) encoded in that word form.

Chintang, a Sino-Tibetan language of the Kiranti branch (approx. 6000 speakers, Eastern Nepal), has higher verb and noun synthesis than Japanese (as just mentioned, measured in number of such categories expressed in the most complex word form; compare Bickel et al., 2007 for Chintang, and Kuno, 1973; Tsujimura, 1996 for Japanese), with up to 10 morphemes per word, versus up to 5 for Japanese [1]. As shown in Stoll et al. (2017), there are 148 unique grammatical elements that can occur together with a verb stem in the corpus (120 grammatical markers and 28 secondary verb stems,

[1] Phonological complexity (phonemic inventory and syllabic structure) is similar across the two languages (Bickel et al., 2007; Shibatani, 1990; Tsujimura, 1996).

called V2, that expand the lexical or grammatical meaning of the main verb). Although some forms of verbs are rarely used, they constitute a part of the adult grammar and are eventually acquired by children. Here are two sample adult utterances from the Chintang corpus (CLDLCh1R01S02.0044 and CLDLCh1R01S02.0057 respectively):

(1) a. ahã hun miʔmuŋ namba-ŋa thok-u-ŋs-e-kha
no DEM a.little father.in.law-ERG.A dig-3P-PRF-IND.PST-NMLZ

'The father-in-law has dug it'

b. ba yaŋ yug-a-yakt-a-kha ni ekchin-a
DEM.PROX ADD sit-PST-IPFV-PST-NMLZ EMPH little.while-NTVZ
ekchin-a-kha
little.while-NTVZ-NMLZ

'He used to sit sometimes'

Japanese has moderate verb synthesis, expressing categories such as tense, voice, mood and polarity. A maximally inflected Japanese verb form would include 4-5 categories (Bickel & Nichols, 2013b; Hinds, 1986; Shibatani, 1990). Thus, overall, Japanese has fewer forms both in its noun and verb paradigms and a smaller number of morphosyntactic features expressed, especially in the verb. Here are two sample adult utterances from the Japanese corpus (MYJCu44.1390521 and MYJCu832.1419814 respectively).

(2) a. usagi-chan doko da
rabbit-FAM where be.PRES

'Where is Missy Rabbit?'

b. ocha mo doozo shi-te
tea too handing_over do-IMP

'Go ahead, make tea again'

2.2 Data

The Chintang recordings took place in a predefined week every month with several separated recordings amounting to approximately 4 hours per month involving 6 target children from 0;6-4;4 (Stoll et al., 2017). During the recordings, which were audiovisual, the children were mainly playing outside of their houses. Relatives, other children, and neighbors are part of their daily lives and this was captured in the recordings.

The Japanese data consist of 2 corpora, MiiPro (Miyata & Nisisawa, 2009, 2010; Nisisawa & Miyata, 2009, 2010) and Miyata (Miyata, 2004a, 2004b, 2004c). Recordings took place indoors, mostly at home and there was often just one caregiver conversing with the child. They contain data for 7 Japanese children aged 1;4-5;1 years old. For the MiiPro corpus, the recordings took place every week from 1;2 to 3;0 and later every 1 or 2 months, and lasted 70 minutes per session. For the Miyata corpus, recordings took place every week and lasted 40-60 minutes.

After data collection, both corpora were first transcribed orthographically, and later annotated morphologically. More information on the annotation process for Chintang can be found in Stoll and Schikowski (in press) and Gaenszle et al. (2005); see Miyata and Naka (2006) for information on annotation of the Japanese corpora. In the Chintang corpus, the transcription was done by native speakers and was susceptible to their impression of what an utterance was. This mainly corresponds to clauses marked by intonation. As for the Japanese corpus, there is no information on how utterance boundaries were defined. Documentation suggests the data were transcribed based on the Wakachi format (Miyata & Naka, 1998). As for the morphological annotation, for Chintang, most of the corpus was hand segmented and manually annotated for morphology and parts-of-speech by trained linguistic students. The training took several weeks and was supervised by an expert in this language. A small part of the morphological annotation was generated automatically based on a morphological tagger (Ruzsics & Samardzic, 2017; Samardzic et al., 2015). Japanese morphological tagging was done with the morphological tagger in CHILDES (JMOR, Miyata and Naka, 2014). For information on the data, see also the ACQDIV manual (Schikowski et al., 2018).

2.3 Modeling segmentation

Word segmentation algorithms usually take as input phonological, symbolic text-like representations such as phonemes or syllables, with few exceptions (e.g., Ludusan et al., 2015 and Roy and Pentland, 2002, who applied segmentation algorithms on raw speech data). There is evidence that even newborns have access to syllables (or vowels) as perceptual units (Jusczyk et al., 1995), and that representation of phoneme sequences is available as early as by four months (Seidl et al., 2009). It can be challenging to represent certain cues in text-like representations, such as coarticulation, which will therefore not be studied here. Here, we limit our study to phonemized representations (Moran & Cysouw, 2018), even though it would be possible to study word-level prosody (Börschinger & Johnson, 2014; Gambell & Yang, 2005a), because some languages do not have lexical stress, such as French (Dupoux et al., 1997).

So far only a couple of papers have applied more than one algorithm to the same corpus; when they do, they find widely varying performances across the different algorithms. By and large, algorithms based on local cues and employing sublexical information reportedly yielded lower scores than lexically driven ones both in English (Cristia et al., 2018) and Japanese (Ludusan et al., 2017). In this study, we sought to directly assess variability in performance across languages and included algorithms of both types.

Additionally, there can be enormous differences in performance within the same algorithm depending on the parameters used (e.g., Gervain and Erra, 2012; Saksida et al., 2017). For example, Saksida et al. (2017) documented that different measures such as forward transitional probabilities, backward transitional probabilities and mutual information and especially the threshold parameter (absolute, relative) affect the results of word segmentation. Given these previous results, we will consider a diverse set of algorithms and their parametrization below, fully expecting them to vary in performance.

The computational algorithms used here have been repeatedly used in the past, and were chosen to represent diverse and cognitively plausible segmentation methods,

spanning the three main classes mentioned above: lexical, sublexical, and baseline. More details on the algorithms can be found in Section 4.2, and we therefore provide only a brief conceptual presentation here.

The lexical representative is a version of the Adaptor Grammar introduced previously (Goldwater et al., 2009; M. Johnson, 2008). In a nutshell, the hierarchical grammar we provided was the unigram one, in which sentences are sequences of words and words are sequences of phones, with the lexicon being composed of frequent phone sequences. We had two sublexical algorithms, the first one being the Diphone Based Segmentation algorithm (DiBS). It is based on the intuition that phone bigrams spanning utterance boundaries probably span word breaks (Daland, 2009; Daland & Zuraw, 2013). The second sublexical algorithm is actually highly parametrizable: the Transitional Probabilities algorithm family (TP) assumes that word-internal pairs of syllables tend to co-occur more frequently than word-external pairs (Saffran, Newport et al., 1996), with four different versions resulting from the crossing of 2 parameters with two levels each (Gervain & Erra, 2012; Saksida et al., 2017). In addition, two baselines were included in this study. The first baseline treats each utterance as a word, based on findings that children recognize words in isolation before they do so in sentences (Depaolis et al., 2014). The second baseline treats each syllable as a word, given that infants might track syllable units from early on (Bertoncini & Mehler, 1981; Jusczyk et al., 1995).

3 Key Questions and Predictions

The key questions motivating this study are: Do languages which vary in morphological complexity differ in segmentability? Do algorithms segment out morphemes in complex languages? And which factors could explain performance differences in segmentation?

Morphological complexity should affect segmentation through several pathways, the first being via the distributional properties of the lexicon. Specifically, languages varying in morphological complexity differ in the frequency of lexical units (words and

morphemes). Corpora of morphologically rich languages such as Chintang contain fewer repetitions of each word type, as well as a higher proportion of hapaxes – forms that occur only once – than languages with little morphology (Stoll et al., 2017). For example, there was a higher proportion of hapaxes in a Japanese than an English corpus, and a lower likelihood of correct identification by a lexicon-building algorithm for hapaxes than words with more repetitions (Boruta et al., 2011). Some algorithms might thus detect frequently occurring word parts such as morphemes, instead of words e.g. they could detect a root separately from its suffixes. Lexical approaches, in particular, tend to recycle existing units, favoring repetition. AG finds the most likely segmentation using a lexicon, whose types have been assigned probabilities based on their frequency distributions. Thus, it could break words up into their component morphemes. This behavior would be rewarded when evaluated on morpheme boundaries, but penalized when evaluated on word boundaries.

Aside from these factors, which are purely morpholexical, there could be phonological factors confounded with morphological complexity, such as word length and segmentation ambiguity.[2] Languages varying in morphological complexity should also differ in the length of lexical units. First, morphologically complex languages such as Chintang usually have longer words and longer sentences. Longer sentences mean a more challenging segmentation task, because there are more places in which to erroneously insert a boundary, or miss inserting one. Second, long strings, which can often be decomposed in a number of different morphemes, may have more alternative parses than short ones. Consider the Japanese utterance “iruka”, which can have several different parses (Scherling, 2016), including one in which it is a single word (“iruka”, meaning “dolphin”); or a phrase (“is it?”, where “iru” is the verb *to be* and “ka” is a particle indicating a question). Fourtassi et al. (2013) used the concept of

[2] A host of other factors affecting segmentation and varying across languages have been proposed and studied in the past, such as head direction (Gervain & Erra, 2012) and input representation (Kastner & Adriaans, 2017). However, these factors are orthogonal to the present study (i.e., they are not necessarily confounded with morphological complexity). Therefore, they will not be discussed any further.

entropy (Shannon, 1948) to estimate the different possible segmentation parses (segmentation ambiguity), and showed that segmentation score differences they found for Japanese versus English could be accounted for by that factor.

Thus, lower performance is predicted for Chintang compared to Japanese, the morphologically less complex language. Algorithms might segment out morphemes, as they are shorter and more frequent than words. This could lead to oversegmentation (i.e. splitting a word up) for Chintang, especially within AG. Last, a number of factors are predicted to affect segmentation performance; unit frequency, unit length, utterance length and inherent segmentation ambiguity.

4 Methods

In this section, we detail several stages of analysis: corpus preparation, phonologization, description, segmentation, and evaluation, followed by statistical analyses of the results. Corpus preparation and phonologization were carried out using custom scripts written mainly in bash. Corpus statistics, unsupervised word segmentation, and evaluation employed the WordSeg package (Bernard et al., 2018)[3]. The WordSeg package provides a collection of tools for text based word segmentation. Finally, statistical analyses were performed in R (R Core Team, 2013). All scripts are available at <anonymized for review>. More details can be found at https://osf.io/e8d2r/?view_only=f9d7b6a307734268bd8a515c55255b69. This OSF page contains scripts, results including segmentation performance and statistics, and other supplementary material.

4.1 Corpus preparation

Neither of the two languages exists in open source phonologization or text-to-speech programs, so we applied grapheme-to-phoneme rules to derive the phonological representation[4] (Moran & Cysouw, 2018). We also cleaned the text from

[3] Available from https://github.com/bootphon/wordseg/.

[4] Japanese had been transcribed in Latin script.

any punctuation and annotations. All utterances containing "???" (which indicates incomprehensible speech or impossible morpheme annotation) were removed from both word- and morpheme-level analyses. We also removed utterances where one of the morphemes had been transcribed into an abstract, unpronounceable code (such as FS_N or kV), from both analyses.

Following Phillips and Pearl (2014b), we syllabified the corpora using the Maximal Onset Principle. According to this principle, the beginning of a syllable should be as large as legally possible (Bartlett et al., 2009). We syllabified as follows, for each language separately. First, we made a list of vowels present in the corpus. Second, we made a list of all valid word-initial onsets, defined as all consonants up to the first vowel of the word or morpheme. Third, each utterance was processed from right to left until a vowel was found, at which point consonants to its left would be clustered to the maximally large onset appearing in the list just mentioned phillips-syllabifier. Notice that this procedure does not syllabify over morpheme or word boundaries. Both corpora are larger than those frequently used for modeling studies (Phillips & Pearl, 2014b; Saksida et al., 2017). This allowed us to further divide each corpus into ten equal subsets, based on their length measured in number of utterances, in order to better estimate the variation in the properties of the segmentation algorithms. After pre-processing, the entire Japanese corpus was 84518 utterances long and had 155805 word tokens. The Chintang corpus was 152571 utterances long and had 426288 word tokens. Table 1 gives properties of the subsets after pre-processing. Right before segmentation, within-utterance word boundaries were removed from the corpora, and only utterance boundaries remained.

Mean subset stats	Chintang	Japanese
# utt	15257	8451
# wtokens	42628 (1371)	15579 (1569)
# wtypes	8194 (690)	1634 (361)
# whapaxes	5180 (514)	812 (206)
# wtokens/utt	2.79(0.09)	1.84 (0.19)
# wtypes/utt	0.54 (0.05)	0.19 (0.04)
# m/utt	4.72 (0.29)	1.96 (0.20)
# syll/utt	5.62 (0.25)	3.38 (0.41)
# phon/utt	11.87 (0.60)	6.58 (0.83)
# mtokens/wtokens	1.69 (0.07)	1.07 (0.03)
# mtypes/wtypes	0.25 (0.02)	0.84 (0.03)
# syll/w	2.01 (0.05)	1.83 (0.07)
# phon/w	4.25 (0.12))	3.56 (0.15)
# syll/m	1.24 (0.03)	1.70 (0.04)
# phon/m	2.56 (0.05)	3.33 (0.09)

Table 1

Corpus features: Means (and standard deviation) across the ten Chintang and Japanese subsets (see main text for explanation). # stands for amount, "utt" stands for utterance. "wtokens", "wtypes", "whapaxes" stand for word tokens, word types and word hapaxes. "m", "syll" and "phon" stand for morphemes, syllables and phonemes.

4.2 Algorithms

A brief, cognitively-focused introduction to the five algorithms follows. For technical details, please refer to the WordSeg documentation (wordseg.readthedocs.io; Bernard et al., 2018) and the work cited for each algorithm.

The first algorithm, a member of the **Adaptor Grammar** family, adopts a lexical approach (Goldwater et al., 2009; M. Johnson & Demuth, 2010; M. Johnson et al., 2007). The Adaptor Grammar (AG) is a generalized version of probabilistic context-free grammars (PCGF, M. Johnson et al., 2007). We use a very simple hierarchical grammar with only a few rules: Sentences are composed of one or more reusable words (or morphemes), and words/morphemes are composed of one or more phonemes. Each utterance is parsed as a sequence of words/morphemes, each word/ morpheme is composed by phonemes, and a given word/morpheme of this sequence would be generated either by choosing an existing form from a lexicon based on previous occurrences, or by considering it as a novel item and inserting its phonemic form in the lexicon. The PCFG regenerates the corpus by repeatedly applying this grammar, which is a set of rewrite rules with assigned probabilities. The rules fit the corpus based on how elements have already been written in the past, according to the Pitman-Yor stochastic process, which favors the reuse of frequently occurring rules (M. Johnson et al., 2007). This process is conceptually related to Zipf's Law, a feature of natural languages, which states that in a large corpus, the frequency of any word is inversely proportional to its rank in the frequency table (Zipf, 1935). AG would thus tend to create a lexicon of moderate size comprised mostly of short words (Perfors & Navarro, 2012).

DiBS (Daland, 2009) performs segmentation using phone bigram probabilities. As a consequence, this algorithm requires that the input be coded as a sequence of phonemes. The intuition behind this algorithm is that certain sound sequences almost never occur within words (or morphemes), so if observed they probably indicate a word boundary. For instance, when [pd] occurs in English, the probability that there is a word boundary is very high: $Pr(\#|pd) \approx 1$.

The unsupervised version (phrasal DiBS) treats utterance edges as a proxy for word edges, assuming that phone sequences frequently spanning utterance boundaries likely also span word boundaries. The algorithm estimates the necessary parameters from data using the Formula 1, where $f(x\#_p y)$ is the number of [xy] sequences with an uttterance boundary in the middle and $f(xy)$ is the number of [xy] sequences in any position, and where x is one phone, y is another phone, $(\#_p)$ is an utterance boundary and $(\#_w)$ is a word boundary:

$$\mathrm{p}(\#_w|\mathrm{xy}) \approx \frac{\mathrm{f(x}\#_p\ \mathrm{y)}}{\mathrm{f(xy)}} \tag{1}$$

When $p(\#_w|xy)$ is higher than a threshold parameter, the system breaks the sequence by positing a word boundary. This threshold is estimated with Formula 2:

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

where the total number of words is Nw, the number of phones is Np and the number of utterances is Nu.

The **TP family** assumes that "word-internal pairs of syllables tend to co-occur more frequently than word-external pairs which are relatively unconstrained" (Saffran, Newport et al., 1996, pp.610), thus the transitional probability between adjacent syllables is higher word-internally than at word boundaries (cf. Gervain and Erra, 2012 for evidence that this may not be the case). The basic minimal units are syllables and not phonemes, unlike in the other two algorithms. Forward transitional probabilities (FTP) are defined as:

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

where $f(AB)$ is the frequency of a syllabic sequence AB and $f(A)$ is the frequency of the syllable A. Backward TP (BTP) is similar, except that the denominator is the frequency of the second syllable instead.

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

Also, algorithms in the TP family require another parameter, namely the threshold used to decide whether to add a word (or morpheme) boundary or not. One

possibility is to use TP with a Relative threshold, i.e. BTPr and FTPr, which leads to placing a word/morpheme boundary wherever the TP value of a syllable pair is lower than the TP of the neighboring syllable pairs, as follows. Given a syllable sequence ($WABY$) where W, A, B, Y stand for syllables, a break will be posited between A and B if $TP(WA) > TP(AB)$ and $TP(BY) > TP(AB)$. Another possibility is to use TP with an Absolute threshold (BTPa and FTPa), which would posit boundaries using a threshold, that is the sum TP value of all syllable pairs over the number of different syllable pairs. For example, given a corpus consisting of a syllable sequence ($WABY$) where W, A, B, Y stand for syllables, the absolute threshold is $TPa = \frac{TP(WA)+TP(AB)+TP(BY)}{3}$. A break will be then posited between A and B if $TP(AB) < TPa$.

Finally, we applied two segmentation baselines. The baselines capture simple segmentation strategies. The baseline called **Syll=Unit** uses $p = 1$ to cut at all syllable boundaries, thus treating every syllable as a word (or morpheme). The baseline **Utt=Unit** labels only utterance boundaries as word (or morpheme) boundaries ($p = 0$).

4.3 Evaluation

The output of each algorithm is evaluated using word (or morpheme) token F-scores, derived from precision and recall, as standard for segmentation algorithm evaluation (Phillips & Pearl, 2015). Precision (Formula 5) checks how many words/morphemes in the group of those *segmented by the algorithm* are correct. Recall (Formula 6) checks how many words/morphemes in the group of those existing in *the original gold corpus* were correctly segmented by the algorithm. True positives are the words/morphemes segmented by the algorithm which are indeed found in the input corpus. False positives are the words/morphemes segmented by the algorithm which are actually not in the input corpus. False negatives are words/morphemes in the input corpus that were not in fact segmented by the algorithm.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{5}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{6}$$

The token F-score balances how accurate and complete the set of identified word/morpheme tokens is (Phillips & Pearl, 2015). It is the harmonic mean of precision and recall, as shown in Formula 7.

$$\text{F-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{7}$$

5 Results

Fig. 1 illustrates the token precision and recall for each subset of the corpora. The point of this figure is to demonstrate that precision and recall are correlated to a considerable extent. The correlation between precision and recall emerges because there is no trade off between false negative and false positive results; when segmenting text, a given parse results in neither or both kinds of errors. This is because if a boundary is posited, then if this boundary is correct, it will increase both precision and recall. If the boundary is incorrect, it will reduce both precision and recall.

Since precision and recall are highly correlated, we focus on the more commonly reported token F-scores. Fig. 2 shows for each language, word token F-scores within each of the 10 subsets, as well as results for the entire corpus, which are nearly always contained in the range of variation of the subsets. The F-scores are presented numerically in the online supplementary material. Similarly to what we found, Bernard et al. (2018) documented that variation in corpus size beyond the first 5k utterances seems to play a negligible role in performance of these segmentation systems, as replicated here. Fig. 2 suggests that there were strong interactions between the three factors of interest (language, algorithm, and evaluation level), which are tested statistically in the next section.

Before proceeding with this statistical evaluation, we perform some descriptive observations. Average performance across algorithms on the word level was .48 for Japanese and .33 for Chintang; and on the morpheme level, this was .49 and .41, respectively. Thus, performance for Japanese was similar across levels, while performance for Chintang was worse for words than for morphemes, and, on morpheme level, it was close to Japanese for some of the algorithms.

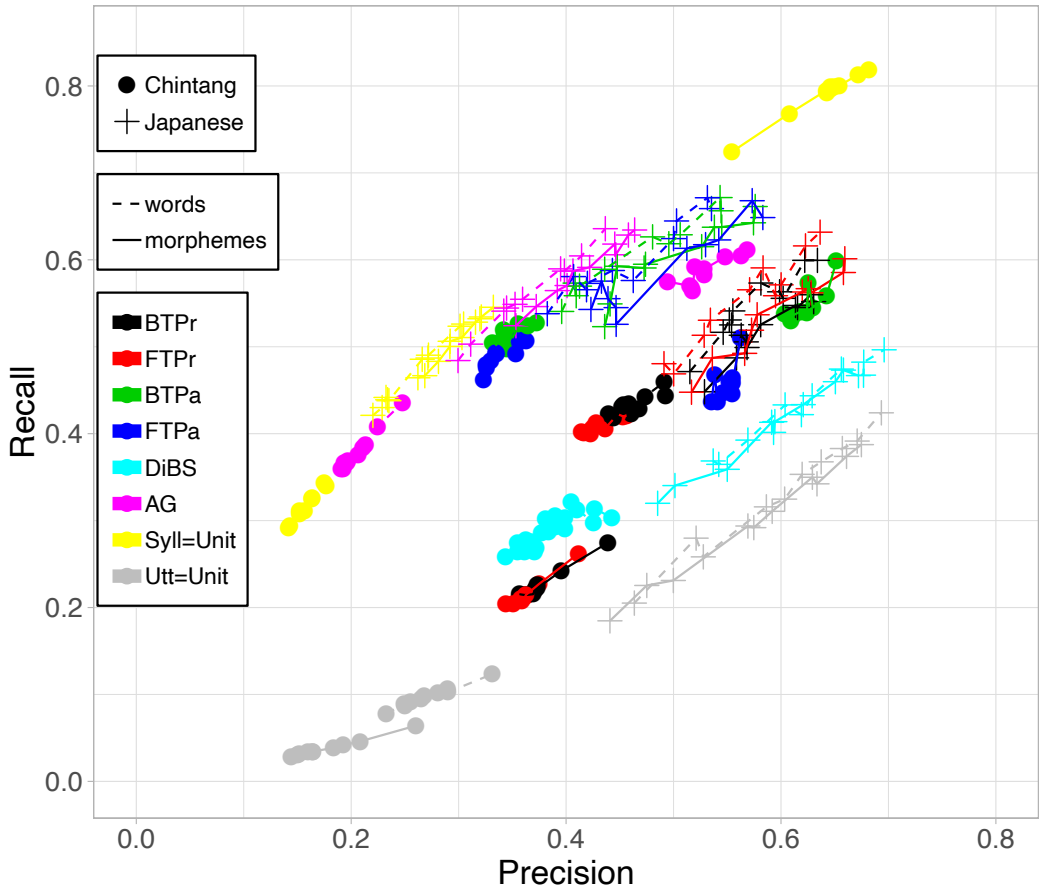

Figure 1. Precision and recall across languages and algorithms for each evaluation level. Algorithms are marked by color. Languages are marked by shape. BTPr, FTPr, BTPa or FTPa indicate the segmentation result for one of the different versions of TP. AG are the results of the unigram Adaptor Grammar. Syll=Unit and Utt=Unit are the results of the baselines. Each dot indicates the results for one of the ten subsets of a given corpus.

To put these descriptive results in the context of the broader literature, we discuss some observations clustered on the basis of the different algorithms, since previous work exclusively employed one algorithm. At points, we need to focus on the word level for this comparison, since previous work has systematically evaluated performance quantitatively on this level, and this level alone.

Focusing first on AG, the sizable performance difference between the two languages found on the word level was reversed on morphemes: AG-word had an average score of .44 for Japanese and .27 for Chintang, whereas for AG-morpheme, the performance for Chintang was higher than that for Japanese, with an average score of

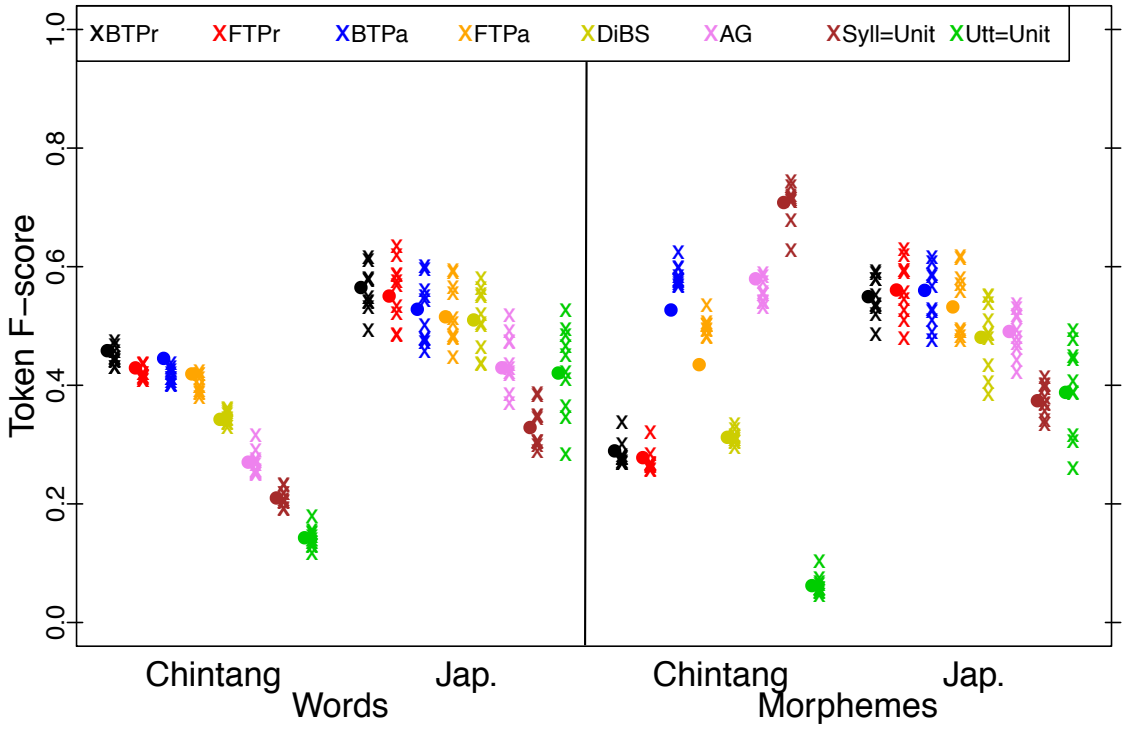

Figure 2. Token F-scores across language (Chintang, Japanese) and level (words, morphemes). Algorithms are marked by color. BTPr, FTPr, BTPa and FTPa indicate the segmentation result for the different versions of TP. AG are the results of the unigram Adaptor Grammar. Syll=Unit and Utt=Unit are the results of the baselines. Filled circles indicate the results for the corpus segmented as a whole. Each "x" shows the result for one subset. Jap. stands for Japanese.

.56 versus .49. As for comparisons with previous work, we found that our AG-word scores were much lower than the .77 documented for English (Fourtassi et al., 2013). Zooming out for a moment from our key question, we also notice that AG achieved higher scores than DiBS and TP *only* in the Chintang morpheme level.

Turning now to TP, absolute-threshold TPa had the highest performance. The higher scores for TP are not just due to our addition of TPa, since even TPr outperforms AG. This is a matter that should be investigated further. TP had relatively smaller language differences compared to AG. TPa-morpheme scores were higher than TPa-word scores, whereas the opposite was true for the relative-threshold

TPr. Worse performance for morphemes than words for TPr is reasonable in hindsight, given that in this implementation a boundary can only be posited in relatively long strings of syllables (see also Gambell and Yang, 2005b). We did not observe much difference between forward and backward alternatives (cf. Gervain and Erra, 2012 for arguments that this parameter should matter for languages varying in head direction, and Saksida et al., 2017 for other data showing that it may not). As for other comparisons with previous work, the best performance for TP in this paper was .63, well below the .85 recorded for English by Saksida et al. (2017).

DiBS showed a language effect that was stable across words and morphemes. For the word level, our scores were .35 and .51 for Chintang and Japanese respectively, close to those for English CHILDES corpora (.43, Daland and Pierrehumbert, 2011).

Finally, the baseline scores ranged from .06 (Utt=Unit for Chintang morphemes) to .71 (Syll=Unit for Chintang morphemes). Both did a better job segmenting Japanese than Chintang on the word level. However, results were different on the morpheme level. Chintang morphemes are on average shorter than Japanese ones (see Table 1; 1.24 versus 1.70 syllables per morpheme), whereas the opposite is true for words (2.01 versus 1.83 syllables per word). As a result, on morphemes, performance is very high when boundaries are systematically posited after every syllable (Syll=Unit), and very low when no boundary is posited at all (Utt=Unit). This might explain why, on morphemes, Chintang outperformed Japanese with Syll=Unit, but had a lower score than Japanese with Utt=Unit.

5.1 Regression on token F-Scores

A regression predicting F-scores from language, level, algorithm, and their interactions accounted for most variance in the data, $R^2 = .93$ ($F(31, 288) = 134.95$, $p < .001$).[5] Even though the presence of significant interactions precluded a direct

[5] The function was: lm(token fscores $\sim$ language $*$ level $*$ algorithm + (1/file), subsets). Token F-scores are the F-scores to be predicted by language, level, and algorithm as fixed effects, and subset as random factor. The data frame contains 320 observations (2 languages x 2 levels x 10 subsets x 8 algorithms).

interpretation of the main effects, the regression confirmed an advantage for Japanese, with a positive coefficient estimating the language effect. Interestingly, the coefficient for this language effect was smaller than that for evaluation level. Further information on the regression results can be found in Tables 2 and 3 and more detailed outcomes in the online supplementary material.

5.2 Proximal causes

In Section 3, we inquired whether specific features related to morphological complexity would affect segmentation. Mean values of such features for the Chintang (word and morpheme) and Japanese (word and morpheme) subsets are shown in Table 4. The regressions introduced next are ran on the subset versions. One prediction pertained hapaxes and repetitions. Specifically, we mentioned that a higher proportion of hapaxes and fewer repetitions of each word token (since each lexeme can have different surface forms) might lower performance in lexical algorithms.

Therefore, we first searched whether type/token ratios could account for (some of) our results. We measured the Moving-Average Type-Token Ratio (MATTR) for each of the 20 subsets (10 for Chintang and 10 for Japanese), in the word and morpheme level gold versions. WordSeg's MATTR computes the type-token ratio in a window of 10 units, shifting this window one unit at a time, and returning this moving average. Thus, it controls for corpus length differences. In a regression across subsets predicting F-scores from MATTR, language, and algorithm (for each subset)[6], MATTR had a non-significant coefficient of -0.396 (SE= 0.452, p-value=0.383) for the regression on words, and -1.176 (SE= 0.888, p-value=0.188) for morphemes.

---

[6] The function was: lm(token fscores ~MATTR* language * algorithm + (1/file) , subset=c(level=="words or morphemes"), subsets). Token F-scores are the F-scores to be predicted by feature (in this example, MATTR), language, and algorithm as fixed effects and subset as random factor.

factor	SumSq	Df	F value	Pr(>F)
lang	0.06	1	43.61	0
level	0.14	1	91.83	0
algo	0.88	7	84.88	0
lang:level	0.06	1	39.66	0
lang:algo	0.11	7	11.08	0
level:algo	1.88	7	182.23	0
lang:level:algo	0.75	7	72.97	0

Table 2

Analysis of variance (ANOVA type III) for all factors and interactions on a linear regression with token F-scores as the dependent variable, rounding to two decimals. "lang" stands for language,"algo" stands for algorithm. "Pr(>F)" stands for the significance probability value associated with the F value, "SumSq" for sum of square values, "Df" for degrees of freedom.

We also measured the hapax ratio by dividing the amount of hapaxes by the total number of unit types for each of the 20 subsets, in the word and morpheme level gold versions. In a regression similar to the one above (with language and algorithm), but now incorporating hapax ratio as an additional predictor, the coefficient for the hapax ratio was also non-significant, -0.682 (SE=1.044, p-value=0.514) for the word-level regression, and -0.275 (SE=1.18, p-value=0.816) for the morpheme-level regression.

In addition, we had predicted that word length side effects of complexity could also matter. Indeed, morphological complexity is correlated with word length, but not morpheme length, according to Table 1. We thus measured the average unit (word or morpheme) length by dividing the total number of phone tokens by the number of unit tokens, separately in the word and morpheme level gold versions of the subsets. A new pair of regressions was thus fit across subsets. Token F-scores were predicted from unit length in addition to language and algorithm, on the word level, and separately on the morpheme level. The unit length factor had a non-significant coefficient estimate of -0.079 (SE =0.084, p=0.35) for the word-level regression; and 0.331 (SE=0.189, p-value=0.082) for the morpheme-level regression.

Last, sentence length as operationalized by the number of syllables per utterance was measured by dividing the total number of syllable tokens by the number of utterances for each subset. A similar pair of regressions revealed this predictor was not significant for the word level (estimate=-0.047, SE=0.026, p-value=0.067) but was significant for the morpheme level (estimate=-0.055, SE=0.024, p-value=0.025). Details on the percentage of variance explained for all regressions can be found in Table 5.

Factor	AG	BTPa	FTPa	BTPr	FTPr	DiBS	Syll=Unit	Utt=Unit
lang	***	***	***	***	***	***	***	***
level	***	***	***	***	***		***	**
lang:level	***	***	**	***	***		***	

Table 3

*Analysis of variance (ANOVA type III) significance results for linear regressions where F-scores are predicted by language, evaluation level, and their interaction within each algorithm separately. "lang" stands for language. p<0.001=***, p<0.01=**, p<0.05=*.*

5.3 Is the language effect due to entropy?

We further investigated whether language effects may be due to one potential confound, in particular the possibility that one of the languages is intrinsically more ambiguous to segment (sentences having different possible segmentation parses). To this end, we followed Fourtassi et al. (2013) and estimated the segmentation entropy of the corpora (Normalized Segmentation Entropy) using WordSeg's descriptive toolset. In Fourtassi's study, English was less ambiguous than Japanese, with entropies of .0021 and .0156 respectively on the word level. The segmentation entropy of our Japanese ACQDIV corpus was 0.028 (word level), thus close to their Japanese results. Surprisingly, the segmentation entropy for Chintang was 0.007 - even less ambiguous than English. We also found overall higher entropy levels when inspecting the morpheme level, with smaller language differences. Even more surprising, in a pair of regressions where entropy, language and algorithm (for each subset) were included, entropy failed to explain a significant proportion of the variance, with a coefficient

estimate of 11.012 (SE=9.033, p-value=0.225) on the word level, but explained a part of the variance on the morpheme level: -5.618 (SE=2.237, p-value=0.013).

5.4 Over-, under- and missegmentation

A breakdown of segmentation performance as a function of part of speech and algorithm is provided in online supplementary materials. Over-, under- and missegmentation cases are reported in Table 6. In the current study, we operationalize oversegmentation as the splitting up of a unit in one or more sub-parts (regardless of whether these are reasonable smaller units or not). We consider undersegmentation the clustering together of two or more words. All other differences from the gold segmentation were labeled missegmentation. The following example illustrates how they were measured: If the input sentence "the dog ate the other dog" is returned as "thedog at ethe other d og", then the score will be 1/6 correct segmentation, 1/6 over-segmentation ("d og"), 2/6 under-segmentation (one for each input word, "the" and "dog"), and 2/6 mis-segmentation ("at ethe").

Factor	Chin. w	Jap. w.	Chin. m.	Jap. m.
MATTR	0.866	0.772	0.837	0.774
prop. hapax	0.615	0.512	0.302	0.455
phones/w	4.248	3.569	2.559	3.333
syll/utt	5.615	3.382	5.831	3.341
entropy (NSA)	0.007	0.028	0.036	0.030

Table 4

MATTR, proportion of hapaxes (prop. hapax), phonemes per word (phones/w), syllables per utterance (syll/utt) and segmentation entropy (NSA) for the entire Chintang word, Chintang morpheme, Japanese word and Japanese morpheme corpus, rounding to 3 decimals.

As for language effects, it was hypothesized in Section 3 that there might be more cases of word level oversegmentation in a morphologically complex language, because algorithms would break apart morphological affixes. As predicted, oversegmentation

rates were higher for both languages when evaluating on words, and, more precisely, they were higher for Chintang than Japanese. This is because word level evaluation considers word oversegmentation an error, but morpheme level evaluation does not penalize it. However, while oversegmentation was substantially reduced on morphemes, it did not disappear.

Finally, previous studies suggested that lexical algorithms might break apart morphological affixes (Section 1.2). We reasoned that this would lead lexical algorithms to oversegment more than sublexical ones (Section 3). This was true for our data, where AG showed distinctive oversegmentation patterns compared to DiBS and TP, and almost no undersegmentation. As far as sublexical algorithms are concerned, DiBS exhibited undersegmentation for both levels and languages. TPa-word tended to oversegment, but TPa-morpheme would undersegment. TPr-morpheme would also undersegment morpheme sequences, particularly with the Chintang corpus, where 70% of its tokens were undersegmented. Unsurprisingly, Syll=Unit tended to oversegment, whereas Utt=Unit was mostly undersegmenting.

Factor	R^2 Factor w.	R^2 Factor*lang*algo w.	R^2 Factor m.	R^2 Factor*lang*algo m.
MATTR	.381	.935	.073	.956
prop. hapax	.370	.901	.056	.935
phones/w	.402	.941	.051	.961
syll/utt	.427	.975	.080	.983
entropy (NSA)	.431	.970	-.003	.980

Table 5

Percentage of variance explained (R^2) predicting F-scores (word and morpheme level) either from the factor given in the first column, or from language, algorithm, the factor and all their interactions in the second column. The R^2 for lang x algo alone is .91 for words and .94 for morphemes (the same for all rows). "lang" stands for language, "algo" for algorithm, "syll/utt" for syllables per utterance, "w" for words and "m" for morphemes.

5.5 Summary

In sum, we observed differences in language, level, and algorithm type in the expected directions. Overall *word* segmentation performance for Japanese was better than performance for Chintang. However, average Chintang scores improved on the *morpheme* level and this reduced the average score difference with Japanese. Surprisingly, factors we had postulated as proximal causes for word segmentability variation (type/token ratio, hapax ratios, word and utterance length) did not explain significant variance. The potentially confounded factor of segmentation entropy did not behave as predicted, suggesting that our original regression (with language, level, and algorithm) was sufficient. Oversegmentation rates were higher for Chintang than Japanese in the word level, especially for the lexical algorithm.

algo	Chin. words				Chin. morph.				Jap. words				Jap. morph.			
	ov	un	mis	cor	ov	un	mis	cor	ov	un	mis	cor	ov	un	mis	cor
AG	54	0	7	38	14	10	15	61	39	3	3	55	30	3	7	59
BTPa	28	9	11	53	4	47	2	47	23	13	2	61	19	16	3	62
FTPa	27	11	13	49	5	53	4	38	24	13	3	60	19	17	5	58
BTPr	7	20	29	45	2	70	5	23	9	25	11	55	6	31	11	52
FTPr	7	20	31	42	1	70	7	22	9	21	15	55	6	27	13	53
DiBS	6	47	17	30	7	47	18	27	2	49	6	43	2	51	7	41
Syll=Unit	68	0	0	32	21	0	0	79	53	0	0	47	49	0	0	51
Utt=Unit	0	90	0	10	0	96	0	4	0	68	0	32	0	71	0	29

Table 6

Percentage of oversegmented, undersegmented, missegmented and correctly segmented word and morpheme tokens for each algorithm, level, and language. "algo" stands for algorithm, "morph." stands for morphemes, "Chin." for Chintang and "Jap." for Japanese. Also, "ov" stands for oversegmentation, "un" for undersegmentation, "mis" for missegmentation and "cor" for correctly segmented.

6 Discussion

In this study, a set of algorithms was applied to corpora of two morphologically diverse languages, and the output was assessed against gold standard segmentation at the word *and* morpheme level. Given the details of Chintang morphology, it was hypothesized that such a rich morphology must pose significant problems for the uninformed learner who is trying to segment the input, but these problems would be mitigated if we consider morphemes instead if words as a segmentation goal.

Results summarized thus far support the prediction that languages varying in morphological complexity might vary in segmentability, but several aspects of these results strongly suggest that the answer is not simple. In our study, the language effect was not the same for all algorithms, and was even reversed when both algorithm and level were varied. The language effect was also smaller than the one for level or algorithm type. Indeed, performance for Chintang was substantially improved when evaluating on morphemes. In other words, differences within-language (across algorithms) seem to be more important than those between languages, and morphological complexity is far from being the sole or major determinant for segmentation.

To further address our research questions, we consider the results within each algorithm next. Our strongest predictions pertained to the lexical algorithm, AG, whose results matched our predictions well. In AG, we observed higher performance for Japanese than Chintang with words as the gold standard, but this difference was reversed with morphemes. This observation is consistent with the proposal that AG, and probably lexical algorithms in general, are ideal to recover recombinable units. Thus, it seems that lexical algorithms might work well for languages like Chintang, improving performance on the morpheme level.

Turning to the sublexical algorithms, even though DiBS is a purely phonotactic-based algorithm, it seems to have been affected by language differences. The algorithm was robust across evaluation levels, which had no impact on the segmentation performance. The most complex patterns of results were found for the

other sublexical algorithm, TP. All four versions of the algorithm (BTPa, FTPa, BTPr, FTPr) yielded divergent patterns. This is in accordance with previous findings, where notable differences in performance were found depending on the parameters used (e.g., Saksida et al., 2017). The language as well as the evaluation level had a significant effect on performance, and an interaction between language and level was observed for all versions.

This study attempted to associate segmentability to language features predicted to have an impact on segmentation. Word length, utterance length, and even corpus entropy explained only a small proportion of the structured variance, and none significantly beyond the factors of language and algorithm. Entropy, in particular, had been postulated in previous cross-linguistic work comparing English and Japanese (Fourtassi et al., 2013), languages which diverge both in morphology and phonotactics, whereas the languages studied here had similar phonotactics. Further work varying these parameters independently may be needed to pin down the importance of factors such as word length, utterance length, and corpus entropy (see Caines et al., 2019, for a similar approach on properties explaining variation in segmentation performance).

One surprising result pertains to the follow-up analyses which investigated the explanatory value of type/token ratio (measured as MATTR) and hapax ratio, which we had suggested as potentially proximal causes for segmentability differences across the languages. It seems that, while morphological complexity has an impact on segmentation, this might not be via the causal paths we had identified, since none could independently explain away the language effect in a multivariate regression. Future work may need to assess whether such variables jointly could explain away language differences, or whether this effect is due to other features that we have not yet considered (see Loukatou et al., 2019, for a similar result on differences between adult- and child-directed speech).

Before closing, we would like to bring up a number of limitations for this study. First, the research was conducted using transcriptions of speech spoken around (and not only to) children varying in age, from a few months to five years old, with the Japanese

children being in general older than the Chintang children. Further research with more homogeneous addressees may provide more stable results. Second, two different datasets compose the ACQDIV Japanese corpus, as mentioned in Section 2.2. This might be the cause for the ostensively more variable results of the Japanese subsets. Third, many Japanese utterances had not been morphologically transcribed, so they had to be excluded.

Moreover, since Chintang is spoken in a multilingual setting, annotators transcribed all speech including non-Chintang words, either because they are recent loanwords or because of code-switching into Nepali. Chintang speakers are bilingual in the morphologically simpler Nepali and children encounter Nepali from early on (Stoll et al., 2015). In fact, approximately 36% of the Chintang utterances had non-Chintang single- or multi-word insertions. For our analyses, we chose to report on the results for the whole corpus, because children born into this community do not come with information about which words are loanwords or code-switched. However, we also segmented a version of the corpus consisting of only all-Chintang utterances, where utterances with non-Chintang insertions had been removed. The performance usually increased by .01-.06 in token F-scores for the all-Chintang corpus, but did not alter our conclusions above. Detailed results can be found on the online supplementary material.

Also, speech transcriptions were used for this study. However, other salient features for segmentation include supra-segmental, speech-related features such as prosody or intonation. Even though there is some literature looking at word segmentation from speech (Ludusan et al., 2015), this task remains challenging for computational modeling.

Additionally, we would like to mention that the algorithms were evaluated on words, as defined by their conventional orthographic representations. However, wordhood and morphemehood are debated issues in linguistics and psycholinguistics, without cross-linguistically valid definitions, as mentioned in Section 1.2.

Finally, we studied two languages which differ in morphological synthesis, admittedly considering only one dimension of morphological complexity. We would

suggest that the level of allomorphy, meaning how many different realizations exist for a single morpheme, or the fusion of the affixes with each other, could also affect segmentation. Further research is needed to show the effects of these specific morphological aspects, although this would ideally involve recovery of morphological paradigms, and not just segmentation as done here.

Clearly, our work barely scratches the surface not only in terms of segmentation differences and similarities across languages, but also in terms of possible evaluation targets for language acquisition segmentation models. We look forward to further research incorporating more languages, in order to investigate the impact of different linguistic traits, and hope future work retains our strategy of employing a range of plausible algorithms and evaluating on different linguistic levels.

References

Aikhenvald, A. (2007). Typological distinctions in word-formation. *Language Typology and Syntactic Description, Volume III: Grammatical Categories and the Lexicon* (pp. 1–64). Cambridge University Press.

Allen, S. (1996). *Aspects of argument structure acquisition in inuktitut* (Vol. 13). John Benjamins Publishing.

Bartlett, S., Kondrak, G. & Cherry, C. (2009). On the syllabification of phonemes. *Proceedings of Human Language Technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics*, 308–316.

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

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J. & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., Reixachs, L. F., Dupoux, E., Daland, R., Cao, X. N. & Cristia, A. (2018). Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior research Methods*.

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

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, I. P., Rai, M., Rai, N. K. & Stoll, S. (2007). Free prefix ordering in Chintang. *Language*, *83*(1), 43–73.

Bickel, B. & Nichols, J. (2007). Inflectional morphology [2nd edition]. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 169–240). Cambridge University Press.

Bickel, B. & Nichols, J. (2013a). Fusion of selected inflectional formatives. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures*

Online. Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/20

Bickel, B. & Nichols, J. (2013b). Inflectional synthesis of the verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/chapter/22

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., Hildebrandt, K., Rießler, M., Bierkandt, L., ZúñIGA, F. & John, B. L. (2017). The AUTOTYP typological databases. *Version 0.1. 0. Online: https://github. com/autotyp/autotyp-data/tree/0.1. 0.*

Bickel, B. & Zúñiga, F. (2017). The 'word' in polysynthetic languages: Phonological and syntactic challenges. In M. Fortescue, M. Mithun & N. Evans (Eds.), *The Oxford Handbook of Polysynthesis* (pp. 158–185). Oxford University Press.

Black, A. & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 124–129.

Blanchard, D., Heinz, J. & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *37*(3), 487–511.

Börschinger, B. & Johnson, M. (2014). Exploring the role of stress in Bayesian word segmentation using Adaptor Grammars. *Transactions of the Association for Computational Linguistics*, *2*(1), 93–104.

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.

Boruta, L., Peperkamp, S., Crabbé, B. & Dupoux, E. (2011). Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation. *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics*, 1–9.

Brown, P. (1998). Children's first verbs in Tzeltal: Evidence for an early verb category. *Linguistics*, *36*(4), 713–754.

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.

Çöltekin, Ç. (2011). *Catching words in a stream of speech: Computational simulations of segmenting transcribed child-directed speech* (Doctoral dissertation). University of Groningen.

Cristia, A., Dupoux, E., Ratner, N. B. & Soderstrom, M. (2018). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, 1–10.

Daland, R. (2009). *Word segmentation, word recognition, and word learning: A computational model of first language acquisition* (Doctoral dissertation). 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? *Association for Computational Linguistics*, 873–877.

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

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.

Dixon, R. M. & Aikhenvald, A. Y. (2002). *Word: A cross-linguistic typology*. Cambridge University Press.

Dryer, M. S. & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. http://wals.info/

Dupoux, E., Pallier, C., Sebastian, N. & Mehler, J. (1997). A destressing "deafness" in French? *Journal of Memory and Language*, *36*(3), 406–421.

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. *Proceedings of the joint meeting of the Association for Computational Linguistics and the Human Language Technology Conference*, 130–138.

Fourtassi, A., Börschinger, B., Johnson, M. & Dupoux, E. (2013). WhyisEnglishsoeasytosegment. *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, 1–10.

Gaenszle, M., Bickel, B., Banjade, G., Lieven, E., Paudyal, N., Rai, A., Rai, I., Rai, M., Rai, N. K., Rai, V. S., Gautam (Sharma), N. P. & Stoll, S. (2005). Research report: the Chintang and Puma Documentation Project (CPDP). *European Bulletin of Himalayan Research*, (28), 95–103.

Gambell, T. & Yang, C. (2005a). Mechanisms and constraints in word segmentation. *Unpublished manuscript*.

Gambell, T. & Yang, C. (2005b). Word segmentation: Quick but not dirty. *Unpublished manuscript*.

Gervain, J. & Erra, R. G. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, *125*(2), 263–287.

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

Hallé, P. A., Durand, C. & de Boysson-Bardies, B. (2008). Do 11-month-old French infants process articles? *Language and Speech*, *51*(1-2), 23–44.

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

Hinds, J. (1986). *Japanese*. Routledge.

Höhle, B. & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, *6*(2), 122–127.

Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of memory and language*, *44*(4), 548–567.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using Adaptor Grammars. *Proceedings of the Tenth Meeting of Association for Computational Linguistics Special Interest Group on Computational Morphology and Phonology*, 20–27.

Johnson, M. & Demuth, K. (2010). Unsupervised phonemic Chinese word segmentation using Adaptor Grammars. *Proceedings of the 23rd International Conference on Computational Linguistics*, 528–536.

Johnson, M., Griffiths, T. L. & Goldwater, S. (2007). Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, 641–648.

Jusczyk, P. W., Jusczyk, A. M., Kennedy, L. J., Schomberg, T. & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 822–836.

Kastner, I. & Adriaans, F. (2017). Linguistic constraints on statistical word segmentation: The role of consonants in Arabic and English. *Cognitive Science*, *42*(2), 494–518.

Ketrez, F. N. & Aksu-Koç, A. (2009). Early nominal morphology in Turkish: Emergence of case and number. In M. Voeikova & U. Stephany (Eds.), *The development of nominal inflection in first language acquisition: A cross-linguistic perspective* (pp. 15–48). Mouton de Gruyter.

Kim, Y. J. (2015). *6-month-olds' segmentation and representation of morphologically complex words* (Doctoral dissertation). University of California, Los Angeles.

Kim, Y. J. & Sundara, M. (2015). Segmentation of vowel-initial words is facilitated by function words. *Journal of child language*, *42*(4), 709–733.

Kuno, S. (1973). *The structure of the Japanese language*. MIT press.

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. *Proceedings of the West Coast Conference on Formal Linguistics*, 237–247).

Loukatou, G., Le Normand, M. & Cristia, A. (2019). Is it easier to segment words from infant- than adult-directed speech? Modeling evidence from an ecological French

corpus. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

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. *Proceedings of the 55th annual Meeting of the Association for Computational Linguistics (volume 2: short papers)*, 178–183.

Ludusan, B., Seidl, A., Dupoux, E. & Cristia, A. (2015). Motif discovery in infant-and adult-directed speech. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 93–102.

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.

Marquis, A. & Shi, R. (2015). The beginning of morphological learning: Evidence from verb morpheme processing in preverbal infants. *Cognitive Science Perspectives on Verb Representation and Processing*, 281–297.

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*(4), 465–494.

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.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* (pp. 23–42). John Benjamins.

Mintz, T. H. (2013). The segmentation of sub-lexical morphemes in English-learning 15-month-olds. *Frontiers in Psychology*, *4*(24).

Miyata, S. (2004a). *Japanese: Aki Corpus*.

Miyata, S. (2004b). *Japanese: Ryo corpus*. Pittsburgh PA: TalkBank, 1-59642-056-1.

Miyata, S. (2004c). Japanese: Tai corpus.

Miyata, S. & Naka, N. (1998). Wakachigaki Guideline for Japanese: WAKACHI98 v.1.1. *The Japanese Society for Educational Psychology Forum Report No. FR-98-003, The Japanese Association of Educational Psychology.*

Miyata, S. & Naka, N. (2006). *JMOR03*. http://childes.psy.cmu.edu/morgrams/jpn.zip.

Miyata, S. & Naka, N. (2014). JMOR06.2: The Japanese morphological analysis program based on CLAN.

Miyata, S. & Nisisawa, H. (2009). *Japanese–MiiPro–Asato Corpus*.

Miyata, S. & Nisisawa, H. (2010). *Japanese–MiiPro–Tomito Corpus*.

Moran, S. & Cysouw, M. (2018). *The unicode cookbook for linguists: Managing writing systems using orthography profiles*.

Moran, S., Schikowski, R., Pajović, D., Hysi, C. & Stoll, S. (2016). The ACQDIV database: Min (d) ing the ambient language. *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, 4423–4429.

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.

Nichols, J. (2009). Linguistic complexity: A comprehensive definition and survey. In G. Sampson, D. Gil & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 109–124). Oxford University Press.

Nichols, J., Witzlack-Makarevich, A. & Bickel, B. (2013). The AUTOTYP genealogy and geography database: 2013 release. *Zurich: University of Zurich*.

Nisisawa, H. & Miyata, S. (2009). Japanese–MiiPro–Nanami Corpus.

Nisisawa, H. & Miyata, S. (2010). *Japanese–MiiPro–ArikaM Corpus*.

Norris, D., McQueen, J. M., Cutler, A. & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology*, *34*(3), 191–243.

Pearl, L. & Phillips, L. (2018). Evaluating language acquisition models: A utility-based look at Bayesian segmentation. In A. Villavicencio & T. Poibeau (Eds.),

Language, cognition, and computational models (pp. 185–224). Cambridge University Press.

Perfors, A. & Navarro, D. J. (2012). What Bayesian modelling can tell us about statistical learning: What it requires and why it works. *Statistical Learning and Language Acquisition*, *1*, 383–408.

Phillips, L. & Pearl, L. (2014a). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. *Proc. of 5th workshop on Cognitive Aspects of Computational Language Learning*, 9–13.

Phillips, L. & Pearl, L. (2014b). Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. *Proceedings of the Cognitive Science Society*, 2775–2780.

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

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

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

Ruzsics, T. & Samardzic, T. (2017). Neural sequence-to-sequence learning of internal word structure. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 184–194.

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

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

Saksida, A., Langus, A. & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, *20*(3), 1–11.

Samardzic, T., Schikowski, R. & Stoll, S. (2015). Automatic interlinear glossing as two-level sequence classification. *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, 68–72.

Sampson, G., Gil, D. & Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press.

Scherling, J. (2016). The creative use of English in Japanese punning. *World Englishes*, *35*(2), 276–292.

Schiering, R., Bickel, B. & Hildebrandt, K. A. (2010). The prosodic word is not universal, but emergent. *Journal of Linguistics*, *46*(3), 657–709.

Schikowski, R., Moran, S. & Stoll, S. (2018). Manual for the ACQDIV corpus.

Seidl, A., Cristià, A., Bernard, A. & Onishi, K. H. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, *5*(3), 191–202.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shi, R., Cutler, A., Werker, J. & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America*, *119*(6).

Shi, R. & Gauthier, B. (2005). Recognition of function words in 8-month-old French-learning infants. *The Journal of the Acoustical Society of America*, *117*(4), 2426–2427.

Shi, R. & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, *11*(3), 407–413.

Shi, R., Marquis, A. & Gauthier, B. (2006). Segmentation and representation of function words in preverbal French-learning infants. *Proceedings of the 30th Annual Boston University Conference on Language Development*, *2*, 549–560.

Shi, R., Werker, J. & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy*, *10*(2), 187–198.

Shi, R., Werker, J. F. & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–B21.

Shibatani, M. (1990). Japanese. In B. Comrie (Ed.), *The world's major languages* (pp. 855–880). Croom Helm.

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

Slobin, D. I. (1985). *The crosslinguistic study of language acquisition: Theoretical issues*. Lawrence Erlbaum Associates.

Slobin, D. I. (2014). Before the beginning: The development of tools of the trade. *Journal of Child Language*, *41*(S1), 1–17.

Stoll, S. (2015). Crosslinguistic approaches to language acquisition. In E. L. Bavin & L. R. Naigles (Eds.), *The Cambridge Handbook of Child Language* (pp. 107–134). Cambridge University Press. https://doi.org/10.1017/CBO9781316095829.006

Stoll, S. & Bickel, B. (2013). Capturing diversity in language acquisition research. In B. Bickel, L. Grenoble, D. Peterson & A. Timberlake (Eds.), *Language Typology and Historical Contingency: In honor of Johanna Nichols* (pp. 1–22). John Benjamins.

Stoll, S. & Lieven, E. (2014). Studying language acquisition cross-linguistically. In H. Winskel & P. Pradakannaya (Eds.), *South and Southeast Asian Psycholinguistics* (pp. 19–35). Cambridge University Press.

Stoll, S., Mazara, J. & Bickel, B. (2017). The acquisition of polysynthetic verb forms in Chintang.

Stoll, S. & Schikowski, R. (in press). Child corpora. In P. Magali & S. T. Gries (Eds.), *Practical handbook of corpus analysis*. Springer.

Stoll, S., Zakharko, T., Moran, S., Schikowski, R. & Bickel, B. (2015). Syntactic mixing across generations in an environment of community-wide bilingualism. *Frontiers in Psychology*, *6*, 82.

Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Blackwell.

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

Willits, J. A., Seidenberg, M. S. & Saffran, J. R. (2014). Distributional structure in language: Contributions to noun–verb difficulty differences in infant word recognition. *Cognition*, *132*(3), 429–436.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin.

4. Is word segmentation child's play in all languages?*

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

***Loukatou, G., Moran, S., Blasi, D., Stoll, S. & Cristia, A. (2019). Is word segmentation child's play in all languages?** ***Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics (ACL)*.**

***Loukatou, G. (2019). From phonemes to morphemes: relating linguistic complexity to unsupervised word (over) segmentation.** ***Typology for Polyglot NLP workshop, co-located with the 47th Annual Meeting of ACL*.**

Is word segmentation child's play in all languages?

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

Steven Moran
Damián E. Blasi
Sabine Stoll
University of Zurich

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

Abstract

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

1 Introduction

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

1.1 Unsupervised bottom-up segmentation across languages

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

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

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

1.2 The present work

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

As for languages, we used the ACQDIV

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

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

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

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

The present study addresses the following questions:

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

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

2 Methods

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 Results

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

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

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

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

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

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

4 Discussion

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

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

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

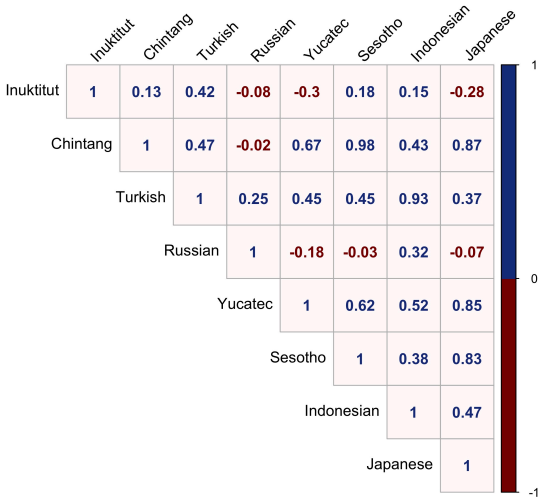

Figure 1: Correlation matrix of the rank ordering in algorithms' performance across languages.

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

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

To sum up, this study provides evidence that, if infants do anything similar to one or more of the algorithms proposed in previous natural language processing research and investigated here, then they would be well-equipped to get a head start in segmenting word-like units regardless of what their native language is. Experimental evidence suggests slight variation in the timing of acquisition of different linguistic features, as a function

of factors such as the transparency of forms, and the complexity of paradigms (e.g., Slobin 1985). Given the small differences found across our unsupervised word segmentation algorithms, such variation might come from something else, such as meaning acquisition, which would require algorithms different from the ones we explored here.

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

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

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

Acknowledgments

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

References

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

Shanley Allen. Unpublished. Allen Inuktitut Child Language Corpus.

Mathieu Bernard, Roland Thiolliere, Amanda Saksida, Georgia R. Loukatou, Elin Larsen, Mark Johnson, Laia Fibla Reixachs, Emmanuel Dupoux, Robert Daland, Xuan Nga Cao, and Alejandrina Cristia. 2018. Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior research Methods*.

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

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

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

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

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

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

Katherine Demuth. 2015. Demuth Sesotho Corpus.

Katherine A. Demuth. 1992. Acquisition of Sesotho. In Dan Isaac Slobin, editor, *The crosslinguistic study of language acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates, Hillsdale, NJ.

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

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. WhyisEnglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10.

Judit Gervain and Ramón Guevara Erra. 2012. The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125(2):263–287.

David Gil and Uri Tadmor. 2007. The MPI-EVA Jakarta Child Language Database. a joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology and the Center for Language and Culture Studies, Atma Jaya Catholic University.

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

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

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

Mark Johnson. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27. Association for Computational Linguistics.

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

Aylin C. Küntay, Dilara Koçbaş, and Süleyman Sabri Taşçı. Unpublished. Koç university longitudinal language development database on language acquisition of 8 children from 8 to 36 months of age.

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

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

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

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

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

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

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(3):545–564.

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

Steven Moran, Robert Schikowski, Danica Pajović, Cazim Hysi, and Sabine Stoll. 2016. The ACQDIV database: Min(d)ing the ambient language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

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

Barbara Pfeiler. Unpublished. Pfeiler Yucatec Child Language Corpus.

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

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

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

Amanda Saksida, Alan Langus, and Marina Nespor. 2017. Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20(3):1–11.

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

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

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

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

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

Sabine Stoll, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski, and Balthasar Bickel. 2015. Audiovisual corpus on the acquisition of Chintang by six children.

Sabine Stoll and Roland Meyer. 2008. Audiovisional longitudinal corpus on the acquisition of Russian by 5 children.

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

Appendix

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

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

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

From phonemes to morphemes: relating linguistic complexity to unsupervised word over-segmentation

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

Abstract

Previous work has documented variation in word segmentation performance across languages, with a trend to yield lower scores for languages with elaborate morphological structure. However, segmenting smaller chunks than words, "oversegmenting", is reasonable from a computational point of view. We predict that oversegmentation would be encountered more often in complex languages. In this work in progress, we use a dataset of 9 languages varying in complexity and focus on cognitively-inspired word segmentation algorithms. Complexity is defined by Compression-based, Type-Token Ratio and Word Length metrics. Preliminary results show that a possible relation between morphological complexity and oversegmentation cannot be predicted exactly by none of these metrics, but may be best approximated by word length.

1 Introduction

The issue of word segmentation is open in the NLP community (e.g., Harris (1955)). Its implementations include processing languages with no orthographic word boundaries, such as Chinese and Japanese. It is also a key problem humans face when acquiring language.

Previous work documented variation in the success rate of segmentation across languages, and a trend to yield lower scores for languages with elaborate morphological structure. This is true for both cognitively inspired (Johnson, 2008; Fourtassi et al., 2013; Loukatou et al., 2018)) and other models (Mochihashi et al., 2009; Zhikov et al., 2013; Chen et al., 2011). Evaluation is conventionally based on orthographic word boundaries.

Do these models manage to learn more linguistic structure, that what is actually described in these accuracy scores? Segmenting smaller meaningful chunks than words is reasonable from a computational point of view: morphologically complex languages often feature multimorphemic, long words, and algorithms might break words up into component morphemes, treating frequent morphemes as words. Finding out morphemes might be useful for later linguistic analysis, especially for languages with rich morphological systems , and such morphemes could be used as cues to further bootstrap segmentation. Thus, a "useful" error in segmentation could be oversegmentation (Gervain and Erra, 2012; Johnson, 2008), the percentage of word tokens returned as two or more subparts in the output.

We thus predict that oversegmentation might be encountered more often in complex languages. To test this, we need data from languages varying in complexity. Since there is no standard way to define complexity, for this study, three metrics are used: first, the Moving Average Type-token Ratio (500-word window) (Kettunen, 2014), and second, two versions of compression-based complexity (Szmrecsanyi, 2016)[1]. The two metrics are normalized (0=least complex, 1=most complex) and their average score is attributed to each language. Third, we look at word length, since, in general, longer words could attract more division.

2 Methods

We use the ACQDIV database (Moran et al., 2016) of typologically diverse languages, with transcriptions of infant-directed and -surrounding speech recordings, from Inuktitut (Allen, 1996), Chintang (Stoll et al., 2015), Turkish (Küntay et al., Unpub-

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

lang	% over	% corr	% total	compr.	MATTR	w length
Inuktitut	42	17	59	1	0.90	8.56
Chintang	26	25	51	0.56	0.87	4.39
Turkish	26	25	51	0.44	0.86	4.92
Yucatec	19	27	46	0.42	0.92	3.80
Russian	29	22	51	0.41	0.91	4.47
Sesotho	26	24	50	0.31	0.86	4.28
Indonesian	25	29	54	0.28	0.85	4.11
Japanese	20	26	46	0.14	0.87	3.94
English	6	51	57	0.02	0.39	3.04

Table 1: Percentage of average oversegmented, correct word tokens and their sum are given per language in the first columns. Complexity scores for the three metrics are also given.

lished), Yucatec (Pfeiler, 2003), Russian (Stoll and Meyer, 2008), Sesotho (Demuth, 1992), Indonesian (Gil and Tadmor, 2007) and Japanese (Miyata and Nisisawa, 2010; Nisisawa and Miyata, 2010). In order to compare with a previously studied language, we included the English Bernstein corpus (MacWhinney, 2000).

Several models have been proposed as plausible strategies used by learners retrieving words from input. We used a set of these strategies (Bernard et al., 2018). Two baselines were Base0, treating each sentence as a word, and Base1, treating each phoneme as a word. DiBS[2] (Daland, 2009) implements the idea that unit sequences often spanning phrase boundaries probably span word breaks. FTP[3] (Saksida et al., 2017) measures transitional probabilities between phonemes and cuts depending on a local threshold (relative, FTPr) or a global threshold (absolute, FTPa). Adaptor Grammar (AG) (Johnson, 2008) assumes that learners create a lexicon of minimal, recombinable units and use it to segment the input. AG implements the Pitman-Yor process. Finally, PUDDLE[4] (Monaghan and Christiansen, 2010) is incremental, and learners insert in a lexicon an utterance that cannot be broken down further, and use its entries to find subparts in subsequent utterances. Before segmentation, spaces between words were removed, leaving the input parsed into phonemes, with utterance boundaries preserved.

3 Results

Statistics regarding corpora and results are presented in Table 1. In general, languages had similar oversegmentation scores, (ranging from 31% to 51% if we exclude English), which did not exactly follow their complexity ranking. Performance difference across languages decreased when considering oversegmented tokens as correctly segmented.

4 Discussion

Word length had the best prediction of oversegmentation compared to other metrics, compression and MATTR. This shows that longer words have more alternative parses, and this could explain oversegmentation results better than other properties inherent to morphologically complex languages. That said, a possible relation between morphological complexity and oversegmentation, could not be *exactly* explained by none of these complexity metrics.

It was also observed that there was no absolute ranking of complexity across languages; on the contrary, it would change according to the feature studied. In general, cross-linguistic differences were small for such a typologically distinct dataset of languages. Further research might shed light on whether this behavior is due to linguistic properties common across languages, or a confound (e.g. corpus size).

Moreover, discovering meaningful units is of particular importance to language acquisition models, such as the ones implemented here. Infant word segmentation algorithms are cognitively plausible only if they are cross-linguistically valid and offer useful insights to learn all linguistic structures. It would also be interesting to compare performance of these models to state-of-the-art NLP algorithms, such as HPYLM (Mochihashi et al., 2009) or ESA (Chen et al., 2011).

A limitation of this study is that the current implementation of WordSeg does not only look at oversegmentation cases resulting in meaningful, morpheme-like sub-parts. A next step would be to focus on reasonable oversegmentation errors, even though not all of these corpora have morpheme annotations.

Measuring reasonable errors such as oversegmentation could shed light on the segmentability of morphologically complex languages and the cross-linguistic applicability of models. Further research might include over-, but also undersegmentation errors, when two or more words in the input returned as a single unit in the output.

[2]Diphone Based Segmentation algorithm

[3]Forward Transitional Probabilities algorithm

[4]Phonotactics from Utterances Determine Distributional Lexical Elements

References

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

Mathieu Bernard, Roland Thiolliere, Amanda Saksida, Georgia Loukatou, Elin Larsen, Mark Johnson, Laia Fibla Reixachs, Emmanuel Dupoux, Robert Daland, Xuan Nga Cao, and Alejandrina Cristia. 2018. Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior research Methods*.

Songjian Chen, Yabo Xu, and Huiyou Chang. 2011. A simple and effective unsupervised word segmentation approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

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

Katherine A. Demuth. 1992. Acquisition of sesotho. In Dan Isaac Slobin, editor, *The crosslinguistic study of language acquisition*, volume 3, pages 557–638. Lawrence Erlbaum Associates, Hillsdale, NJ.

Abdellah Fourtassi, Benjamin Börschinger, Mark Johnson, and Emmanuel Dupoux. 2013. WhyisEnglishsoeasytosegment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10.

Judit Gervain and Ramón Guevara Erra. 2012. The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125(2):263–287.

David Gil and Uri Tadmor. 2007. The mpi-eva jakarta child language database. a joint project of the department of linguistics, max planck institute for evolutionary anthropology and the center for language and culture studies, atma jaya catholic university.

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

Mark Johnson. 2008. Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27. Association for Computational Linguistics.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Aylin C. Küntay, Dilara Koçbaş, and Süleyman Sabri Taşçı. Unpublished. Koç university longitudinal language development database on language acquisition of 8 children from 8 to 36 months of age.

Georgia Loukatou, Sabine Stoll, Damian Blasi, and Alejandrina Cristia. 2018. Modeling infant segmentation of two morphologically diverse languages. *TALN*.

Brian MacWhinney. 2000. *The CHILDES project: tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.

Padraic Monaghan and Morten H Christiansen. 2010. Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, 37(3):545–564.

Steven Moran, Robert Schikowski, D Pajović, Cazim Hysi, and Sabine Stoll. 2016. The ACQDIV database: Min (d) ing the ambient language. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, pages 4423–4429.

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

Barbara Pfeiler. 2003. Early acquisition of the verbal complex in yucatec maya. *Development of verb inflection in first language acquisition*, pages 379–399.

Amanda Saksida, Alan Langus, and Marina Nespor. 2017. Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental Science*, 20(3):1–11.

Sabine Stoll, Elena Lieven, Goma Banjade, Toya Nath Bhatta, Martin Gaenszle, Netra P. Paudyal, Manoj Rai, Novel Kishor Rai, Ichchha P. Rai, Taras Zakharko, Robert Schikowski, and Balthasar Bickel. 2015. Audiovisual corpus on the acquisition of chintang by six children.

Sabine Stoll and Roland Meyer. 2008. Audio-visional longitudinal corpus on the acquisition of russian by 5 children.

Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, 57:71.

Valentin Zhikov, Hiroya Takamura, and Manabu Okumura. 2013. An efficient algorithm for unsupervised word segmentation with branching entropy and mdl. *Information and Media Technologies*, 8(2):514–527.

5. Segmenting word and sub-word units in an artificial language experiment *

Abstract: The purpose of this study is to investigate segmentation in two meaningful unit levels, words and stems, among adults exposed to an artificial language. The structure of the language has properties similar to those of most natural languages; meaningful units are not only words, but also stems and affixes. We ask whether human adults can segment both words and stems in these settings.

***Loukatou, G. & Cristia, A. Segmenting word and sub-word units in an artificial language experiment. [in writing] - Preregistration: https://osf.io/fuydc**

Segmenting words and morphemes in an artificial language

Introduction

Humans segment word-like units out of continuous streams of speech when exposed to artificial languages. Words are important components of linguistic structure, but they are not the only meaningful, recombinable units in running speech. In human languages, morphemes are minimal meaningful units, and they also need to be segmented during language learning.

Word segmentation is an important learning task, were word boundaries are identified in continuous speech. Artificial languages have been widely used to investigate word segmentation (e.g. Saffran et al., 1996 and follow-up work). It is assumed that some learning mechanisms are shared between artificial and natural language learning (Gómez & Gerken, 2000; Reber, 1967). When used and interpreted properly, artificial languages can help obtain better experimental control over the input to which learners are exposed (Fedzechkina et al., 2016; Folia et al., 2010), and isolate specific learning factors (Hayakawa et al., 2020), especially for segmentation.

The current standard for artificial language studies on segmentation is to focus on words as the target level of segmentation (e.g. Cunillera et al., 2010; Estes & Lew-Williams, 2015; Finn & Hudson Kam, 2008; Frank et al., 2013; Johnson & Jusczyk, 2001; Karuza et al., 2013; Kurumada et al., 2013; Lew-Williams & Saffran, 2012; Saffran et al., 1997; Thiessen & Erickson, 2013; Tyler & Cutler, 2009 -but see segmentation of multi-word units by Siegelman & Arnon, 2015). In most studies, the words composing the artificial language have no sub-units (stems or affixes) that could be found within other words. Participants seem to be able to segment words out of artificial languages based on several different cues (transitional probabilities cues, Saffran et al., 1996; speech cues, Johnson & Jusczyk, 2001; mapping to word referents, Cunillera et al., 2010).

However, some studies suggest that word segmentation is not the only task participants do, when exposed to an artificial language. They can segment speech, but also generalise non-adjacent dependencies (Frost & Monaghan, 2016), map the word forms to word referents (Cunillera et al., 2010) or learn the overall linguistic structure (Siegelman & Arnon, 2015). Moreover, properties related to linguistic structure, such as non-adjacent dependencies and the number of different words seem to affect segmentation (Frank et al., 2010; Frost & Monaghan, 2016).

There has been an effort to investigate word segmentation in artificial languages with structures that resemble more those of natural languages (e.g. variable word length and frequency, Frank et al., 2010; Hoch et al., 2013; Johnson & Tyler, 2010; Kurumada et al., 2013; Schuler et al., 2017). The findings suggest that word segmentation can indeed be affected by morphological features. One feature that may also affect segmentation, is the existence of affixes that cannot stand alone. Many natural languages have this -- in fact, English does have a few affixes that cannot stand alone, such as "s" in "stands". This cross-linguistically frequent feature, could also be captured by artificial languages, and discovered by listeners when segmenting meaningful units. However, we are unaware of a study specifically testing whether listeners do segment out stems separately from affixes.

Previous literature on morpheme learning is mostly based on behavioural studies not related to segmentation. Several studies show that both adults and children can parse non-words containing both stems and suffixes (Fedzechkina et al., 2012; Finley & Newport, 2011; Finley & Wiemers, 2013; Hudson Kam & Newport, 2009). School-aged children (Finley & Newport, 2011) and adults (Finley & Newport, 2010) used distributional information within words to segment them into stems and suffixes. It was suggested that the participants inferred a pattern within words, and did not simply memorize whole words. Even 12-month-old children could distinguish between grammatical and ungrammatical sequences after exposure to an artificial language (Gomez & Gerken, 1999), which supports a more general pattern-based abstraction in artificial language learning (R. Gomez et al., 2000; Marcus, 1999). All of this work strongly suggests participants are analyzing the words they are exposed to, but they may not be segmenting items out specifically (for instance, they may instead calculate distances with known tokens, rather than extracting the parts).

It stands to reason that learners *do* segment morphemes. Morphemes, like words, are building blocks of language and the ability to extract them is fundamental. Nonetheless, it is still unclear whether humans first focus on bigger (words) or smaller (stems and affixes -- henceforth "morphemes") blocks, or whether they process both levels in a complementary way. For instance, infants start segmenting word-like units into a lexicon by the age of 6-8 months (Bergelson & Swingley, 2013), but they also seem to recognize morphemes early on in speech, in experimental settings. For example, children learning French can parse verbs into stems and suffixes by 11 months of age (Marquis & Shi, 2015) and children learning English segment suffixes at 15 months of age (Mintz, 2013, see also Gomez & Gerken, 1999; Mintz et al., 2002). Children learning Hungarian, an agglutinative language with rich morphology, can decompose new words into stems and suffixes by 15 months of age (Ladányi et al., 2020). Those authors argue for the relevance to focus on morphemes in early language

learning, and suggest that the existence of a large number of morphemes in a language may even trigger a more analytic processing.

Despite the wealth of evidence that morphemes may also be readily segmented, no previous study has looked at both word and morpheme segmentation, even though the aforementioned literature strongly suggests that humans should process both words and morphemes in language learning. Consequently, it is still unclear whether humans can segment morphemes out of running speech in an artificial language, and whether, if exposed to a language with both words and morphemes, they would succeed better in segmenting one or the other.

This study

The purpose of this study is to investigate segmentation in two meaningful unit levels, words and stems, among adults exposed to an artificial language. The structure of the language has properties similar to those of most natural languages; meaningful units are not only words, but also stems and affixes (which we will call "morphemes" here). We can then ask whether human adults can segment both words and stems in these settings. The study is preregistered in https://osf.io/fuydc. We note that due to the COVID-19 crisis, we were unable to complete the study. Therefore, we report fully on the design and planned analyses, but only report results for two pilots.

Three key questions will be addressed in this study:

1. Can participants segment out whole words in a language where there are affixes? If participants can segment out words, then they will choose words more than non-words -- items that are not structurally words or morphemes. If there is no preference between the two, this would mean that participants do not segment out words.
2. Can participants segment out stems in a language where there are affixes? If participants can segment out stems, then they will choose stems more than non-stems -- items that are not structurally words or morphemes. If there is no preference between the two, this would mean that participants do not segment out morphemes.
3. Which units, words or morphemes, are better segmented? If participants segment words better than stems, there will be a difference in the two tests above in terms of preference: They will prefer words to non-words more than they prefer stems to non-stems.

Predictions

The predictions of the three key questions are the following:

1. We predict that participants will segment out words in a language where words are composed of morphemes. This is predicted, because in previous artificial language studies word segmentation has been successful in other languages.
2. We predict that participants will segment out stems in a language where words are composed of morphemes. There are no previous studies on morpheme segmentation of an artificial language. This prediction is based on previous findings showing that humans are sensitive to morphological properties, in artificial and natural language settings (e.g. language acquisition).
3. We predict that participants will segment equally well stems and words. No previous studies have looked at segmentation of both words and morphemes. This prediction is based on previous findings that humans can rely on several available cues to understand a language, sometimes performing more than one task at the same time.

Methods

Participants will hear sentences of an artificial language, where words contain more than one morpheme. Their preference for words versus non-words and for stems versus non-stems will then be measured. All materials and scripts can be found in OSF LINK.

Building the language

Each participant is assigned to one of four artificial languages, A, B, C or D. The languages were generated using different orders of syllable concatenation, in order to control for any effects that could result from a specific concatenation. A counterbalancing procedure was used to create the languages. Language A was generated with random concatenation of eighteen syllables. The syllables were "glu", "sin", "ga", "kli", "ten", "ko", "blu", "tun", "man", "blo", "ti", "gle" ,"da", "pun", "go", "kan", "fen" and "bi", and were chosen so that no syllable is a word in French. The first five syllables were used to create three noun stems, the next five syllables were used to create three verb stems, and the following syllables were used for the optional elements, the noun and verb singular and plural affixes and the aspect affixes. Each syllable would have only one use in the vocabulary.

Language B was generated by inverting the order of the syllables (for example, the last syllable “bi” would go first, the second-to-last syllable “fen” would go second). Language C was created by inverting the order of syllables from the middle to the beginning, and then from the end to the middle (for example, the syllable “blo” would go first, the syllable “man” second, the syllable “ti” last). Language D was created by starting from the middle to the end, and then from the beginning to the middle (for example, “blo” would go first, “ti” would go second, “man” last).

The text was converted to speech using the mac Speech Synthesis tool. Specifically, the Mexican voice was used because the voice has a flat prosody, and in order to avoid accidental insertion of prosodic cues in a language the participants may be exposed to. The speed of speech was fixed to 140.

Training

The auditory stimuli are presented as utterances with clearly marked utterance boundaries, as cued both by silence and the end of a scene). Words and morphemes have meanings. In order to underline the existence of morphemes in the language, morphemes have meanings. The participant listens to a sentence while watching a video portraying the action described in the sentence. For example, if a sentence contains a ‘dog’ noun stem, and a ‘walk’ verb stem, the video would show a dog walking. The noun stem is followed by a number suffix, and the video shows one or two dogs. The verb stem is followed by an aspect suffix, and the video shows a dog walking continuously or once. Participants are not given the sentence in writing, nor any breakdown of the stems versus suffixes, nor of the conceptual morphemes just mentioned.

The structure of each sentence is portrayed in Figure 1. Scenes of the videos are pictured in Figure 2. Each sentence consists of a noun stem with its number affix and of a verb stem with an aspect affix and a number affix (not the same one as for the noun). The sentence can have one optional element (akin to an adverb because it can occur at the beginning or the end, but having no meaning). The training phase contains in total 96 sentences. The content of the sentences is built to allow for certain requirements to be respected during the test phase.

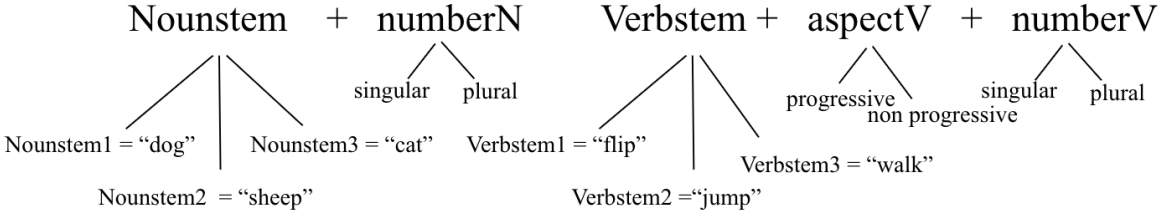

Figure 1. The utterance structure.

One requirement is having the same sum of transitional probabilities (TPs) between syllables for both items in a test trial. Statistical learning using TPs is a primary source of evidence for segmentation in laboratory experiments for infants (Estes & Lew-Williams, 2015; R. L. Gomez, 2012; Pelucchi et al., 2009; Romberg & Saffran, 2010) and adults (Frank et al., 2010; Perruchet & Desaulty, 2008; Toro et al., 2005). In those studies, participants segment out words based on TPs.

However, since in this experiment we do not test the well-established use of TPs in segmentation, the TP information should be controlled for in the test. To this end, we adjusted the combination of stems and affixes during training. Stems have TP=1. When comparing a stem to a non-stem, some meaningless pair of syllables appearing next to each other should also have TP=1. This happens by presenting some verb stem with the same aspect affix. For example, the noun stem (‘dog’) appears systematically in singular, and the noun stem (‘sheep’) appears systematically in plural. Similarly, the verb stem (‘flip’) appears systematically with a non-progressive affix, and the verb (‘jump’) with a progressive affix. For these two noun and verb stems, TP between stem and affix equals 1. The noun (‘cat’) and verb (‘walk’) appear with all combinations of affixes and in the same number of times for each affix. For them, TP between stem and affix equals 0.5. The average TP between a noun affix for number and the first syllable of a verb stem is 0.33.

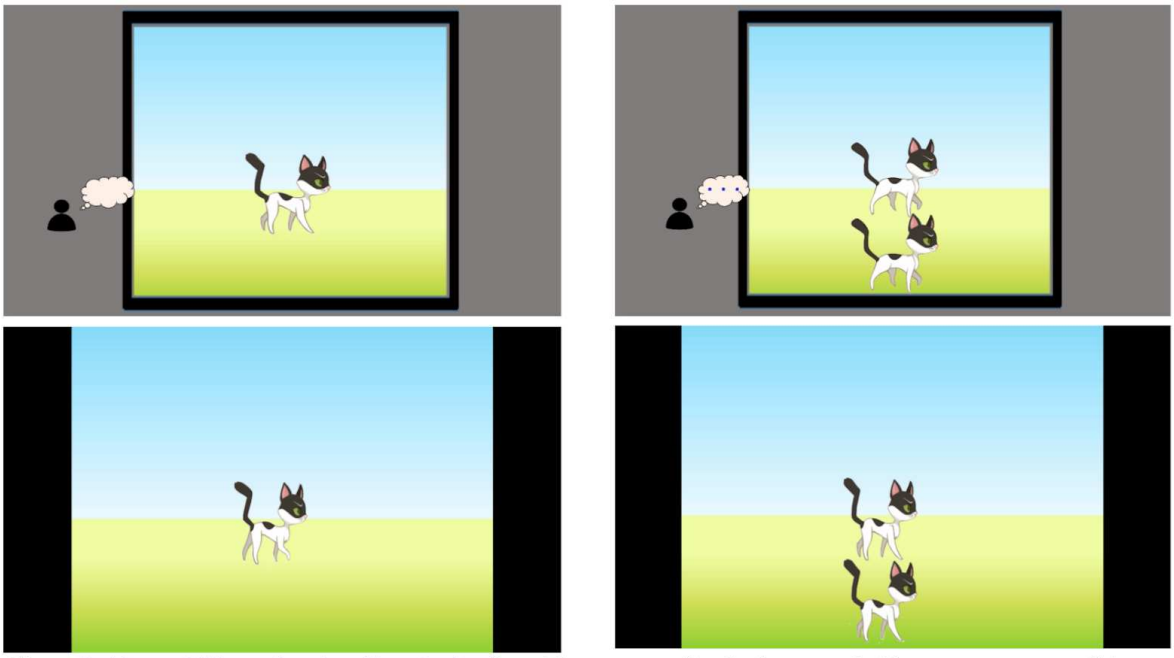

Figure 2. Above: Videos after the pilots. In the first video, a cat walks. In the second video, two cats are walking. Below: The same videos before the pilots.

Testing

In total, the test phase consists of 30 paired-forced-choice trials. Videos are not shown during the test phase, and stimuli are presented purely auditorily. Fourteen trials have word and non-word pairs (six nouns versus non-nouns + eight verbs versus non-verbs). Sixteen trials have stem and non-stem pairs (eight noun stems versus non-noun stems + eight verb stems versus non-verb stems). A noun word is a noun stem with an affix indicating number. A verb word is a verb stem with an affix indicating aspect and an affix indicating number. Stems are presented bare.

Once the familiarization period is over, participants are asked the following: ‘Vous allez maintenant entendre 2 sons. Quel son ressemble le plus à la langue que vous venez d'entendre? Ne réfléchissez pas trop et allez-y avec votre instinct!" (“You are now going to hear two sounds, which one sounds better for the language you just heard? Don’t overthink it and follow your gut”). They then need to press a button after hearing each trial: the left button if they think that the first item works better, or the right button for the second sound.

At each trial, the participant hears a correct and a false stimulus. Several conditions are respected during the test. Each test item has at least 2 syllables, to avoid missing or mishearing a (very short) sound. Both items have the same length, but also result in the same sum of (forward) Transitional Probabilities, and have the same frequency (or, rarely, the incorrect option has a larger frequency than the correct option). This way, no other cue could affect the preference of one versus the other item, other than the preference for a meaningful versus a meaningless unit.

Participants

Participants should be at least 18 years old, with French as their native language. The experiment lasts 20 minutes and they will be paid for participation. The study was initially designed to be tested in the lab, however, due to the current sanitary situation, it will be updated for online testing. Our target size is 52 participants. We estimated that this would be a sufficient number of participants, given that previous similar studies using artificial language segmentation experiments with adults found and average effect size of 0.46 (Cunillera et al., 2010; Frost & Monaghan, 2016; Hoch et al., 2013; Perruchet & Desaulty, 2008; Toro et al., 2005; Tyler & Cutler, 2009). Eligible for inclusion are all subjects who complete the experiment (meaning that they answer all test questions) and are not interrupted by an external factor (for example someone enters the room) during the experiment. In the

online version of the experiment, a test for attention will be included. Outliers will not be excluded in the analysis.

Statistical Analysis

For the analysis, we will fit a generalized linear mixed effect model using the lme4 library in R[1](R studio team, 2015). The participants' answers (correct/wrong) are the dependent variable, the level (stem/word) and number of trial (1st, 2nd…) are fixed variables. The random effect of the participant is included, with level and number of trial as random slopes.

Results

Prior to the onset of the COVID-19 crisis, two pilot studies were conducted, each with eleven participants. The first pilot indicated three issues with the study. First, many participants reported that the successive presentation of similar sounds during the test phase was distracting. Second, some participants felt that the familiarization period was too long. Third, some participants did not notice a difference in verb aspect in the videos, e.g. they considered the 'walk' and 'walking' videos as describing the exact same action. We addressed all issues in a second pilot. The repetition of the same stimulus or of a stimulus with the same stems is now avoided in successive trials. A message to the participant appears in the screen after completing 25%, 50% and 75% of the training, congratulating them for completing the corresponding part of the study. The duration of the action in the videos with progressive actions was increased. However, the latter effect persisted for some participants in the second pilot, so we have now added a "narrator" figure next to the videos.

We present next an analysis of the quantitative results of the second pilot. Figure 1 and 2 show the distribution of correct and wrong answers for the two levels (word and stem). Visualisation of the results shows that there were more correct than wrong answers for both levels. With respect to the three key questions, we observe that:

1. Based on the intercept in a regression with the word level as baseline, the word trial coefficient is 0.83 (SE=0.436, p-value=0.057). Thus, recognition of words is marginally significant in this pilot.
2. Based on the intercept of a regression with the stem level as baseline, the stem trial coefficient is 0.633 (SE=0.342, p-value=0.064). Thus, recognition of stems is marginally significant in this pilot.

[1] *glmer(formula= thisresplog ~ level + numberoftrial +(1 + level+ numberoftrial|uniqueid), control = glmerControl(optimizer = "bobyqa"), family = binomial(link = "logit"), data = pilot_long_log)*

3. Based on the level as a fixed effect in a regression with the stem level as baseline, the coefficient for level is 0.197 (SE=0.261, p-value=0.451). Thus, recognition of words versus stems is not significantly different in this pilot. The trend is for more accurate responses for word trials than stem trials.

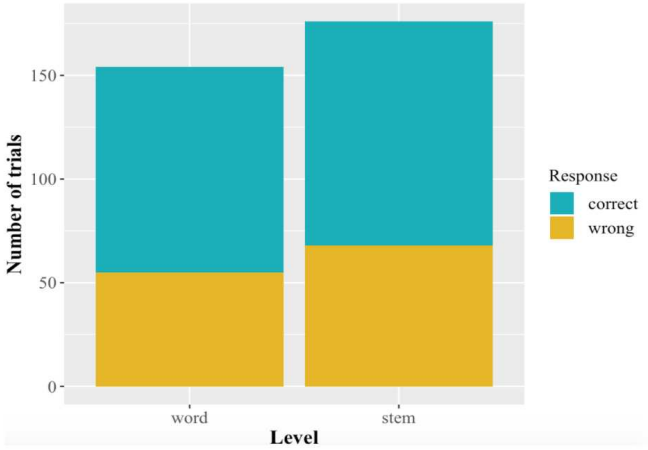

Figure 3. Number of total correct and wrong answers per level for all participants. There are slightly more trials for stems than words (14 word trials and 16 stem trials per participant).

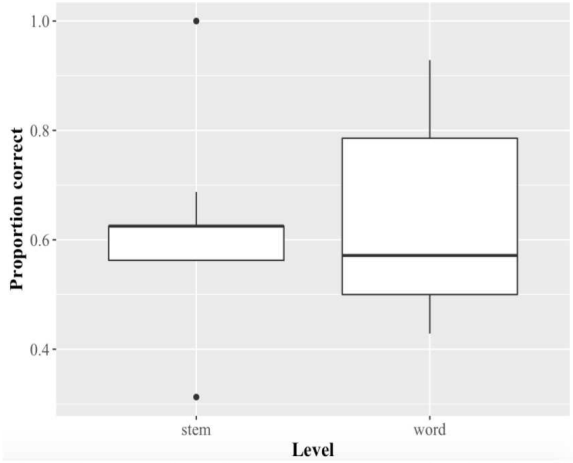

Figure 4. Proportion of correct answers per level and participant.

In a further exploratory analysis, we ask whether there is a difference in segmentation between verbs and nouns. Verbs might be more difficult to segment than nouns, for example due to the presence of more morphemes per word on average. We investigate this by including type (noun /verb) as a fixed effect variable and as a random slope. In the pilot results, the type coefficient is -0.268 (SE=0.407, p-value=0.51). It seems that nouns and verbs have a similar level of difficulty.

Discussion

Before analysing the results, we need to emphasize that only eleven subjects participated in the pilot study analyzed. Consequently, this analysis is underpowered. We aim to resolve this issue in the main study. Nevertheless, we attempt a tentative interpretation of the preliminary results below.

Based on the results for the first key question of the study, we observed that participants tend to prefer words to non-words, meaning that they segment words out of running speech in this artificial language. These results agree with our predictions. However, what makes the difference with previous studies is that participants were not based on TP or other specific cues (e.g. length, frequency) to choose the correct answer, as both words and non-words in each test trial had the same cues. The participants would only be able to segment words and morphemes out of speech after identifying them as meaningful units of the language, through a more abstract process of pattern recognition and/or association with meaning.

Based on the results for the second key question of the study, we also observed that participants tend to prefer stems to non-stems, meaning that they segment morphemes out of running speech in this artificial language. These results agree with our predictions. Similarly as above, participants could not have been based on TP or other specific cues during the test, since these cues were controlled for. This result is consistent with the idea that participants consider minimal meaningful (or at least recombinable) units when hearing an unknown language.

Importantly, the results for the third key question of the study show that, when participants get (briefly) exposed to an artificial language, they focus on at least two levels, words and morphemes, in order to extract meaningful units of the language. This corresponds to previous observations that humans exposed to languages with natural language characteristics and thus rich structure, can perform more than one task simultaneously, and may take advantage of the rich morphosemantic information of a language, in order to decipher the input. Similar findings are crucial for language acquisition, as children are also frequently exposed to natural languages with rich structure. If humans are capable of attending to more than one level and segmenting out morphemes, then children would also have a head start when exposed to such natural languages.

Last, no significant differences were found when segmenting nouns or verbs, showing that the two word classes were overall processed in a similar way. However, visualisation of the results shows larger variance in answers for the word than the morpheme level. This should be investigated further

in the main study. Moreover, if the main study confirms the patterns observed in the pilot, we plan to conduct future studies, in order to further test whether participants successfully associate unit forms with their referents in the video, and whether participants can successfully segment out morphemes and words after addition of more morphemes. If the main study disproves the patterns observed in the pilot, we plan to facilitate the task, by removing some morphemes. Planned conditions of the main study also include checking the relevance of meaning in segmentation, by removing the videos (and all their semantic information) and presenting the language in an audio-only condition.

Last, some limitations of the study should be mentioned. Our artificial language, even though inspired by natural language characteristics, is simple enough to be learnable after a short exposure. The goal of this study is to inform questions concerning the relationship between language segmentation and language structure, and to invite for further research on both morpheme and word segmentation, in both artificial and natural language settings. Moreover, we could measure the learning output after exposure to the language, but we cannot be sure what mechanisms were used by the participants to segment it. Further research can focus on this aspect, for example by modeling language learning with specific segmentation mechanisms, in order to compare the results.

References

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397. https://doi.org/10.1016/j.cognition.2013.02.011

Cunillera, T., Laine, M., Càmara, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language*, *63*(3), 295–305.

Estes, K. G., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1517–1528. https://doi.org/10.1037/a0039725

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902. https://doi.org/10.1073/pnas.1215776109

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case ... Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking. *Cognitive Science*. https://doi.org/10.1111/cogs.12346

Finley, S., & Newport, E. L. (2010). Morpheme segmentation from distributional information. In *Boston University Conference on Language Development (BUCLD) Online Proceedings Supplement*.

Finley, S., & Newport, E. L. (2011). *Morpheme segmentation in school-aged children*. University of Rochester Working Papers in the Language Sciences.

Finley, S., & Wiemers, E. (2013). Rapid learning of morphological paradigms. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Folia, V., Uddén, J., De Vries, M., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. *Language Learning*, *60*(SUPPL. 2), 188–220. https://doi.org/10.1111/j.1467-9922.2010.00606.x

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

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and Long-Term Retention of

Large-Scale Artificial Languages. *PLoS ONE*, *8*(1), e52500. https://doi.org/10.1371/journal.pone.0052500

Frost, R. L. A., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74. https://doi.org/10.1016/j.cognition.2015.11.010

Gómez, R., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186. https://doi.org/10.1016/S1364-6613(00)01467-4

Gomez, R., Gerken, L., & Schvaneveldt, R. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition*, *28*(2), 253-263.

Gomez, R. L. (2012). *Rebecca Gómez, Statistical learning in infant language development*. In *The Oxford handbook of psycholinguistics*. Oxford University Press.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Hayakawa, S., Ning, S., & Marian, V. (2020). From Klingon to Colbertian: Using Artificial Languages to Study Word Learning. *Bilingualism: Language and Cognition*, *23*(1), 74–80. https://doi.org/10.1017/S1366728919000592

Hoch, L., Tyler, M. D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic Bulletin & Review*, *20*(1), 142–147. https://doi.org/10.3758/s13423-012-0309-8

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*(4), 548–567. https://doi.org/10.1006/jmla.2000.2755

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*(1), 46–54. https://doi.org/10.1016/j.bandl.2012.11.007

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. https://doi.org/10.1016/j.cognition.2013.02.002

Ladányi, E., Kovács, Á. M., & Gervain, J. (2020). How 15‐month‐old infants process

morphologically complex forms in an agglutinative language? *Infancy*, *25*(2), 190–204. https://doi.org/10.1111/infa.12324

Lew-Williams, C., & Saffran, J. R. (2012). All words are not created equal: Expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–246. https://doi.org/10.1016/j.cognition.2011.10.007

Marcus, G. F. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77–80. https://doi.org/10.1126/science.283.5398.77

Marquis, A., & Shi, R. (2015). The Beginning of Morphological Learning: Evidence from Verb Morpheme Processing in Preverbal Infants. In R. G. de Almeida & C. Manouilidou (Eds.), *Cognitive Science Perspectives on Verb Representation and Processing* (pp. 281–297). Springer International Publishing. https://doi.org/10.1007/978-3-319-10112-5_13

Mintz, T. H. (2013). The Segmentation of Sub-Lexical Morphemes in English-Learning 15-Month-Olds. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00024

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*(4), 393–424. https://doi.org/10.1207/s15516709cog2604_1

Pelucchi, B., Hay, J., & Saffran, J. (2009). *Statistical Learning in a Natural Language by 8‐Month‐Old Infants*. *Child development*, *80*(3), 674-685.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & Cognition*, *36*(7), 1299–1305. https://doi.org/10.3758/MC.36.7.1299

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behavior; New York*, *6*(6), 855–863.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78

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

Saffran, Jenny R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental Language Learning: Listening (and Learning) Out of the Corner of Your Ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development : The Official Journal of the Society for Language Development*, *13*(4), 357–374. https://doi.org/10.1080/15475441.2016.1263571

Siegelman, N., & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input

facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75. https://doi.org/10.1016/j.jml.2015.07.003

Thiessen, E. D., & Erickson, L. C. (2013). Beyond Word Segmentation: A Two- Process Account of Statistical Learning. *Current Directions in Psychological Science*, *22*(3), 239–243. https://doi.org/10.1177/0963721413476035

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34. https://doi.org/10.1016/j.cognition.2005.01.006

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

6. Conclusions

With these three studies, we attempted to contribute to the literature of cross-linguistic learnability. First, we used input from longitudinal recordings of typologically diverse languages for modeling experiments in Chapters 3 and 4. We also created a language inspired by features of morphologically rich languages for an artificial language experiment in Chapter 5.

Second, we identified specific diversifying features across languages in the morpholexical level e.g. word token frequency. We argued that these features should affect word segmentation. In Chapter 3, we observed that after selection of two languages in theory very diverse in terms of morphological synthesis, Chintang and Japanese, the between-corpus differences of relevant features, such as morpheme to word ratio, were smaller than expected. This is evidence that theoretical complexity of a system does not necessarily manifest in the same way in everyday speech, and even less, probably, when this speech is directed to children. Child-directed speech has long been considered a simplified register (e.g. Genovese et al., 2020), where lexically and syntactically complex elements are less frequent. Thus, differences across linguistic systems may get attenuated within child-directed speech. Future research should address this by quantifying differences between everyday overheard speech versus speech directed to children. We look at this in detail in Part 2.

Moreover, the features we identified did not explain away the effect of language in Chapter 3. This is evidence that in natural language, such factors may be confounded with other linguistic aspects, and thus their importance is less pertinent than what would be found in a carefully controlled experiment testing a specific feature. One other interpretation of the results is that morphological diversity, as expressed by these characteristics, is simply not a factor for segmentation. Future research should consider including more typological features accounting for diversity, such as phonological and syntactic properties, as well as interactions between them.

Third, in the same chapter, we found that differences across models are larger than differences across languages. This is an important finding in itself and invites for further research, as it shows that diversity in segmentability was mainly not due to the differing input, but due to the strategies that were used to process it. Also, the outcome of this study was indicative of some order of performance across strategies: some models performed better than others for the two languages, and seemed to be more affected by the language effect than others (at least by the morphological effect, as was studied here). We decided to investigate this matter further, by testing more languages, which we did in Chapter 4.

In Chapter 4, we identified three specific criteria of evaluation for the performance of our segmentation strategies. First, the models should be cross-linguistically stable. Do they perform relatively well or badly across languages? Second, the models should have a high segmentation performance, which we defined as performing above chance. Third, we included the concept of error plausibility, with respect to over- and under-segmentation. We believe that these criteria are indicative of a viable learning model, and they may be used for evaluation in future modeling studies.

Most models respected the viability criteria, which is supporting evidence for the performance of several segmentation strategies in cross-linguistic, naturalistic input. However, performance for two languages, Russian and Inuktitut, diverged compared to the rest of the languages. This is a matter that should be investigated further. One possible explanation is the existence of a corpus artefact: Perhaps these two corpora have some particularities which cause the failure of the models, regardless of language. Future research could reproduce this study using parallel corpora (a collection of corpora, each of whom is the translation of the other) - this way, we can make sure whether this divergence is due to the challenging structures of these particular languages or not.

Moreover, in Chapters 3 and 4 we made use of corpora whose languages were supposed to be maximally different in specific aspects. We believe that cross-linguistic comparison in future studies should ideally happen under these terms. This way, we can identify the main source(s) of difference between languages. Cross-linguistic comparisons including *any* language, risk to yield underinformative and not easily interpretable results.

Fourth, in Chapters 3 and 5, we addressed the issue of comparability. In Chapter 3, we asked whether difference in segmentability across languages is mitigated when we reframe the concept of word segmentation as segmentation of meaningful units. For this, we conducted a modeling study. Results from the first study showed that language differences are indeed reduced when segmentation is not strictly based on one level. We further argued that considering morphemes as well as words may be beneficial for early language learning, especially for morphologically rich languages.

In Chapter 5, participants got exposed to an artificial language where segmentation could happen across a word and a sub-word level. Preliminary outcomes are evidence that adult participants adapt fast to the particularities of the linguistic structure, and consider both minimal meaningful units and words when listening to an unknown language. Words may thus not be considered as a unique (or

even standard) segmentation level across languages. Future research on younger participants is needed in order to confirm whether this holds true during early language learning.

Finally, the use of ecological input from different languages automatically meant that we would use input also from different cultures and upbringing settings. This aspect of diversity becomes obvious when we compare input from Chintang and Japanese: a considerable part of input heard by Chintang-learning children is either overheard, or addressed to them by other children, and not adults. On the contrary, input heard by Japanese-learning children is mostly directed to them by their own parents. This can be problematic, as input may differ depending on the speech register and the speaker, even within the same language. Thus, *cross-linguistic aspects in acquisition interact with cross-cultural aspects*. We thus decided to address this issue in the next part of the dissertation, Part 2. Are there differences in input from different registers and speakers? How large are these differences cross-culturally? And are they relevant to segmentation?

Part 2

7. Diversity across cultures

What is heard by children in North American or European middle class families is only one way children are talked to. As noted by Ochs & Schieffelin (1984), “the general patterns of white middle-class caregiving that have been described in the psychological literature are characteristic neither of all societies nor of all social groups” (p.283). Given this variation, it is surprising that most research published in language acquisition journals is drawn from one specific sample: Western, Educated, Industrialized, Rich and Democratic (WEIRD) communities (Diesendruck, 2007). Generalization of WEIRD findings to other populations is not obvious (Nielsen et al., 2017).

In this section, we are going to look at cross-cultural diversity. Child input varies enormously within and across cultures (Fouts et al., 2012; Low & Stocker, 2005). The difference in the extent and manner of adults’ talk to children (Hoff, 2006), can be found across many dimensions (e.g. number of siblings, rural vs urban place, literacy, multilingualism, income and socioeconomic status, parental education), as we detail in Chapter 7.1. The dimensions can be so many, that the acronym WEIRD / non-WEIRD may in fact be too abstract to capture cultural diversity in some cases. In 7.2, we turn to evidence from acquisition. These variables set the stage for the investigations in Chapters 8 and 9.

7.1 Input across cultures

7.1.1. Variation in quantitative and qualitative features of language input

In this section, we will look at quantitative and qualitative differences in input speech across cultures. The issue of quantity of speech has received a great deal of attention in recent literature, and will be described in this paragraph. Hoff (2006) considers that the *amount* at which adults engage children in communicative interaction is a cultural variation in input. North American mothers talk to their children since birth (Snow, 1977). In other cultures, such as the Mayans of Mexico (Casillas et al., 2019), caregivers rarely address their children. Yucatec Mayan children hear less speech, only a small amount of which is directly addressed to them (Shneidman & Goldin−Meadow, 2012).

Next, we mention some differences across cultures with respect to *quality* of input. Speech addressed to children is frequently discussed, due to its specific speech features observed in Euro-American families. Caregivers use ‘baby language’, or else child-directed speech (CDS), which seems to be

linguistically adapted for children and prefered by them (Soderstrom, 2007). It has reduced vocabulary, shorter utterances, longer pauses, more repetitions and rephrasings than adult directed speech (Hoff, 2006).

Even though these CDS features have been detected in several languages, such as French, German, Italian, Japanese, British and American English, some appear more extreme in American English than in other languages (Fernald et al., 1989); CDS is subject to cross-cultural variability. For example, in Inuit villages, baby language is not a desirable speech register at all (Crago et al., 1993). CDS from adults to young children is far from universally uniform (Lieven, 1994). This subject will be further discussed in Chapter 8. Cross-cultural diversity has also been identified in the subjects of conversation. For example, EuroAmerican parents seem to provide more information about objects than African American or Japanese parents (Lawrence & Shipley, 1996; Toda et al., 2009).

7.1.2. Tracing the sources of input variation across and within cultures

Overall, cross-cultural diversity is correlated to factors within cultures, such as socioeconomic status (SES) and parental education. These factors will not be dealt with in this dissertation, but we describe them briefly here. Hart & Risley (1995) documented that 2-year-old children growing up in American English high-SES families heard more words than those in low-SES families. SES differences have also been found in vocabulary (more word types and tokens in high SES speech, Hart & Risley, 1995), and syntax (Huttenlocher et al., 2010). Similarly, college-educated mother input is more in quantity, lexically and syntactically richer, and contains more questions than input from high-school educated mothers (Hoff-Ginsberg, 1991).

The number of siblings is another factor. Firstborn children seem to receive more speech from their mothers than later borns (Hoff-Ginsberg, 1998; Oshima-Takane & Robbins, 2003; Snow, 1972), and more complex speech than later borns (Hoff-Ginsberg, 1998). However, Woollett (1986) notes that language environments for a younger sibling should be stimulating for learning; speech between older siblings and the mother is very interactive, frequently referring to events, objects, or the younger sibling. Thus, such environments could provide developmentally complex and salient models of language.

Last, the above mentioned aspects of cross-cultural diversity relate to the issue of dyadic studies. Previous literature has focused on mother-child interactions (but see Weisner et al., 1977; Woollett, 1986), despite the fact that most children around the world grow up in polyadic situations, and nonparental caretaking is common across societies (Lieven, 1994). Caregiving in diverse cultures

relies heavily on extended family (e.g. Dilworth-Anderson, 1992), thus linguistic input may not necessarily come from the mother. In Indigenous Australian communities, for example, older children look after younger children. In Arnhem Land, children enter a peer group since they are two years old (Hamilton, 1981). Connelly (1984) observed that, in some Lesotho villages, siblings and peers as young as 2;1 years old have caregiving roles, speaking to the younger child with simplified speech (also see Demuth, 1992). This input is likely to differ from parental input (Loakes et al., 2013), but there are hardly any studies on its nature. We further address this issue in Chapter 8.

7.2 Language acquisition across cultures

7.2.1 Analysing previous learning outcomes

We still know very little about what children hear in diverse cultures, and evidence on learning outcomes in acquisition across cultures is also limited. Of these studies, even fewer studies have looked at differences in segmentation outcomes, since previous research has focused on differences in lexical and syntactic development through production. For example, one thing that varies as a function of culture seems to be the amount and content of children's productive vocabulary (Hoff, 2006; Tardif et al., 1999).

Moreover, most cross-cultural work derives from ethnographic studies and is anthropologically-oriented, with little systematic research on language learning. Some language information provided may sometimes be qualitative observations of the investigators.

One exception is a recent longitudinal paper that studied language learning of Tseltal Mayan children (Casillas et al., 2019). These children are infrequently addressed and hear little adult CDS. Even though the authors expected a divergence in lexical development, compared to middle-class English norms, Tseltal children learned language early on and produced their first words at the same age as English children. Obviously, these children extract enough information from the linguistic environment, even though they are not directly addressed to. Similar results were reported by Shneidman & Goldin−Meadow (2012). Interestingly, these authors found that children's early words were predicted by adult CDS (and not overheard speech or speech by other children), even though they heard very little of it (compared to input of middle-class Euro-American children).

What previous ethnographic work exists, shows similar patterns of language acquisition; children learn language early on despite hearing little CDS (Brown & Gaskins, 2014; Liszkowski et al., 2012;

Ochs & Schieffelin, 1984). Crago et al. (1997) argued that Inuit children, who receive little directed input, acquire Inuktitut at ages comparable to middle class North American children. Lieven (1994) also observed that across-cultures children tend to learn language at around the same time.

A way that children may extract information is from non-directed/overheard speech, by listening to nearby speech addressed to other people. Children seem to be good at observing and learning from the interactions and behaviors taking place around them (León, 2011; Rogoff, 2003). According to behavioral experiments, two-year olds learn referents presented in a third party conversation (Akhtar et al., 2001), even if they are simultaneously engaged in another activity, or if the label is in a non salient position in the sentence (Akhtar, 2005). Dunn & Shatz (1989) provided naturalistic evidence that two-year-old children understand much of the conversations they overhear. According to Barton & Tomasello (1991), 19-month-old English-speaking children are capable of participating in mother-sibling-child conversations. Children were as likely to respond to comments directed to another person as they were to those directed to themselves. Moreover, in artificial language studies, children also implicitly acquire at least some complex structures (Finn et al., 2014).

7.2.2 Linking input to learning outcomes

We discuss here some variables of input speech previously mentioned to account for lexical and morphosyntactic development, and which may differ cross-culturally. The overall quantity of CDS is one of them; more input seems to lead to faster vocabulary growth for English-learning children (Diesendruck, 2007; Hart & Risley, 1995; Weizman & Snow, 2001), accounting for 16% of the variance in 2-year-old children's utterance length growth over the following 9 months (Barnes et al., 1983).

CDS is supposed to facilitate early word learning in English-speaking (Bakermans-Kranenburg et al., 2004; Cartmill et al., 2013; Rowe, 2008) and Spanish-speaking families (Weisleder & Fernald, 2013), by providing a simple-to-learn language model. Specifically, utterance repetitions are positive predictors of English-learning children's grammatical development, predicting 18-40% of variance (Hoff-Ginsberg, 1986), and frequency and diversity of verb frames in input predict child verb use (Naigles & Hoff-Ginsberg, 1998). The number of word types produced by mothers predict the number of word types of their two-year-olds ten weeks later (Hoff & Naigles, 2002; Hoff-Ginsberg, 1986). Long utterances and questions in CDS contribute to syntactic development (Choi & Gopnik, 1995; Hoff-Ginsberg, 1998; Huttenlocher et al., 2010). We hypothesize that some of these or other features of CDS may also predict segmentation, being one of the first tasks children need to tackle.

Nevertheless, CDS also has words, sounds and sentences which are rather complex (CVC syllabic forms, sound sequences, …)(Gierut, 2007). Fernald & McRoberts (1996) documented that boundary markers in English CDS are not reliable enough for bootstrapping, as utterances often have non-canonical structures. Ludusan et al. (2015) showed a lower recall and cluster collocation in English CDS than adult-directed speech (ADS), even though it had better prosodic boundary information than ADS. These outcomes are inconsistent with the view that IDS is clearer and simpler than ADS, at least as far as segmentation is concerned.

With respect to modeling segmentation, some work has addressed the issue of learnability in diverse registers. Ludusan et al. (2017), based on Japanese CDS and ADS laboratory-collected data, found a smaller difference in segmentability between these two registers than previously reported (Batchelder, 2002). A smaller and even reversed difference in segmentability was found for ecological CDS and ADS of English-speaking families (Cristia et al., 2019).

Future studies need to investigate the features of CDS, and quantify their exact impact on learnability. However, it is important to do so considering that, cross-culturally, CDS features vary; CDS may be produced by different speakers, and consist of very different qualitative features.

7.2.3 Comparing learning outcomes across cultures

Cross-cultural outcomes should be evaluated with caution. An ethnocentric bias hindering comparability has been brought up in the field, as measures of affection are sometimes adjusted to Western ways of thinking (Rogoff, 2003; Rothbaum et al., 2000). For example, Quiché CDS may not have some features frequently found in English CDS, but it has eight different special features, including whispering, initial-syllable deletion, a verbal suffix appearing only in CDS, and fixed word order (Pye, 1986).

The use of standardized tests (e.g. reading and vocabulary tests) has also raised issues of comparability. For example, children speaking African American English (AAE) receive the same language tests as other American English-learning children, but their free play consists mostly of code-switch, complex syntactic forms and a special use of semantics (Craig & Washington, 2004). Interestingly, some children shift to using fewer AAE features once at school, and these children outperform their peers who do not shift to AAE on standardized tests.

One other issue concerns object-oriented and high-density CDS activities such as book reading, which are rare in some communities. It can be challenging to compare speech to children across different activities, and in different, culturally-appropriate routines. Woollett (1986) reported that mother to child interactions are largely dependent on context. Goldfield (1993) documented that toy play incites the production of more nouns than verbs, whereas the opposite happens during non-toy play (physical play). Similar patterns have been reported for English and Mandarin Chinese (Tardif et al., 1999), Korean (Choi & Gopnik, 1995) and Japanese-learning children (Ogura et al., 2006). The acquired vocabulary thus depends on caregiver-child *interaction norms*. Whereas for English-learning children, much time spent with caregivers consists in naming objects, and many words acquired are basic-level object nouns, this may not be the case across communities.

Last, researchers typically record one or a few hours of a child's input. However, the input of a child may differ depending on the time of the day (Casillas et al., 2019; Soderstrom & Wittebolle, 2013; VanDam et al., 2016). Recordings at specific times reflect temporary conditions, such as a particular conversation during the session (Huttenlocher et al., 2010). Few sessions cannot fully capture the diversity of input and production, sometimes even due to corpus artefacts and contexts.

7.3 Future research

Montag et al. (2018) emphasizes that there is a lot we don't know about language learning, and that "we need to understand all this if we are to tell parents how they should talk to their children" (p.22) - also see Leffel & Suskind, (2013) and Roberts & Kaiser (2011). Casillas et al. (2019) concludes that more quantitative and reproducible methods in diverse contexts are needed, in order to learn more about language learning. One way to investigate different contexts is by testing scaled-up, diverse input. Quantitative metrics may include corpus analysis, a useful means of quantifying diversity in input, and modeling. **In Chapters 8 and 9, we address these by implementing segmentation modeling and corpus analyses in diverse, naturalistic settings.**

Due to the WEIRD bias and other methodological issues of previous studies discussed above, it is still unclear what input is heard across cultures. Previous evidence, though, suggests that there is no standard way that children are addressed to. In particular, the amount of directed input varies and the role of CDS is not obvious when we look at language learning at scale (Gierut, 2007). In different environments, CDS is rare, or it does not have facilitating features. However, children growing up in these environments somehow learn language. Moreover, there is still little information on other speech registers often present in children's ambient input, such as overheard speech. **In Chapters 8 and 9, we address this by comparing child-directed and child-overheard speceh, as they were**

heard by children in a non-WEIRD and a WEIRD culture. We attempt a comparison of specific features, and ask whether they can explain away segmentability differences in Chapter 9.

Finally, acquisition does not always happen through a mother-child dyadic interaction, and the role of different speakers should also be taken into consideration, especially in cross-cultural studies. Living in households with more people than a typical North American family might suggest that children could get enough language input, even if it does not come from the direct family. For example, other children and adults often have prominent caregiving roles. **We address this by comparing the contribution of other speakers' child-directed and overheard input in a non-WEIRD and a WEIRD culture, in Chapter 8.**

8. Child-directed and overheard input from different speakers in two maximally distinct cultures *

Abstract: Mother-child dyad speech has long been the focus of early input studies, despite evidence suggesting that non-maternal input can be important for language outcomes. Additionally, in many communities (particularly non-WEIRD ones), interaction occurs with multiple speakers rather than mostly the mother. Yet, few studies describe CDS from various speakers, and even fewer investigate this across cultures.

In this study, we analyze speech produced around and to children by their mother, other children and adults, in two diverse cultures. We ask who produces the input, and how much of it is child-directed. We also ask whether different speakers vary in terms of utterance length, function (ratio of questions) and lexical diversity. To answer these questions, we annotated three corpora. The non-WEIRD Demuth corpus in the Sesotho language was recorded in non-industrial South African Lesotho. We also annotated recordings from the WEIRD Lyon and Paris corpora for three children- those with siblings, and the same age range as in the Demuth corpus.

CDS is significantly prevalent over overheard speech for both settings. However, the input composition is dramatically different; maternal input is more dominant in the WEIRD corpora compared to the non-WEIRD one. In the latter, other children's input is more prevalent than maternal input. Interestingly, in terms of speech quality, other children's and adults' CDS present similarities with maternal speech within each culture. These results invite further cross-cultural early input research, in order to check if these speech compositions and qualities are representative of WEIRD and non-WEIRD quantifiable distinctions, and the impact these might have for language development.

***Loukatou, G., Scaff, C., Demuth, K., Cristia, A. & Havron, N. Child-directed and overheard input from different speakers in two maximally distinct cultures. (under review)**

Child-directed and overheard input from different speakers in two maximally distinct cultures

The amount and quality of early language input are factors affecting children's language development, and have drawn substantial attention in research (e.g. Hart & Risley, 1995; Hoff & Naigles, 2002). For example, some research shows that children's vocabulary skills correlate with the amount and quality of input speech that mothers offer children during day-to-day interactions (Hoff, 2003; Hoff & Naigles, 2002; Hart & Risley, 1995). However, previous literature has mostly focused on *maternal* input. Relatively little attention has been devoted to input from other children and adults. In this paper, we will describe input from corpora where other speakers, in addition to mothers, talk to children.

Most previous studies on language input are based on families living in middle-class Euro-American communities. Indeed, Henrich et al. (2010) observed that most participants in psychological studies come from a Western, Educated, Industrialized, Rich and Democratic (WEIRD) population sample, and this bias is also obvious in developmental studies (Nielsen et al., 2017). Recent evidence points to the fact that input differs depending on the culture of the family (e.g. Cristia et al., 2019). What limited literature exists supports the idea that different populations also have different norms about who is expected to speak to children.

Sources of input

In most middle-class Euro-American families, parents are expected to have absolute responsibility over their children. Consequently, most studies focus on parental input, rather

than input from non-parental speakers (e.g. Bakermans-Kranenburg et al., 2004; Huttenlocher et al., 2010; Ispa et al., 2004; Pan et al., 2005). Moreover, in these families, the mother typically has the role of primary caregiver (e.g. Roopnarine et al., 2005). As a result, mother-child dyad speech has long been emphasized in early input studies.

However, the focus on the mother as the primary caregiver might not reflect universal human tendencies. There is evidence suggesting that mothers are not always the sole, or even principal caretakers (e.g. Shneidman & Goldin−Meadow, 2012; Weisner et al., 1977). Caretaking of a child by an individual who is not the mother is referred to as allomaternal care. Responsibility for care can be shared among a circle of individuals, kin and non-kin, older siblings, peers or cousins, as part of common daily routine around the world (Fouts et al., 2012). Even in middle-class Euro-American families, siblings might play some caretaking role. For instance, in an experimental setting with 57 American mothers and their preschool children, Stewart & Marvin (1984) found that when the mother left the room, 51% of older siblings engaged in caretaking activities. Sibling caretaking has also been documented in blue collar African-American and Latino families, where most play also happens among siblings (Zukow-Goldring, 2002). Lower income Euro-American families also seem to rely more than middle income families on extended kin for child care (Hofferth, 1995).

Across diverse cultures, the need to study more speakers than the mother is even more evident. Using an ethnographically-detailed sample of 186 societies, Barry & Paxson (1971) found that only 46.2% of the societies had mothers as principal caretakers. After infancy, this proportion decreased by another 19.4% (see also Weisner et al., 1977). During infancy, adult family members are the principal companions or caretakers in 39.8% of the societies (32.3%

are mothers or other females), children rank second (16.7% females, 24.8% overall) and other females, including employees, third (9.1%).

More generally, anthropological evidence finds that siblings are frequent allomaternal carers across cultures, for example in Ngoni of Malawi, (Read, 1968); Dusun of Malaysia, (Williams, 1971); Java, (Geertz, 1989); Kwoma of New Guinea (Whiting, 1941); Polynesia (Martini & Kirkpatrick, 1992). In Hawaiian families, caretaking is shared among parents, neighbors, kin, and almost always children (Gallimore et al., 1974). These results also relate to household size, as larger household size means more opportunities for allomothering. For example, the total number of siblings and incidence of sibling caretaking seem to correlate in Hawaiian-American families (Gallimore et al., 1974). Families tend to be larger cross-culturally than is the case for the United States (Burch, 1970, 1979).

Even when children in the environment are not allomaternal carers, they may play with the younger child, and therefore have an opportunity to provide linguistic input. Play in children groups of the same or mixed age is another way to expose oneself to language. Children growing up in middle-class Euro-American families interact in groups of other children on many occasions, such as in preschool and kindergarten (even though these interactions are usually monitored by adults). According to O'Shannessy (2013), in many Australian-Indigenous communities, children spend a great deal of time interacting with other children. In Arnhem Land, children are absorbed into peer groups by the age of two years (Hamilton, 1981). In Guatemala, children at the age of two years start spending their time with other children and seldom look for adult attention (Rogoff, 1981).

Other adults should also be considered as a potentially important source of input. For example, according to the Census 2000 (Simmons & Dye, 2003), 5.8 million grandparents in the United States were either primary caregivers raising grandchildren, or living as coresidents and helping to care for grandchildren (see also Standing et al., 2007). Caregiving by other adults, such as grandmothers, is common in many cultures, including many African communities (e.g. Thupayagale−Tshweneagae, 2008).

Despite these facts, few studies describe input to children produced by other family members such as siblings (but see Hoff-Ginsberg & Krueger, 1991; Weppelman et al., 2003), and other adults (but see Shute & Wheldall, 1999, 2001). Our study will take a step into filling this gap by analyzing the amount and quality of input speech children receive in diverse communities by mothers, other children, and adults.

Quantity of input

Across cultures, the total amount of input directed to children differs. Children may be addressed by their caregivers only rarely, sometimes because they are not seen as communicative partners (Lieven, 1994). This has been noted, for example, in Gusii mothers in Kenya (Richman et al., 1992), Gapuners in Papua New Guinea (Kulick, 1992), Kaluli in Papua New Guinea (Ochs & Schieffelin, 1994), Samoans in Western Samoa (Ochs & Schieffelin, 1994) and Javanese speakers in East Java (Wolff & Poedjosoedarmo, 1984). This difference might be especially relevant when studying the role of different speakers, because, as mentioned above, cultures vary in who spends time with the child.

A study based on daylong recordings found that Tseltal Mayan children are only infrequently directly spoken to by adults: a day-wide average of 3.63 min per hr (Casillas et al., 2019), which is approximately a third of what found for North American children (11.36 min per hr, Bergelson et al., 2019), but is comparable to that for Tsimane children (Cristia et al., 2019) and Yucatec Mayan children (Shneidman & Goldin−Meadow, 2012). Meanwhile, Tseltal children hear a lot of other-directed speech (ODS), averaging 21.05 min per hr, which is more than has been previously reported for other cultural settings (e.g., Bergelson et al., 2019).

Child-directed speech is not an exclusive source of input, and these studies do not rule out that other registers and speech from other speakers might also contribute to the child's input (Soderstrom, 2007). Even though they are not directly addressed by adults, Tseltal children somehow extract enough information to produce canonical babbling, first words, and word combinations at approximately the same ages that North American English-learning children do (Casillas et al., 2019). Some researchers argue that at least some amount of early language learning must be based on overheard, rather than child-directed speech (e.g. Ochs & Schieffelin, 1994). Lieven (1994) suggests that the child-centered style of speaking is one way of enabling children to learn language, but it is not essential, and concludes that children around the globe tend to learn language at approximately the same time, despite the many diverse ways of speaking to (and around) them.

Indeed, children are good at observing and learning from interactions taking place around them (e.g. Rogoff, 2003 - see also behavioral experiments from Akhtar, 2005 and naturalistic evidence from Barton & Tomasello, 1991 and Dunn & Shatz, 1989). We should also consider the fact that ODS is not necessarily speech directed to adults, but it can also be directed to

other children, especially in communities where children spend a lot of time together. Speech directed to other children, even though overheard, may be easier to follow and of higher relevance to the child than speech directed to adults.

Whereas in US families, siblings seem to address their younger siblings much less than their mothers do (Oshima-Takane & Robbins, 2003), a study on the Tsimane forager-horticulturalist society in Bolivia found that while 42% of child-directed speech came from the mother, 37% came from other children, and the rest from other adults (Scaff et al., in prep.). According to Shneidman & Goldin−Meadow (2012), the amount of utterances spoken by other children was higher in a Yucatec Mayan village than in Chicago (68% vs. 10% at 24 months), and unlike in Chicago, a sibling talked to the target child more than an adult would. Yakanarra children interlocutors also address markedly more input to younger interlocutors than older interactants do (Loakes et al., 2013). Mothers were *not* the major source of language input for Luo, Koya and Samoan children (Snow & Ferguson, 1979).

Quality of input

Quantity is not the only aspect of input that matters. The *quality* of input (here, its morphosyntactic, lexical and interactional properties) is important. The input addressed to a child is called child-directed speech (CDS). CDS is preferred by children to adult-directed speech (ADS) (The ManyBabies Consortium et al., 2020), it seems to promote language learning (Soderstrom, 2007) and can differ in quality from ADS along many aspects.

Some of the earliest work on CDS reported that it is a register adjusted to child listeners, exhibiting syntactic, phonological and lexical simplification. Specifically, maternal CDS is

characterized by a large number of questions (e.g. Kruper & Užgiris, 1987; Toda et al., 1990), repetitions (e.g. Hoff, 2006), shorter utterances and low type-token ratio (e.g. Henning et al., 2005). CDS may also be helpful for word learning, due to a preponderance of single word utterances (e.g. Brent & Siskind, 2001).

Although American English is the most studied language with respect to CDS, there is evidence for some properties of CDS in a variety of languages, including French (Loukatou et al., 2019; Veneziano & Parisse, 2010), German (Fernald & Simon, 1984), Japanese (Fernald & Morikawa, 1993), Spanish (Weisleder & Waxman, 2010), Hebrew (Adi-Bensaid et al., 2015), Turkish and Mandarin Chinese (Shi et al., 1998), British English (Shute & Wheldall, 1999), and Australian English (Lee et al., 2014). CDS word tokens and types are a better predictor of vocabulary acquisition than overall heard tokens and types (Brent & Siskind, 2001; Shneidman & Goldin−Meadow, 2012).

Cross cultural differences are found not only for the overall amount of speech around the child, the proportion of CDS across cultures, and the speakers who speak around the child, but also in the *way* speakers address children. Some distinguishing features of CDS can be found across cultures. For example, Ngaanyatjarra children in Indigenous Australia are addressed with phonologically simplified CDS, with repetitions and slower speech rate (e.g. Kral & Ellis, 2008). Harkness (1977) observed that Kipsigis caregivers in Kenya adjust the length and complexity of their utterances to their children's length of utterance. Fisher & Tokura (1996) reported that Japanese- and English-speaking mothers add prosodic cues at utterance edges, and alter phonetic cues relevant to their language.

However, speech addressed to children is not always simplified. For example, contrary to studies in middle-class Euro-American communities, Javanese children are spoken to with complex honorific forms that they are supposed to use in order to talk to their superiors, and with longer and morphologically more complex utterances than used in ADS (Smith-Hefner, 1988). Similarly, Kaluli and Samoan caregivers do not engage in morphosyntactic simplification (Ochs & Schieffelin, 1994), and little language input in Inuktitut-speaking communities is morphosyntactically simplified (Allen & Crago, 1997). Quiche´ Mayan CDS does not have some of the properties noted in American English CDS; but Quiche´-Mayan mothers use other features not found in English CDS, such as whispering, initial-syllable deletion and a suffix only found in CDS (Pye, 1986). Thus, there seems to be cross-cultural variation of features within CDS. It is noted that, despite less simplification in CDS, there is some documentation that Kaluli, Inuktitut and Samoan children learn language within the range of normal developmental variation (Allen & Crago, 1997; Ochs & Schieffelin, 1994).

While some cultures have been mapped with respect to their distinguishing qualities of CDS, less is known about the way different *speakers* use CDS. That is, do other children and adults also engage in such qualitative modifications of speech? Are the characteristics of CDS similar for non-maternal speakers across societies? For example, very few studies have investigated CDS produced by other children. With respect to siblings in US families, Dunn & Kendrick (1982) and Shatz & Gelman (1973) reported that siblings adjusted their speech when talking to their younger siblings. They used short sentences, simple verb tenses, and repeated their sentences twice as much as when they were talking to their mothers. In Lesotho, Connelly (1984) observed that even two-year-old Sesotho-learning children speaking to younger children adjust their speech, though Demuth (1986) reported that

Sesotho-speaking siblings learn complex linguistic forms, such as relative clauses (e.g., “Bring that thing you found”), more from siblings than from adults. Furthermore, in a task-oriented study where seven-to-eight-year old siblings from US families played with their toddler siblings using toys, they were found to be less adept than their mothers in adjusting speech to their younger siblings (Hoff-Ginsberg & Krueger, 1991). Other studies have confirmed this, showing that preschool children correct their siblings’ syntax less often than their parents do, ask fewer questions, and provide fewer corrective repetitions (Dunn & Kendrick, 1982; Mannle et al., 1991; Strapp, 1999).

Like mothers, fathers and grandmothers seem to adopt a simplified speech register towards children in French, Italian, German, Japanese, British and American English (Fernald et al., 1989). Both similarities and differences in speech have been observed between mother and other adult speakers in American English and French (Pancsofar & Vernon-Feagans, 2006; Rondal, 1980). In Fernald et al. (1989), both fathers and mothers used shorter utterances when speaking to their children than when speaking to an adult. Fathers' speech at 24 months was predictive of children's language development at 36 months in American English (Tamis−LeMonda, 2004). However, in a study by Rondal (1980), French fathers’ speech was more lexically diverse and contained longer utterances than mother’s speech, and McLaughlin et al. (1983) found that mothers tune their language (American English) more to the child's linguistic abilities than fathers do.

The present study

As this brief summary hopefully illustrates, there are still large gaps in our knowledge about the nature of language input across cultures. There is strong evidence that CDS from mothers

is crucial to acquisition. At the same time, in Euro-American families, and especially in less-studied diverse cultures, linguistic input may come from other people, very often children. It is still unknown what properties characterize this kind of input, and whether it is helpful for acquisition. Research is thus needed to provide answers about whether the child is even addressed in these contexts (or is only exposed to surrounding speech), who the speakers are, and how similar their speech is to maternal speech. This should be investigated both for cultures where the mother is considered the primary caregiver, and for cultures where children spend most of their time with other people. Our work will thus focus on the distribution and features of child directed and other directed speech produced by different speakers in naturalistic, cross-cultural recordings.

In this study, we take a step towards describing children's input across cultures. We describe children's input in middle-class European families in two cities in France, and that of a non-industrial southern African community in Lesotho, where it has been documented that children interact with more speakers than just the parents. We use the same descriptive metrics in order to comprehensively document the linguistic input children receive in both kinds of communities, from parents and other people. We detail the amount and quality (here, the morphosyntactic, lexical and interactional features of speech) of linguistic input spoken by different people, separately for child-directed speech, and speech directed to others.

We specifically ask, for both communities:

1. How do different speakers contribute to the linguistic input heard by the child?

2. How much input heard by children is directed to them, and how much is directed to other children and adults, when different speakers are taken into account?

3.What are the qualitative properties of this input, and how do they differ across speakers and cultures?

According to Demuth (1986), rural families in Lesotho often lived in extended family units, with allomaternal carers participating in the caregiving of the child. Many Sesotho-speaking men were employed in South Africa and were rarely at home. Siblings typically have a two-and-a-half years age difference, and children as young as 2;1 often spend time in peer groups and with younger siblings (Demuth, 1992).

Demuth recorded spontaneous interactions of children in rural Lesotho (the Demuth Sesotho Corpus). According to her observations (Demuth, 1986, 1992), teaching language to the child was an important responsibility of the community, and even children seemed to adjust their speech when talking to younger children, modifying its phonology and syntax. The recorded corpus was used by the investigator in order to study grammatical phenomena, focusing on the target children's production and the use of prompts by caregivers.

In this study, we make use of the same corpus, together with two French corpora (the Paris and Lyon corpus), in order to investigate our research questions. We annotated these corpora in order to quantify the contribution of different speakers to the overall input, and to analyse the quantity of directed and overheard registers, and specific qualitative speech properties for each register and speaker. Sessions with similar recording methods and target children with older sibling(s) and the same age range were chosen for both corpora, and were annotated by native French and proficient English speakers.

Method

Data

Sesotho corpus

The Demuth Sesotho Corpus was recorded in a rural community in southern Africa. It was compiled by Katherine Demuth in the country of Lesotho in 1980-1982 (Demuth, 1992). Data were collected in a Lesotho mountain village of 550 people in the district of Mokhotlong. The language spoken is Sesotho (also called Sotho, or Southern Sotho), a southern Bantu language used by three million speakers. The corpus can be found on CHILDES (MacWhinney, 2000).

The Sesotho corpus is a longitudinal study of three target children as they interacted with their caregivers. The children were aged from 2;1 to 3;2 for Hlobohang (boy, 11,221 input utterances after excluding target child and investigator speech), 2;1-3;0 for Litlhare (girl, 12,669 input utterances after excluding target child and investigator speech) and 2;1-3;2, for ‘Neuoe (girl, 10,502 input utterances after excluding target child and investigator speech). Three-to-four- hour recordings of spontaneous speech took place every month for each child with the presence of the investigator. Hlobohang and ‘Neuoe each had an older cousin in the same household, and Lithlare had an older brother. The transcriptions were morphologically coded and translated to English with the help of the children’s mothers and grandmothers.

French corpus

In order to obtain data comparable to the Sesotho corpus in size, number of target children, target child age, and presence of siblings, we merged recordings from two French corpora, the Paris corpus (Morgenstern & Parisse, 2012) and the Lyon corpus (Demuth & Tremblay, 2008). Both corpora are on CHILDES (MacWhinney, 2000). The need for two corpora is due to the fact that most target children in these corpora were first-born and were thus excluded for the present purposes, which aimed to look at input in contexts where there were more people than just the mother. Families with more than one child are not considered atypical in France; in 2013, 6 out of 10 families had more than one child (INSEE, 2013). The French corpora were recorded with a similar method to the Sesotho corpus, as described below.

The Lyon corpus was co-created by Katherine Demuth and Harriet Jisa in order to study the acquisition of morphophonological elements in French. It was compiled by Jisa and colleagues at the University of Lyon 2. The corpus contains longitudinal audio recordings of monolingual, typically developing French-speaking children from one to three years of age, and focuses on spontaneous interactions. Recordings from Anaïs (girl, 7,022 utterances after excluding target child speech and investigator speech) and Theotime (boy, 4,062 utterances after excluding target child speech and investigator speech) were included in this study, to obtain an age range comparable to the Sesotho corpus. Each child was recorded for one hour every two weeks. Anaïs had two older sisters, and Theotime one older sister. Although the research assistant was not always present during recordings, we only kept the recordings where the assistant was present, in order to increase comparability with the Sesotho corpus. The sessions were recorded with a small video recorder placed on a tripod. The child wore a

wireless microphone, and its radio transmitter was placed inside a child pack worn by the child. The sessions were transcribed at the Dynamique du Langage Laboratory.

The Paris corpus was co-created by Aliyah Morgenstern and Chistophe Parisse. The corpus was used to study the presence of pragmatic cues in speech, such as prosody and gestures, and grammatical development. The corpus contains longitudinal audio recordings of monolingual, typically developing French-speaking children. One child in our study comes from this corpus, Anaé (girl, 7,247 utterances after excluding target child speech and investigator speech). The child was videotaped in her home once a month for an hour, in spontaneous interactions, and the data were then transcribed by the investigators. The chosen recordings also span the Sesotho age range. Anaé has two older brothers.

Data Preparation

For both the Sesotho and the French corpus, we quantified the contribution of different speakers to the overall input by measuring the total number of utterances produced by each speaker for the whole corpus, as well as the average number of utterances per recording hour and per session. Speech produced by or addressed to the investigator during the sessions, though generally rare, was excluded from the analysis. We take into account local norms regarding family units. In the Sesotho corpus studied in the current study, for example, one target child (‘Neuoe) was growing up in the same household with her cousin, and was being taken care of by her aunt, whom we consider as a mother for the sake of this study.

The target children were selected by the criteria described above (having older ‘sibling(s)’, same age range, and sessions with similar recording methods) and annotated for speaker and

interlocutor by a native French (for the French corpus) and a proficient English speaker (for the Sesotho corpus, using the English translation). Although the corpora had already been transcribed, indicating who was speaking and what they were saying, the transcriptions did not contain information on who was being spoken to. Therefore, we added this layer of annotation. Our annotators were asked to read the transcriptions of a session, and try to understand who is being spoken to for each utterance. When available, they were provided with descriptions of the situation and comments of the authors and previous annotators. The annotators were instructed to add an addressee annotation only if they were confident (more than 70% sure), indicating whether a sentence is addressed to a target child, or another specific addressee or group of addresses. Specifically for the French corpus, some parts of the recordings had not been transcribed by the original transcribers (usually parts where the target child did not participate in the conversation). Therefore, the annotators also transcribed the missing sections after consulting the videos.

To check accuracy of the addressee annotation, 20% of these annotations were double-checked by a second annotator. For Sesotho, the two annotators agreed in 89% of the cases. Where they did not agree, this often was because one coder made a decision and the other was under 70% sure (57% of disagreements). A further reliability check was then performed by Demuth, who was present during the original data collection, and any questions resolved. For French, the two annotators agreed in 96% of the cases. Since the reliability check between the first annotator and the second annotator (the one annotating 20% of the corpus) were high, we took into account the annotations of only the first annotator, excluding utterances where this annotator was uncertain.

Last, we studied the quality of speech input to children, using a corpus analysis. All scripts related to corpus processing and analysis can be found in this OSF link: https://osf.io/mws9g/?view_only=b116b0c6bb5c48508a547dbc955461b1. We measured *lexical diversity* by counting the moving average type-token ratio (MATTR) and the ratio of hapaxes (words found only once). MATTR gives the mean total number of unique words (types) divided by the total number of words (tokens) per chunk of 10 (and 100) words, thus controlling for differences in corpus size. Ratio of hapaxes gives the ratio of hapax words in the corpus, by dividing the number of hapaxes with the total number of word tokens.

We measured *morphosyntactic complexity*, by counting the preponderance of single word utterances, by dividing the number of utterances containing only one word with the total number of utterances, and by counting the mean utterance length (MLU) in words (and in morphemes for Sesotho). It should be noted that, unlike French, Sesotho contains multi-morphemic prosodic words, since prefixes to the verbs (subject and object person-number-agreement forms and tense markers) and noun-class prefixes are written together as one 'word'. The average number of morphemes per word in the corpus is approximately 1.9. We deal with this difference in three ways: (1) We tried to make our French corpus 'more similar' to the Sesotho, by combining determiners with the following nouns as one prosodic word (details on Demuth & Tremblay, 2008). Nonetheless, the multimorphemic Sesotho 'verb' *ke-tla-mo-otla* "I-will-him-hit" was counted as one 'word', whereas the same phrase was counted as four words in French ("Je vais le frapper"). (2) We provide a supplementary metric for Sesotho and French counting the mean utterance length in morphemes (see Appendix). We do this because MLU might not be as comprehensive a matrix to language complexity in Sesotho as it is in French, and mean utterance length in

morphemes might be a better measure. (3) Since the two languages differ across these aspects, we only focus on language-internal comparisons - we never compare which language uses longer utterances of more morphemes per word, but we do compare between speakers and registers within each language.

Last, we measured *speech elicitation* by looking at the ratio of questions divided by the total number of utterances, and by looking at the ratio of conversational turns. We considered as a conversational turn each utterance followed by an utterance produced by the person initially addressed to. We then divided the number of these utterances by the total number of utterances. The results of the above metrics were descriptively compared across different speakers and speech registers.

Results

We looked at input composition with respect to its speakers, after categorizing the input as child-directed and overheard, and further looking at the addressee of overheard input (other children or adults).

Quantitative features of input

We asked how much different caregivers contribute to the overall linguistic speech heard by the child, both directed and overheard (see Table 1). For the Sesotho-learning children, most speech came from other children. The category “other children” contains all other children present in the linguistic environment of the key children: siblings, cousins, playmates, etc. Speech from other adults varied; for one Sesotho-learning child in particular, Hlobohang, other adult speech (the grandmother) was more abundant than mother’s speech.

French-learning children received most of their input from mothers, and other children had the smallest contribution to overall input. There was variation in input from other adults.

Child	Mother %	Adults %	Children %
Sesotho Hlobohang	14.67	40.29	**45.03**
Sesotho Litlhare	41.68	2.81	**55.51**
Sesotho ‘Neuoe	22.10	11.82	**66.08**
French Anaé	**83.47**	1.0	15.53
French Anais	**67.80**	28.34	3.86
French Theotime	**93.47**	-	6.53

Table 1. Percentage of the total number of utterances produced by mothers, other adults and children for each child. The largest number in each row is in bold.

Second, we asked how much input heard by children was directed to them, how much was overheard, and how child-directed versus overheard speech was distributed across speakers. CDS is speech directed to the child. Overheard speech includes speech directed to other children (OCDS), and to adults (ADS). The majority of input was child-directed for both French-learning and Sesotho-learning children, except for ‘Neuoe, who was the cousin in the family with other children, and most speech around her was OCDS. More details can be found in Table 2.

Child	CDS %	OCDS %	ADS %
Sesotho Hlobohang	**89.83**	6.81	3.33
Sesotho Litlhare	**90.95**	5.36	3.28
Sesotho ‘Neuoe	27.0	**50.29**	22.04
French Anaé	**87.04**	6.28	5.55
French Anais	**94.28**	2.44	3.03
French Theotime	**91.80**	4.06	3.50

Table 2. Percentage of the total number of utterances that was CDS, OCDS or ADS for each child. The largest number in each row is in bold. These three do not add up to 100% because the rest of the utterances were unclassified in terms of addressee.

As can be seen in Table 3, CDS was mostly produced by other children for the Sesotho-learning children. For the French-learning children, the mother was the main source of CDS. Information on OCDS and ADS speakers can be found in the Appendix, Table A1.

	CDS		
	MOT	ADU	OCHI
Sesotho Hlobohang	13.74	37.08	**39.01**
Sesotho Litlhare	40.25	2.06	**48.64**
Sesotho ‘Neuoe	4.89	5.49	**16.61**

French Anaé	**79.88**	.15	7.01
French Anais	**65.37**	27.40	1.51
French Theotime	**90.37**	-	1.43

Table 3. Percentage of speech produced by mothers (MOT), other adults (ADU) and other children (OCHI) comprising CDS for each child. The numbers for 'Neuoe are small because most speech around the child was not CDS. The largest number in each register-based group of cells is in bold.

Qualitative features of input

Next, we asked what the qualitative properties of input are, and how they differ across speakers and cultures. We focused on morphosyntactic, lexical and interactional properties separately. The raw data can be found in the online supplementary material. Morphosyntactic properties are shown in Figure 1. The MLU measured in morphemes is in the Appendix (Figure A1). For both Sesotho and French, CDS has the shorter MLU, followed by OCDS and ADS. In Sesotho, there were slightly more single word utterances in CDS than in OCDS and in ADS. In French, there were no large differences between registers.

Within Sesotho CDS and OCDS, there were no large differences between speakers. Within Sesotho ADS, other child speech had the shortest MLU and highest proportion of one word utterances. Within French CDS and OCDS, there also were no large differences between speakers. We observe that mothers' speech had the longest MLU and least single word utterances. Overall, OCDS is more similar to CDS than ADS, and different speakers provide input of similar qualitative features within each register.

We then analysed lexical properties (Figure 2). The MATTR with a smaller word window offers the advantage of accounting for more data. The MATTR with a larger window is more stable, but is based on fewer sessions (the ones with a lot of speech). The smaller window MATTR is displayed here, whereas the larger window MATTR can be found in the Appendix (Figure A2). The properties were measured on the entirety of the lexicon (closed class words included). For both Sesotho and French, CDS was the least lexically diverse register and the register with the lowest hapax ratio, followed by OCDS and then by ADS. In Sesotho CDS and OCDS, there were small differences between speakers. In French CDS, differences between speakers were larger than in Sesotho, maternal speech being the most lexically diverse, and with the lowest hapax ratio of all speakers.

Last, we analysed some interactional properties for speech elicitation (Figure 3). For Sesotho, the highest proportion of questions and conversational turns (when an utterance was followed by an utterance produced by the person previously addressed to) was found overall in CDS, followed by OCDS, and then ADS. For French, similar ratios of questions and conversational turns were found overall in CDS and OCDS.

Within Sesotho CDS and OCDS, other adults and mothers had higher question ratios than other children. For ADS, adults had the highest question ratios. There were no large differences between speakers with respect to conversational turn ratios for CDS. Within French CDS, other adults and mothers also had higher question ratios than other children, but we see a difference between corpora in terms of conversational turns, with higher ratios for other adults in French but less so in Sesotho.

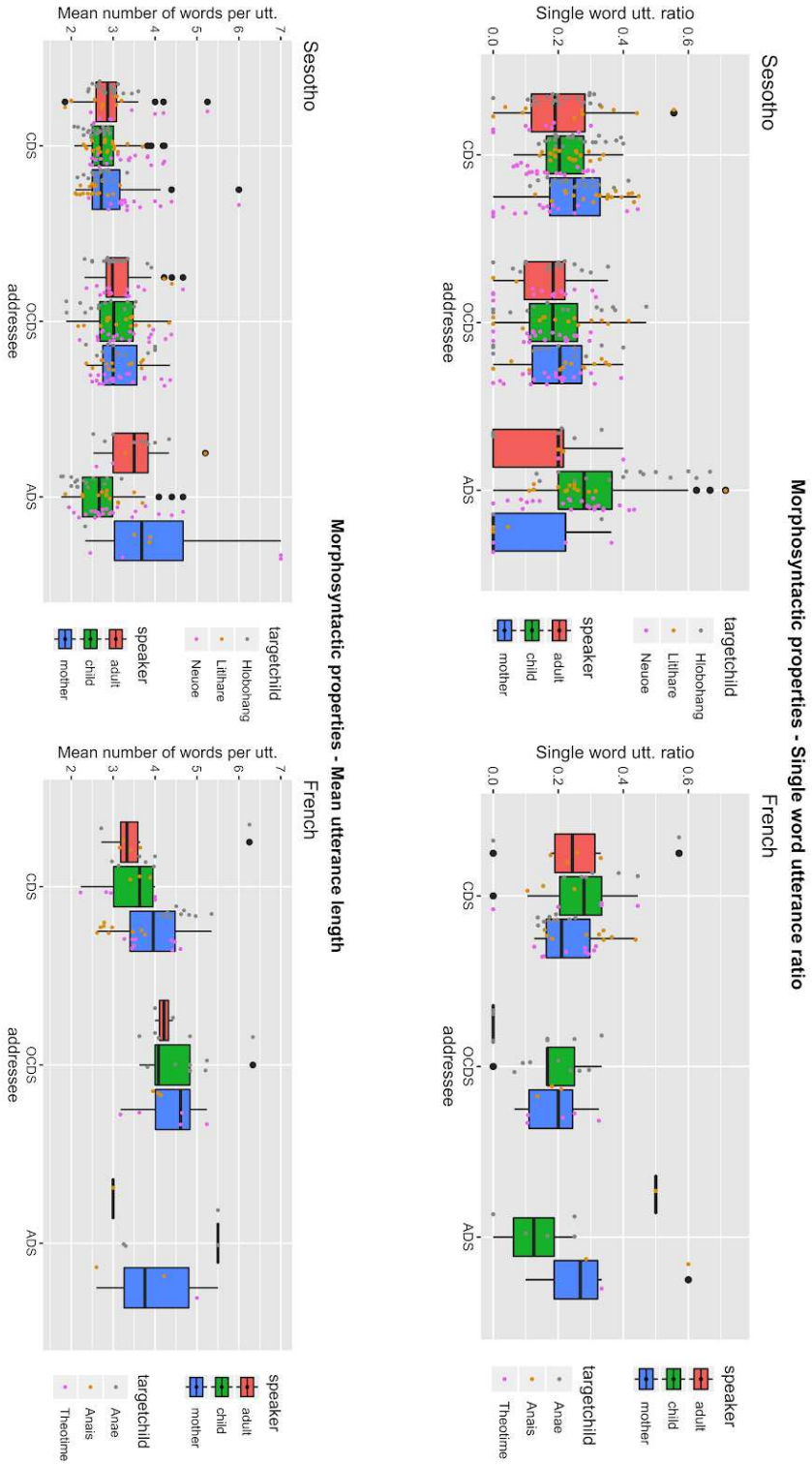

Figure 1. Mean length of utterance in words (bottom) and single word utterance ratio (top) for Sesotho (left) and French (right). Each point is a session (where the corresponding speaker produced at least one utterance). Boxplot colors indicate speakers, adult speakers in red, mothers in blue and other children speakers in green, and boxplot groups indicate the register, CDS at the left, OCDS at the middle and ADS at the right.

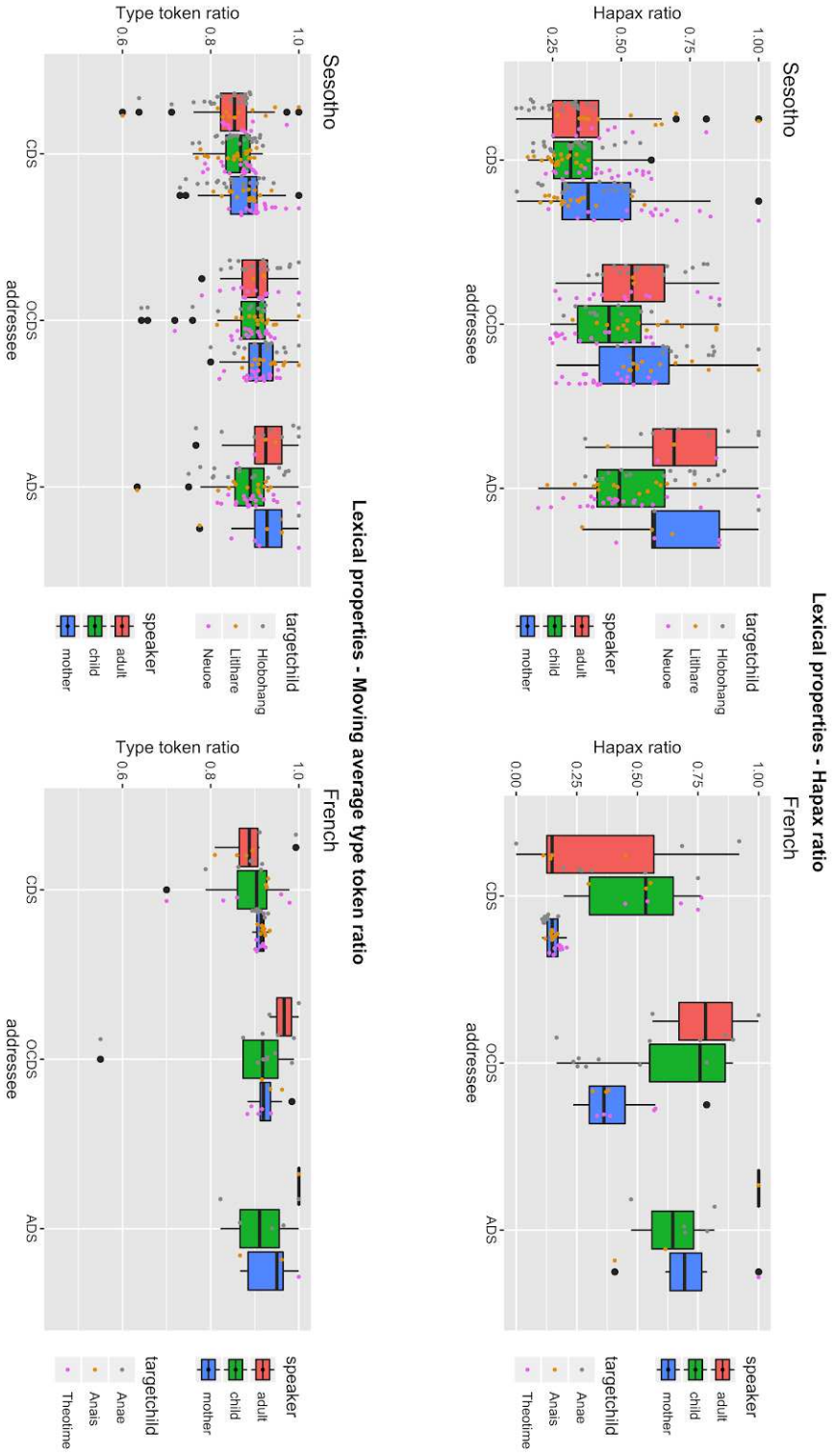

Figure 2. Type-token ratio (bottom) and hapax ratio (top) for Sesotho (left) and French (right). Each point is a session (where the corresponding speaker produced at least one utterance). Boxplot colors indicate speakers, adult speakers in red, mothers in blue and other children speakers in green, and boxplot groups indicate the register, CDS at the left, OCDS at the middle and ADS at the right.

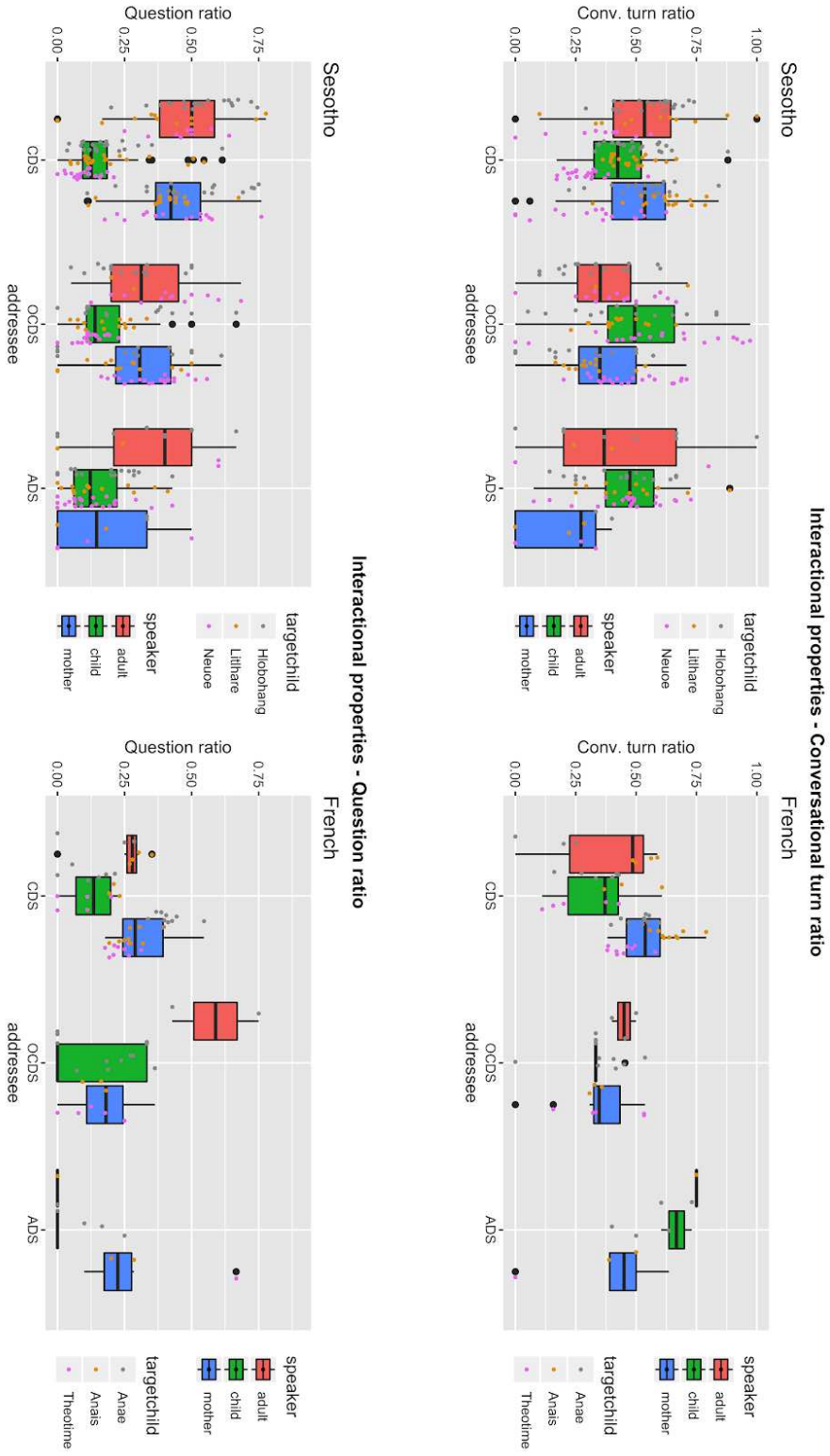

Figure 3. Question ratio (bottom) and ratio of conversational turns (top) for Sesotho (left) and French (right). Each point is a session (where the corresponding speaker produced at least one utterance). Boxplot colors indicate speakers, adult speakers in red, mothers in blue and other children speakers in green, and the x-axis indicates the register, CDS at the left, OCDS at the middle and ADS at the right.

Discussion

In the hope of informing the sparse literature on quantitative and qualitative descriptions of children's input across populations, the current study includes a detailed analysis of input to and around French- and Sesotho-learning children growing up in two very different cultural settings. We used the same descriptive metrics, in order to comprehensively document the linguistic input French and Sesotho-learning children receive from parents and other people. We described the amount and quality (the morphosyntactic, lexical and interactional features of speech) of linguistic input spoken by mothers, other adults and children, separately for child-directed speech, speech directed to other children and speech directed to adults. We found some differences across cultures, but also a great deal of similarity.

The composition of input differed dramatically between cultures. Maternal input was more dominant in the French corpus compared to the Sesotho one. In the latter, other children's input was more prevalent than maternal input (see Shneidman and Goldin-Meadow, 2012 for a similar observation in a Mayan village). Given that all target children had older siblings, the lower proportion of speech from other children in French indicates that French children may have fewer opportunities to hear speech produced by other children compared to Sesotho children. This difference in composition is also consistent with observations that mothers have more children to attend to in Lesotho, and children are surrounded by more children - at least in the rural village context - than in urban France. This is despite the fact that both the Sesotho corpus and the Lyon corpus were ‘designed’ to elicit spontaneous speech interactions between mother and target child; this becomes more challenging in Lesotho from the age of

2;6, either due to the birth of another sibling, or the increasing independence of the child by this ages, facilitating interactions with the larger peer group.

Thus, other children are an important source of input for many learners of Sesotho. This invites further research on how speech from other speakers contributes to children's input (see Sperry et al., 2019 for a similar suggestion), how essential input is from expert (i.e., adult) speakers, and what amounts of adult speech are necessary for language acquisition. That said, maternal speech is proportionally important in both corpora, and other child speech is not prevalent in the French corpora. These results are in line with previous findings for US families, documenting that siblings address their younger siblings much less frequently than what their mothers and other adults do (e.g. Oshima-Takane & Robbins, 2003).

In every other respect, marked similarities between cultures were observed. To begin with, CDS was the most prevalent register in both cultures. All target children in French settings were mostly exposed to speech directed to them. Two out of three Sesotho-learning children were mostly exposed to CDS, the third listening mostly to OCDS. This again points to the need to study diverse cultures; there are not only differences between industrial urban and non-industrial rural communities, but, given the little CDS found in some non-industrial cultures, there are also large differences between different non-industrial rural communities.

Another clear convergence across corpora is that CDS seems to have some similar features (e.g. ratio of questions, type-token ratio) in both French and Sesotho. This agrees with previous work showing that CDS is a simplified register adjusted to child listeners. However, this pattern of simplification is not exhibited across all features. For example, we found a

high ratio of questions in Sesotho. This agrees with previous observations (Demuth, 1992; 1995; Kline & Demuth, 2010). Moreover, for both cultures, OCDS seems to be more simplified than ADS, and has more similar patterns to CDS than what ADS does. This suggests that speech directed to both target and other children from all speakers, has CDS-like characteristics in both societies. As a result, it is conceivable that children who are exposed to a great deal of OCDS could learn more from this type of overheard speech than from ADS. Such a finding points to a need to study the characteristics of OCDS in diverse cultures, children's preference for OCDS over ADS, and whether children learn more from OCDS than from ADS.

In sum, more cross-cultural differences were found with respect to *who* addresses the key children, than with respect to *how* they address them. Child speakers of CDS in both cultures use a register similar to mothers' speech, especially in its morphosyntactic and lexical features (but this is not the case for hapax ratio in French). This suggests that other children also adjust their speech when talking to younger children.

However, we could not say the same with respect to speech-eliciting features. Children in both French and Sesotho have a lower ratio of questions than the mothers, in line with previous observational results (e.g. Dunn & Kendrick, 1982; Mannle et al., 1991). Mothers and adults have a higher ratio of conversational turns than other children in French. Interestingly, this pattern does not appear in Sesotho; mothers and adults have similar ratios of conversational turns to that of other children.

Other adults also use child-directed speech when talking to children, confirming some previous results (e.g. Shute & Wheldall, 1999, 2001). Adult speech addressed to children seems to be less lexically diverse than that addressed to other adults for both languages.

In sum, based on the results of our analysis, all target children grow up in environments where they are often directly addressed, either by their mother, or by children and other adults, and that input from different speakers is not dramatically different in qualitative terms, at least for the majority of the features studied here. This indicates that both French- and Sesotho-learning children grow up in stimulating environments for linguistic development, in agreement with previous anthropological studies (e.g. Connelly, 1984).

Before closing, we would like to mention some limitations of the study. This research is based on existing corpora from previous studies where the number of target children recorded was limited, due to the labor-intensive, longitudinal work involved in collecting daylong recordings and annotating fully spontaneous speech. A larger number of target children might have enabled us to conduct statistical tests on differences between registers, speakers and cultures, whereas the current study may remain descriptive.

Our annotators also suspected that speech from other children may be underestimated in French recordings, since older siblings sometimes seemed discouraged from speaking during the recording. Here is an example from Anaé's recording session 020804: when the older brother comes in the room, the mother says to him "Toi tu commences pas. (=Don't start.)" and after a while she says "Toi tu te tais s'il te plaît. (= Don't speak please.)". It is possible that siblings interact more with each other when they are not recorded, but it is also possible

that the same types of interactions happen when families are not recorded, which would mean that these are genuine cultural differences. It may be the case that French parents usually send away a noisy or agitated child when trying to focus on another child whereas Sesotho parents do not, which would partially explain why Sesotho-learning children receive more speech from other children than French-learning children.

Last, as explained above, the Sesotho sessions were recorded in the presence of the investigator and for comparability reasons, we chose to annotate sessions of French corpora where the investigator was also present during the recordings. Speech produced by or addressed to the investigator was removed in both corpora for this analysis. Future studies might be able to use recording devices where the investigator is not present, since the involvement of a stranger in the recording process might change families' behaviour.

For this study, we focused on the ratio of questions as a speech eliciting feature. However, additional speech-eliciting features and sentence types can be found in speech, such as imperatives. Future studies may want to include the ratio of imperatives in the analyses. Also, in Sesotho, other adults were mostly grandmothers and neighbors, whereas in French they were mostly fathers. In Sesotho, child members were a mixed-age group of siblings, peers and other children, whereas in French child members were mainly older siblings, with ages ranging from five to ten years. Future studies may focus on speech separately for these different speaker roles, looking at similarities and differences between e.g. sibling and peer speech, or father and grandmother speech. Last, we used data-driven measurements, such as length of utterance and lexical diversity for this study. However, language also consists of social patterns and is sensitive to context. For this reason, future studies may need to include

more comprehensive analyses, taking into consideration more features of engagement between partners, and the families' characteristics.

We look forward to further research incorporating more cultures, in order to investigate the impact of non-parental caregivers, and to check whether speech composition and qualities such as the ones found here are representative of quantifiable distinctions in other cultures.

References

Adi-Bensaid, L., Ben-David, A., & Tubul-Lavy, G. (2015). Content words in Hebrew child-directed speech. *Infant Behavior and Development*, *40*, 231–241. https://doi.org/10.1016/j.infbeh.2015.06.012

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209. https://doi.org/10.1111/j.1467-7687.2005.00406.x

Allen, S., & Crago, M. B. (1997). *Linguistic and cultural aspects of simplicity and complexity in Inuktitut (Eskimo) child-directed speech*. 91–102.

Bakermans-Kranenburg, M. J., IJzendoorn, M. H. van, & Kroonenberg, P. M. (2004). Differences in attachment security between African-American and white children: Ethnicity or socio-economic status? *Infant Behavior and Development*, *27*(3), 417–433. https://doi.org/10.1016/j.infbeh.2004.02.002

Barry, H., & Paxson, L. M. (1971). Infancy and Early Childhood: Cross-Cultural Codes 2. *Ethnology; Pittsburgh*, *10*(4), 466–508.

Barton, M. E., & Tomasello, M. (1991). Joint Attention and Conversation in Mother-Infant-Sibling Triads. *Child Development*, *62*(3), 517–529. https://doi.org/10.1111/j.1467-8624.1991.tb01548.x

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019).

What Do North American Babies Hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724. https://doi.org/10.1111/desc.12724

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

Burch, T. K. (1970). Some demographic determinants of average household size: An analytic approach. *Demography*, *7*(1), 61–69. https://doi.org/10.2307/2060023

Burch, T. K. (1979). Household and Family Demography: A Bibliographic Essay. *Population Index*, *45*(2), 173–195. JSTOR. https://doi.org/10.2307/2735726

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early Language Experience in a Tseltal Mayan Village. *Child Development*. https://doi.org/10.1111/cdev.13349

Connelly, M. (1984). *Basotho children's acquisition of noun morphology* [University of Essex]. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.303888

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child‐Directed Speech Is Infrequent in a Forager‐Farmer Population: A Time Allocation Study. *Child Development*, *90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Demuth, K. (1986). Prompting routines in the language socialization of Basotho children. *Language Socialization Across Cultures*, 51–79.

Demuth, K. (1992). *Acquisition of Sesotho.* In *The cross-linguistic study of language acquisition* (pp. 557-638). Lawrence Erlbaum Associates.

Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, *35*(1), 99–127. https://doi.org/10.1017/S0305000907008276

Dunn, J., & Kendrick, C. (1982). The speech of two- and three-year-olds to infant siblings: 'baby talk' and the context of communication. *Journal of Child Language*, *9*, 579–595.

Dunn, J., & Shatz, M. (1989). Becoming a Conversationalist despite (Or Because of) Having an Older Sibling. *Child Development*, *60*(2), 399–410. https://doi.org/10.2307/1130985

Tamis‐LeMonda, C. S., Shannon, J. D., Cabrera, N. J., & Lamb, M. E. (2004). Fathers and mothers at play with their 2‐and 3‐year‐olds: Contributions to language and cognitive development. *Child development*, *75*(6), 1806-1820.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*(3), 637–656.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(03), 477–501. https://doi.org/10.1017/S0305000900010679

Fisher, C., & Tokura, H. (1996). Acoustic Cues to Grammatical Structure in Infant-Directed Speech: Cross-Linguistic Evidence. *Child Development*, *67*(6), 3192–3218. https://doi.org/10.1111/j.1467-8624.1996.tb01909.x

Fouts, H. N., Roopnarine, J. L., Lamb, M. E., & Evans, M. (2012). Infant Social Interactions With Multiple Caregivers: The Importance of Ethnicity and Socioeconomic Status. *Journal of Cross-Cultural Psychology*, *43*(2), 328–348. https://doi.org/10.1177/0022022110388564

Gallimore, R., Boggs, J., & Jordan, C. (1974). *Culture, Behavior and Education: A Study of Hawaiian Americans*. Beverly Hills, CA: Sage.

Geertz, H. (1989). *The Javanese family: A study of kinship and socialization*. Prospect Heights, Illinois : Waveland Press.. http://hdl.handle.net/2027/heb.04452.0001.001.

Hamilton, A. (1981). *Nature and nurture: Aboriginal child-rearing in north-central Arnhem Land*. Australian Institute of Aboriginal Studies. https://trove.nla.gov.au/work/25087621?selectedversion=NBD2032630.

Harkness, S. (1977). Aspects of Social Environment and First Language Acquisition in Rural Africa. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children* (pp. 309–316). Cambridge

University Press.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children* (pp. xxiii, 268). Baltimore: Paul H Brookes Publishing.

Henning, A., Striano, T., & Lieven, E. V. M. (2005). Maternal speech to infants at 1 and 3 months of age. *Infant Behavior and Development*, *28*(4), 519–536. https://doi.org/10.1016/j.infbeh.2005.06.001

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Hoff, E. (2006). How social contexts support and shape language development☆. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E, & Naigles, L. (2002). How Children Use Input to Acquire a Lexicon. *Child Development*, *73*(2), 418–433. https://doi.org/10.1111/1467-8624.00415

Hofferth, S. L. (1995). Caring for children at the poverty line. *Children and Youth Services Review*, *17*(1), 61–90. https://doi.org/10.1016/0190-7409(95)00004-V

Hoff-Ginsberg, E., & Krueger, W. M. (1991). Older Siblings as Conversational Partners. *Merrill-Palmer Quarterly (1982-)*, 465-481.5

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

INSEE. (2013). *Ménages—Familles − Tableaux de l'Économie Française | Insee*. Retrieved December 2, 2020 from https://www.insee.fr/fr/statistiques/1906666?sommaire=1906743

Ispa, J. M., Fine, M. A., Halgunseth, L. C., Harper, S., Robinson, J., Boyce, L., Brooks‐Gunn, J., & Brady‐Smith, C. (2004). Maternal Intrusiveness, Maternal Warmth, and Mother–Toddler

Relationship Outcomes: Variations Across Low-Income Ethnic and Acculturation Groups. *Child Development*, *75*(6), 1613–1631. https://doi.org/10.1111/j.1467-8624.2004.00806.x

Kral, I., & Ellis, E. M. (2008). Children, language and literacy in the Ngaanyatjarra Lands. In *Http://trove.nla.gov.au/version/44685897*. Continuum Publishing Company. https://openresearch-repository.anu.edu.au/handle/1885/38788

Kruper, J. C., & Užgiris, I. C. (1987). Fathers' and mothers' speech to young infants. *Journal of Psycholinguistic Research*, *16*(6), 597–614. https://doi.org/10.1007/BF01067087

Kulick, D. (1992). Anger, gender, language shift and the politics of revelation in a Papua New Guinean village. *Pragmatics*, *2*(3), 281-296.

Lamm, B. (2008). *Children´s ideas about infant care: A comparison of rural Nso children from Cameroon and German middle class children*. (Doctoral dissertation, University of Osnabrück). Retrieved from https://repositorium.ub.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-2008080129

Lee, C. S., Kitamura, C., Burnham, D., & Todd, N. P. (2014). On the rhythm of infant- versus adult-directed speech in Australian English. *Journal of the Acoustical Society of America*, *136*(1), Article 1. http://research.gold.ac.uk/10791/

Leffel, K. R., & Suskind, D. L. (2013). Parent-directed approaches to enrich the early language environments of children living in poverty. *Seminars in Speech and Language*. https://doi.org/10.1055/s-0033-1353443

Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In *Input and interaction in language acquisition* (pp. 56–73). Cambridge University Press. https://doi.org/10.1017/CBO9780511620690.005

Loakes, D., Moses, K., Wigglesworth, G., Simpson, J., & Billington, R. (2013). Children's language input: A study of a remote multilingual Indigenous Australian community. *Multilingua*, *32*(5). https://doi.org/10.1515/multi-2013-0032

Loukatou, G., LeNormand, M.-T., & Cristia, A. (2019). *Is it easier to segment words from infant- than*

adult-directed speech? Modeling evidence from an ecological French corpus. Proceedings of the 41st Conference of Cognitive Science Society.

MacWhinney, B. (2000). *The CHILDES Project: The database*. Psychology Press.

Mannle, S., Barton, M., & Tomasello, M. (1991). Two-year-olds' conversations with their mothers and preschool-aged siblings. *First Language*, *12*, 57–71.

Martini, M., & Kirkpatrick, J. (1992). Parenting in Polynesia: A view from the Marquesas. In J. L Roopnarine & D. Bruce (Eds.), Annual advances in applied developmental psychology: Vol. 5. *Parent-child Socialization in Diverse Cultures* (pp. 199–222). Norwood, NJ: Ablex.

McLaughlin, B., White, D., McDevitt, T., & Raskin, R. (1983). Mothers' and fathers' speech to their young children: Similar or different? *Journal of Child Language*, *10*(1), 245–252. https://doi.org/10.1017/S0305000900005286

Morgenstern, A., & Parisse, C. (2012). The Paris Corpus. *Journal of French Language Studies*, *22*(1), 7–12. https://doi.org/10.1017/S095926951100055X

Nielsen, M., Haun, D., Kärtner, J., & Legare, C.(2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31-38.

Ochs, E., & Schieffelin, B. B. (1994). Language Scoialization and its Consequences for Language Development. In *Handbook of Child Language* (pp. 73–94). London:Blackwell.

O'Shannessy, C. (2013). The role of multiple sources in the formation of an innovative auxiliary category in Light Warlpiri, a new Australian mixed language. *Language*, *89*(2), 328–353. https://doi.org/10.1353/lan.2013.0025

Oshima-Takane, Y., & Robbins, M. (2003). Linguistic Environment of Secondborn Children. *First Language*, *23*(1), 21–40. https://doi.org/10.1177/0142723703023001002

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal Correlates of Growth in Toddler Vocabulary Production in Low-Income Families. *Child Development*, *76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Pancsofar, N., & Vernon-Feagans, L. (2006). Mother and father language input to young children:

Contributions to later language development. *Journal of Applied Developmental Psychology*, *27*(6), 571–587. https://doi.org/10.1016/j.appdev.2006.08.003

Pye, C. (1986). Quiché Mayan speech to children. *Journal of Child Language*, *13*(1), 85–100. https://doi.org/10.1017/S0305000900000313

Read, M. (1968). *Children of Their Fathers: Growing Up Among the Ngoni of Malawi.* International Thomson Publishing.

Richman, A. L., Miller, P. M., Levine, R. A., & Al, R. E. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology*, 614–621.

Rogoff, B. (1981). Adults and Peers as Agents of Socialization: A Highland Guatemalan Profile. *Ethos*, *9*(1), 18–36. https://doi.org/10.1525/eth.1981.9.1.02a00030

Rogoff, B. (2003). *The Cultural Nature of Human Development*. Oxford University Press.

Rondal, J. A. (1980). Fathers' and mothers' speech in early language development. *Journal of Child Language*, *7*(2), 353–369. https://doi.org/10.1017/S0305000900002671

Roopnarine, J. L., Fouts, H. N., Lamb, M. E., & Lewis-Elligan, T. Y. (2005). Mothers' and Fathers' Behaviors Toward Their 3- to 4-Month-Old Infants in Lower, Middle, and Upper Socioeconomic African American Families. *Developmental Psychology*, *41*(5), 723–732. https://doi.org/10.1037/0012-1649.41.5.723

Scaff, C., Stieglitz, J., Casillas, M., & Cristia, A. (in preparation). Language input in a hunter-forager population: Estimations from daylong recordings.

Shatz, M., & Gelman, R. (1973). The Development of Communication Skills: Modifications in the Speech of Young Children as a Function of Listener. *Monographs of the Society for Research in Child Development*, *38*(5), 1–38. https://doi.org/10.2307/1165783

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*(1), 169–201. https://doi.org/10.1017/S0305000997003395

Shneidman, L. A., & Goldin‐Meadow, S. (2012). Language input and acquisition in a Mayan village:

How important is directed speech? *Developmental Science*, *15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Shute, B., & Wheldall, K. (1999). Fundamental frequency and temporal modifications in the speech of British fathers to their children. *Educational Psychology*, *19*(2), 221–233. https://doi.org/10.1080/0144341990190208

Shute, B., & Wheldall, K. (2001). How do grandmothers speak to their grandchildren? Fundamental frequency and temporal modifications in the speech of British grandmothers to their grandchildren. *Educational Psychology*, *21*(4), 493–503. https://doi.org/10.1080/01443410120090858

Simmons, T., & Dye, J. L. (2003). *Grandparents Living with Grandchildren: 2000. Census 2000 Brief*. Retrieved December 2, 2020 from https://eric.ed.gov/?id=ED482412

Smith-Hefner, N. J. (1988). The Linguistic Socialization of Javanese Children in Two Communities. *Anthropological Linguistics*, *30*(2), 166–198.

Snow, C., & Ferguson, C. (1979). *Talking to Children: Language Input and Acquisition*. Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Language Does Matter: But There is More to Language Than Vocabulary and Directed Speech. *Child Development*, *90*(3), 993–997. https://doi.org/10.1111/cdev.13125

Spiro, M. (1958). *Children of the Kibbutz*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674366077

Standing, T. S., Musil, C. M., & Warner, C. B. (2007). Grandmothers' Transitions in Caregiving to Grandchildren. *Western Journal of Nursing Research*, *29*(5), 613–631. https://doi.org/10.1177/0193945906298607

Stewart, R. B., & Marvin, R. S. (1984). Sibling Relations: The Role of Conceptual Perspective-Taking in the Ontogeny of Sibling Caregiving. *Child Development*, *55*(4), 1322–1332. https://doi.org/10.2307/1130002

Strapp, C. M. (1999). Mothers', fathers', and siblings' responses to children's language errors: Comparing sources of negative evidence. *Journal of Child Language*, *26*(2), 373–391. https://doi.org/10.1017/S0305000999003827

The ManyBabies Consortium, Frank, M. C., Alcock, K. J., Arias-Trejo, N., … Soderstrom, M. (2020). Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Thupayagale‐Tshweneagae, G. (2008). Psychosocial effects experienced by grandmothers as primary caregivers in rural Botswana. *Journal of Psychiatric and Mental Health Nursing*, *15*(5), 351–356. https://doi.org/10.1111/j.1365-2850.2007.01232.x

Toda, S., Fogel, A., & Kawai, M. (1990). Maternal speech to three-month-old infants in the United States and Japan. *Journal of Child Language*, *17*(2), 279–294. https://doi.org/10.1017/S0305000900013775

Veneziano, E., & Parisse, C. (2010). The acquisition of early verbs in French: Assessing the role of conversation and of child-directed speech. *First Language*, *30*(3–4), 287–311. https://doi.org/10.1177/0142723710379785

Weisleder, A., & Waxman, S. R. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. *Journal of Child Language*, *37*(5), 1089–1108. https://doi.org/10.1017/S0305000909990067

Weisner, T. S., Gallimore, R., Bacon, M. K., Barry, H., Bell, C., Novaes, S. C., Edwards, C. P., Goswami, B. B., Minturn, L., Nerlove, S. B., Koel, A., Ritchie, J. E., Rosenblatt, P. C., Singh, T. R., Sutton-Smith, B., Whiting, B. B., Wilder, W. D., & Williams, T. R. (1977). My Brother's Keeper: Child and Sibling Caretaking. *Current Anthropology*, *18*(2), 169–190.

https://doi.org/10.1086/201883

Weppelman, T. L., Bostow, A., Schiffer, R., Elbert-Perez, E., & Newman, R. S. (2003). Children's use of the prosodic characteristics of infant-directed speech. *Language & Communication*, *23*(1), 63–80. https://doi.org/10.1016/S0271-5309(01)00023-4

Whiting, J. W. M. (1941). *Becoming a Kwoma: Teaching and learning in a New Guinea tribe*. Pub. for the Institute of human relations by Yale university press.

Williams, T. (1971). *A Borneo childhood:Enculturation in Dusun society*. Holt, Rinehart and Winston.

Wolff, J. U., & Poedjosoedarmo, S. (1984). Communicative Codes in Central Java. *Language*, *60*(1), 196. https://doi.org/10.2307/414227

Zukow-Goldring, P. (2002). Sibling caregiving. In *Handbook of parenting: Being and becoming a parent, Vol. 3, 2nd ed* (pp. 253–286). Lawrence Erlbaum Associates Publishers.

Appendix

	CDS			OCDS			ADS		
	MOT	ADU	OCHI	MOT	ADU	OCHI	MOT	ADU	OCHI
Sesotho Hlobohang	13.74	37.08	**39.01**	1.02	**3.01**	2.77	.21	.54	**2.58**
Sesotho Litlhare	40.25	2.06	**48.64**	2.55	0.23	**2.58**	.39	.35	**2.55**
Sesotho Neuoe	4.89	5.49	**16.61**	18.55	7.12	**24.61**	.29	.16	**21.60**
French Anaé	**79.88**	.15	7.01	**5.59**	.22	.47	.44	.37	**4.73**
French Anais	**65.37**	27.40	1.51	**2.42**	-	.01	.41	.24	**2.38**
French Theotime	**90.37**	-	1.43	4.06	-	-	.12	-	**3.37**

Table A1. Percentage of speech produced by mothers (MOT), other adults (ADU) and other children (OCHI) comprising CDS, OCDS and ADS for each child. The largest number in each register-based

group of cells is in bold. We observe that CDS, OCDS and ADS were mostly produced by other children for all Sesotho-learning children, except for Hlobohang, for whom other adults and other children contributed almost equally to the total OCDS. For the French-learning children, the mother was the main source of CDS and OCDS, and other children were the main source of ADS.

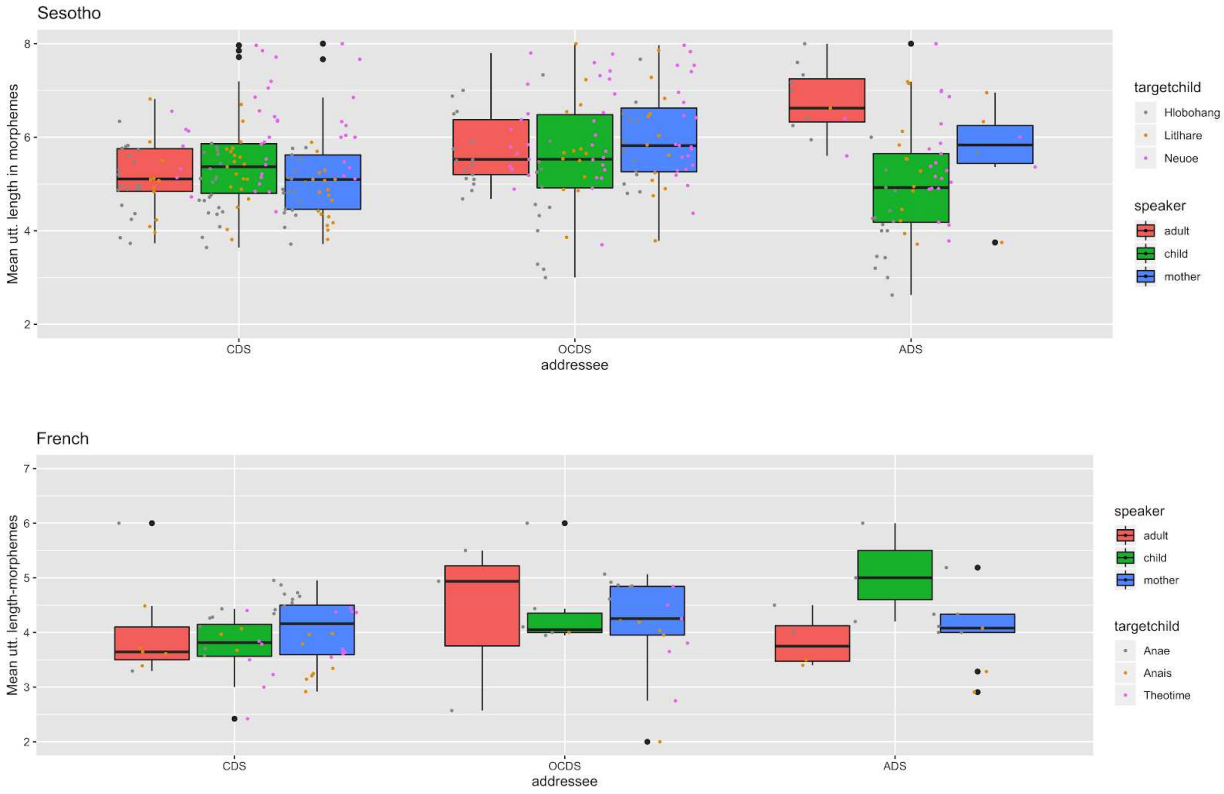

Figure A1. Mean utterance length **in morphemes** for Sesotho (top) and French (down). Morphological analysis for French was generated with CLAN (MacWhinney, 2000). For example, the Sesotho verb 'ke-tla-mo-otla' is counted as four morphemes 'ke tla mo otla'. The same phrase in French 'Je vais le frapper' is counted as five morphemes 'Je vais le frapp er'). Each point is a session. Boxplot colors indicate speakers, adult speakers in red, mothers in blue and other children speakers in green, and the x-axis indicates the register, CDS at the left, OCDS at the middle and ADS at the right.

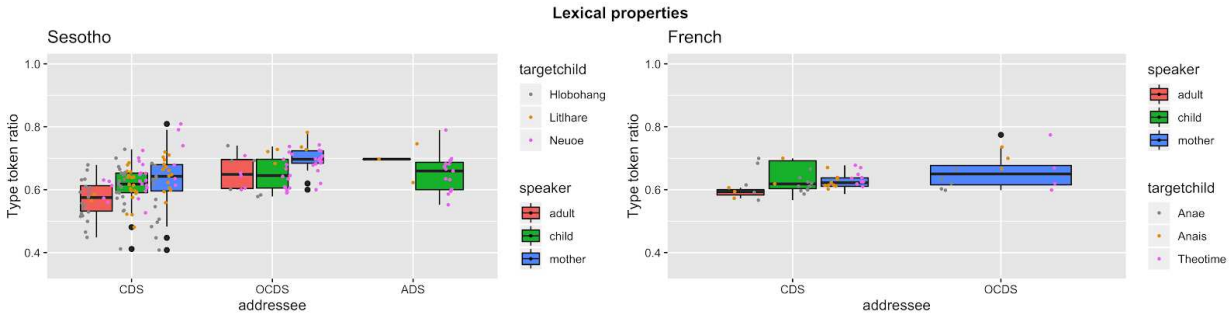

Figure A2. Moving average type token ratio for Sesotho (left) and French (right) within a **window of 100 words**.

9. Is it easier to segment words from infant- than adult-directed speech? Modeling evidence from an ecological French corpus *

Abstract: Infants learn language by exposure to streams of speech produced by their caregivers. Early on, they manage to segment word forms out of this continuous input, which is either directly addressed to them, or directed to other adults, thus overheard. It has been suggested that infant-directed speech is simplified and could facilitate language learning. This study aimed to investigate whether features such as utterance length, segmentation entropy and lexical diversity could account for an advantage in segmentability of infant-directed speech. A large set of word segmentation algorithms was used on an ecologically valid corpus, consisting of 18 sets of recordings gathered from French-learning infants aged 3-48 months. A series of textual analyses confirmed several simplicity features of infant-, compared to adult-directed speech. A small segmentation advantage was also documented, which could not be attributed to any of those corpus features. Some particularities of the data invite further research on more corpora.

***Loukatou, G., Le Normand, M.-T. & Cristia, A. (2019). Is it easier to segment words from infant-directed speech? Modeling evidence from an ecological French corpus. *Proceedings of the 41st Conference of Cognitive Science Society*.**

Is it easier to segment words from infant- than adult-directed speech? Modeling evidence from an ecological French corpus

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

Marie-Thérèse Le Normand (marielenormand@mac.com)
INSERM & LPP (Laboratoire Psychopathologie et Processus de Santé), Université Paris Descartes, Sorbonne
Paris, France

Alejandrina Cristia (alecristia@gmail.com)
Laboratoire de sciences cognitives et psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University
Paris, France

Abstract

Infants learn language by exposure to streams of speech produced by their caregivers. Early on, they manage to segment word forms out of this continuous input, which is either directly addressed to them, or directed to other adults, thus overheard. It has been suggested that infant-directed speech is simplified and could facilitate language learning. This study aimed to investigate whether features such as utterance length, segmentation entropy and lexical diversity could account for an advantage in segmentability of infant-directed speech. A large set of word segmentation algorithms was used on an ecologically valid corpus, consisting of 18 sets of recordings gathered from French-learning infants aged 3-48 months. A series of textual analyses confirmed several simplicity features of infant-, compared to adult-directed speech. A small segmentation advantage was also documented, which could not be attributed to any of those corpus features. Some particularities of the data invite further research on more corpora.

Keywords: language acquisition; infant-directed speech; computational modeling; word segmentation; unsupervised learning

Introduction

Infants acquire language early on, building a vocabulary of several hundred word forms by 11 months of life (Ngon et al., 2013). Since most word forms do not appear in isolation (Brent & Siskind, 2001), much previous work studies how infants segment (i.e., pull out) forms from their caregivers' running input. A close look at this input shows that it is not homogeneous, but instead contains some speech addressed to the infants themselves (infant-directed speech or IDS) and some speech overheard by infants which is addressed to others, including adults (adult-directed speech or ADS). These two speech registers differ along many dimensions, including some that may impact word segmentation.

Broadly, IDS has been claimed to present properties that would facilitate language acquisition, with IDS being phonologically, syntactically, and semantically simplified (Soderstrom, 2007). Other characteristics are more relevant to word segmentation. First, IDS may have a higher proportion of single-word phrases (Brent & Siskind, 2001), and phrases might be shorter in length (Newport, Gleitman, & Gleitman, 1977) than in ADS. In shorter phrases, more words would occur at phrase edges, which should improve segmentation: Phrase edges, easily perceptible, are word boundaries provided "for free". Indeed, infants may be more successful at recognizing and segmenting phrase-final words (E. Johnson, Seidl, & Tyler, 2014). Additionally, shorter phrases entail that the set of possible segmentations for each phrase is smaller, lowering segmentation ambiguity. For instance, Fourtassi, Börschinger, Johnson, and Dupoux (2013) showed that ADS might be more ambiguous to segment, when comparing an ADS to an IDS corpus. Second, words may be shorter (Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011), which should mean that word, morphemes, and syllable boundaries coincide more often and there are fewer places to posit or miss positing a boundary. Third, there may be more repetitions, therefore fewer hapaxes (words uttered only once), and overall less lexical diversity (Soderstrom, 2007). Low lexical diversity means fewer target words need to be found. There might be more cues to help segment out frequently repeated words, than words that appear rarely or once. Indeed, one computational modeling study found that artificially reducing phrase length and increasing word repetition in a corpus improved word segmentation with one word segmentation model (Batchelder, 1997). Based on these hypotheses and previous work, we predict that the task of recovering wordforms is easier in IDS than ADS.

Naturally, IDS features may not be the same across infant ages. IDS addressed to very young infants may differ from that addressed to older infants, possibly resembling ADS more as infants get older. For example, IDS features may become less accentuated as the infant grows up; repetitions might decrease, utterance length and lexical diversity increase with age (Henning, Striano, & Lieven, 2005; Soderstrom, 2007). According to the hypotheses explained above, IDS addressed to younger infants should be "easier" to segment than IDS to older infants.

In this paper, we aim to address the question of whether it is easier to segment wordforms from IDS than ADS, using multiple word segmentation models, and taking into account changes with infants' age. In the next section, we review

previous modeling work more thoroughly, before introducing our own approach.

Previous studies

Some studies tested whether infants learn more from IDS than ADS in an experimental situation. However, improvements for IDS compared to ADS could be due to the fact that infants pay more attention when they listen to IDS, and thus learn more from it. This method cannot reveal whether, above and beyond this attentional effect, there are intrinsic *informational* differences that affect segmentability. Fortunately, there is a complementary method to approach this question with a colder eye, which builds on computational models of word segmentation. The input to such word segmentation models is usually speech transcriptions, in order to control for differences such as attention capture and acoustic implementation. Segmentation models used for this method are based on findings by experimental studies that infants might make use of statistical cues. Computational models of infant word segmentation can be grouped into two conceptual classes: lexical and sublexical. Sublexical models segment based on local cues, such as transitional probabilities and phonotactics. Lexical models build a lexicon based on recurrent chunks of speech identified with Bayesian probabilities or by memorizing isolated words.

Little previous modeling work has specifically compared IDS and ADS. Four representative studies are summarized in Table 1. For these four studies, improved segmentation performance was found for IDS than ADS: 15% for Batchelder (2002), 5-8% for Fourtassi et al. (2013), 2-10% for Ludusan, Mazuka, Bernard, Cristia, and Dupoux (2017) and 3-10% for Daland and Pierrehumbert (2011). A recent paper critiqued this previous work as follows (Cristia, Dupoux, Ratner, & Soderstrom, 2018). IDS mainly involved caregivers addressing their infants during predefined tasks (e.g., a play session in the laboratory) or in short visits to the child's home. In the former case, by constraining the context, the structure and lexicon of caregivers might have been limited and adapted to that task. And in both cases, being observed could affect caregivers' behavior, who might produce less spontaneous and more formal speech. Moreover, ADS was mostly addressed to an unfamiliar person (experimenter). These conversations are likely more formal than ADS between caregivers in daily life, and could increase the complexity of the speech. As shown by E. Johnson, Lahey, Ernestus, and Cutler (2013), IDS differs more from ADS to unfamiliar adults, than ADS to familiar adults. This could result in increased qualitative differences between registers and probably overestimated differences in segmentability.

Indeed, Cristia et al. (2018) recently documented a considerably smaller IDS advantage when modeling segmentation on an ecological English IDS and ADS corpus. The corpus consisted of transcriptions from excerpts of day-long recordings; thus infants' linguistic environment was recorded while they were going on with their daily lives, resulting in realistic IDS and ADS. Across a wide range of lexical and sublexical models, the IDS advantage ranged from -2% to 8%, with only 3 models providing evidence of an advantage greater than a measure of error. Interestingly, the difference between registers was further reduced when IDS was matched to ADS in corpus length.

The present study

We contribute to this literature in three main ways. First, we specifically describe IDS-ADS differences using various corpus description tools. We compare the registers in: phrase length, word length, ratio of single word phrases, intrinsic segmentation ambiguity (using segmentation entropy), lexical diversity (using Moving Average Type-Token Ratio – MATTR–, so as to control for corpus size), and ratio of hapaxes. Some, but not all of these features have been separately looked at in previous studies (i.e. Fourtassi et al., 2013 measured segmentation ambiguity and Batchelder, 1997 measured word and phrase length, repetitiveness). This is the first study to systematically investigate a plurality of language features on the same IDS-ADS corpus. We test whether IDS is simpler than ADS, as far as these features are concerned. Moreover, following Batchelder (2002), we further investigate whether variation in these features can actually account for the segmentability of a register.

Second, IDS corpora coming from a wide infant age range have been used by previous research, but IDS addressed to infants of different ages were, most of the times, merged together. One exception is Batchelder (1997), who documented that IDS to younger children (13-18 months) produced more successful results than IDS to older children (22-25 months), whereas ADS results from mothers of younger versus older infants didn't differ. In this paper, we specifically ask whether some IDS features interact with infant age and whether segmentability of IDS might actually be affected by age. For that, we include IDS and ADS from a wide age range, and further investigate possible correlations between features, segmentation scores, and infant age.

Third, we follow Cristia et al. (2018) by analyzing a completely ecological child-centered corpus, based on excerpts of day-long recordings, and which thus contains natural ADS and IDS as the child hears over the course of the day. The results of our study would provide more evidence to the question whether differences in home-recorded IDS and ADS are smaller than those between less controlled IDS-ADS contrasts (see Table 1).

In addition to these three main contributions, we extend the range of languages studied to European French.

Methods

We segmented IDS and ADS of each infant separately. Scripts used for corpus preprocessing, phonologization, and segmentation as well as results and supplementary material are available at https://osf.io/6vwse/?view_only=0bc4f6c0e23040cbbb92e26d414d4a7a. Statistical analyses were carried out in R (R Core Team, 2013).

Table 1: Summary of design in previous modeling studies comparing IDS and ADS segmentation. In Language(s), Eng stands for English, Jap for Japanese, Span for Spanish. Under IDS and ADS, we describe the corpora. The specific corpora used were: R= RIKEN; H= Hamasaki; C= Spontaneous Japanese; BR= Bernstein Ratner; B= Buckeye; D= Deuchar & Clark 1992, Marrero; M= Miyata 1995; novel= Moon and the Sixpense; short stories were written by Alejandro Dolina (MacWhinney, 1996). Under model, we note the type of model used: lex for lexical and sublex for sublexical.

Study	Language(s)	Infant age(s)	IDS	ADS	model
Batchelder (2002)	Eng.	1;1-1;9	play session (BR)	novel	1 lex
Batchelder (2002)	Span.	1;8-8;0	CHILDES (D)	short story	1 lex
Batchelder (2002)	Jap.	1;3-3;1	home play session (M)	science book	1 lex
Daland et al. (2011)	Eng.	various	all CHILDES	interview (B)	1 sublex
Fourtassi et al. (2013)	Eng.	1;1-1;9	play session (BR)	interview (B)	1 lex
Fourtassi et al. (2013)	Jap.	2;2-3;7	play session (H)	lecture (C)	1 lex
Ludusan et al. (2017)	Jap.	1;6-2;0	play session (R)	lecture (C)	1 lex, 3 sublex

Corpus

Sixteen typically developing native French-speaking infants (eight girls, eight boys; ages 3-48 months, M=20, SD=13), whose families were highly educated, were included. Two of the infants were recorded at two different ages. Each child was recorded 10-16 hours per day, three days a week, in their natural environments. The original recordings are available online (Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2016a, 2016b; VanDam et al., 2016). Next, 18 10-min samples, totaling 3 hours per child (1 hour per day), were selected for orthographic transcription by two native French speakers, as detailed in Canault et al. (2016b). The main criteria for selection reported was that a number of activities were sampled, and that there be a high number of productions by the child and the adult. For the present project, the transcriptions of the first day for all infants were corrected by a native French speaker, who made sure that the definition of utterance was stable (and corrected any other errors, such as misattributions or orthographic errors). The coder annotated whether an adult caregiver's utterance was directed to the target child, an adult, or other, using content and context. Utterances addressed to the target child constituted the IDS corpus and those directed to an adult were the ADS corpus.

Pre-processing

Pre-processing was carried out using custom scripts written mainly in bash and in python, available from `https://github.com/georgialoukatou/French_ADS_IDS_segmentation_Lyon`. All extraneous codes (such as punctuation marks or "xxx", the code indicating that what was said could not be understood by the transcriber) were removed, leaving only the orthographic representation of the adults' speech. The corpora were phonologized with the French voice of the espeak TTS system (Duddington, 2012), using the phonemizer wrapper (Bernard, 2018), which further syllabifies according to the Maximum Onset Principle.

Before segmentation, all spaces between words were removed, leaving the input parsed into minimal units. The minimal units were either phones or syllables. Both phonemes and syllables were tested with all models. Utterance boundaries were preserved as such, since they are supposedly salient to infants (Shukla, White, & Aslin, 2011). This constitutes the input to the model. After preprocessing, the 18 infant-directed corpora contained M=487 (SD 350) utterances (range 84 to 1,172 utterances). The 18 adult-directed corpora contained M=238 (SD 230) utterances (range 15 to 780 utterances).

For comparability with previous work, we evaluate the models' performance using lexical token F-scores, measured by comparing the original version of the input (with spaces between words) against the one returned by the model (with spaces in the hypothesized breaks).

Segmentation

Both corpus description and segmentation were carried out using the WordSeg package (Bernard et al., 2018), available from `https://github.com/bootphon/wordseg/`. Due to space limits, the algorithms are only briefly described here. Full technical details can be found in `https://wordseg.readthedocs.io/`. All algorithms are unsupervised, and inspired in infant experimental work.

We used two representatives of the sublexical word segmentation class contains, called DIBS and TP for short. The Diphone Based Segmentation algorithm (DiBS; Daland & Pierrehumbert, 2011) is based on the idea than a phoneme sequence often spanning phrase boundaries would probably span word breaks.

The Transitional Probabilities algorithm family (TP; Saksida, Langus, & Nespor, 2017) is based on the concept that syllable pairs with lower statistical coherence tend to span word breaks. Forward TP (FTP) measures the frequency of occurrence of the syllabic sequence AB given the frequency of occurrence of the syllable A. Backward TP (BTP) measures the frequency of occurrence of the syllabic sequence AB given the frequency of occurrence of the syllable B. The Relative versions (FTPr or BTPr) threshold TPs against that of neighboring sequences. The Absolute versions

Table 2: Paired t-tests measuring feature differences across IDS and ADS. Word length is measured in phonemes. % 1-w phrase stands for ratio of single word phrases. % hapaxes stands for percent of hapaxes. IDS gives the mean values of each feature on the IDS corpus, with standard deviation in parentheses. ADS shows the mean values of each feature on the ADS corpus with standard deviation in parentheses. The window size for MATTR is 10 words. "p" gives the p-value of the t-test.

Feature	IDS	ADS	p
Word length	2.86 (.08)	2.80 (.11)	.071
Phrase length	5.89 (.85)	6.73 (.86)	*
% 1-w phrase	.18 (.06)	.13 (.05)	**
Entropy	.02 (.004)	.03 (.01)	.31
MATTR	.89 (.03)	.93 (.02)	***
% hapaxes	.39 (.22)	.48 (.27)	***

(FTPa or BTPa) instead threshold on the average of all TPs over the sum of different syllable bigrams.

We used two representatives of the lexical class as well: AG and PUDDLE. Adaptor Grammar (AG) uses the Pitman-Yor process, a stochastic process of probability distribution which prefers the reuse of frequently occurring rules versus creating new ones to build a lexicon, then uses that lexicon to parse the input (M. Johnson, Griffiths, & Goldwater, 2007).

Phonotactics from Utterances Determine Distributional Lexical Elements (PUDDLE, Monaghan & Christiansen, 2010) treats each utterance as a lexical item, unless an already stored item is part of this utterance, and the remainders are phonotactically legal. If so, it breaks up the utterance into segments, and the segments would enter the lexicon as new lexical items.

Finally, two baselines were included: Syll=Word treats each syllable as a word and Utt=Word treats each utterance as a word.

Results

We first investigated whether IDS is simpler than ADS in terms of six corpus features that could affect word segmentation, as described in the reasoning above. The results of paired t-tests comparing the registers for each feature are in Table 2, which shows that four out of six features fit our predictions.

We also noticed that IDS size corpus (M=487, SD=350 per child) was significantly larger than the ADS one (M=238, SD=230), based on a t-test with t(17)=2.63, p=0.02. This may mean that these infants were exposed to more IDS than ADS, similar to what Cristia et al. (2018) found for English.

The performance of all segmentation algorithms for both registers is captured in Figure 1. IDS is easier to segment than ADS when points are above the dotted diagonal line. There was a small IDS advantage for most algorithms, although some showed the opposite effect (DiBSs,

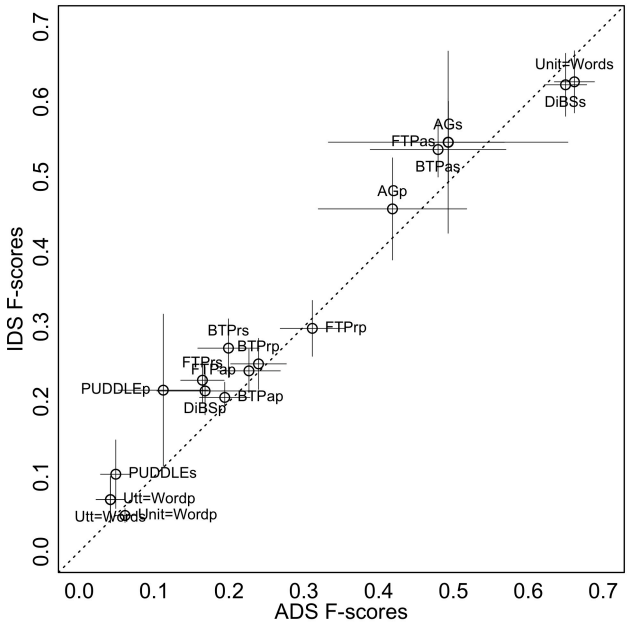

Figure 1: Token F-scores obtained by each algorithm for IDS as function of that for ADS. The final "s" in the model's name means that the basic unit of the corpus was syllables (PUDDLWs, Utt=Words, Unit=Words, DiBSs, FTPas, FTPrs, BTPas, BTPrs, AGs). The final "p" in the model's name means that the basic unit of the corpus was phones (PUDDLWp, Utt=Wordp, Unit=Wordp, DiBSp, FTPap, FTPrp, BTPap, BTPrp, AGp). Error bars show two standard deviations over the 18 corpora.

Unit=Words, Unit=Wordp, FTPrp). We also observe that in many cases the pseudo-confidence intervals cross the diagonal line, suggesting that performance difference is within the range of error. Thus, only FTPrs, BTPrs, Utt=Wordp, PUDDLEp and PUDDLEs showed a clear advantage of IDS. We then tested for overall effects in a linear mixed effect regression model (Bates, Mächler, Bolker, & Walker, 2015) predicting token F-scores from register (IDS or ADS) as a fixed effect, where subject and algorithm (AGs, AGp, DiBSs, DiBSp...) were random effect variables. Register significantly affected token F-scores ($\chi^2(1)=50.87$, $p<.05$, Type II Anova), IDS having a performance advantage of .03 $\pm$ .004 (standard error).

Next, we tested whether this performance advantage was due to one of the above-mentioned corpus properties. To see whether performance differences were due to the artifactual difference in corpus length, we also included the number of utterances as a register feature. Thus, 7 new models, each including one of the features as an additional fixed effect, were fit. We then measured the significance of register and features in the new models with a Type II Anova test (Fox & Weisberg, 2011).

If the advantage of IDS was entirely due to one feature, then register would no longer be significant in these addi-

Table 3: Corpus features predict segmentation scores, but do not replace register. β feat stands for the estimated coefficient of that feature; β rgstr for that of register in the new model (which should be compared to 0.03 at the simple model). p features shows whether feature was significant in new model. p rgst shows whether register remained significant in the new model. N. utts stands for number of utterances.

Feature	Feature		Register	
Feature	β	p	β	p
Word length	.02	.48	.03	***
Phrase length	.01	***	.04	***
% 1-w phrase	.06	.29	.03	***
Entropy	-1.58	***	.03	***
MATTR	.5	***	.05	***
% hapaxes	.03	.18	.03	***
N. utts	.00005	***	.02	***

Table 4: Correlation tests (Spearman) of corpus features and infant age for each register. "coef." stands for correlation coefficient.% 1-w phrase stands for ratio of single word phrases. % hapaxes is the ratio of hapaxes.

Feature	IDS coef.	ADS coef.
Word length	.50*	.06
Phrase length	.34	-.56*
% 1-w phrase	-.37	.12
Entropy	-.50*	.70**
TTR	.44	-.37
% hapaxes	.01	.30

tional analyses. Results (in Table 3) showed that phrase length, segmentation entropy, MATTR, and corpus size accounted for variance in the results, but no single feature rendered register effects non-significant.

Next, we investigated whether IDS features change with infant age, with IDS becoming more ADS like as infants age. Spearman correlation tests between properties and infant age for each register separately (Table 4) did not confirm our predictions: Only word length and entropy (neither of which had emerged as register properties on Table 2) correlated with age in IDS; entropy and phrase length did so for ADS. We have no plausible explanation for these effects.

Two infants were recorded twice at different ages, one at 31 and 38 months, the other at 32 and 40 months. Following a recommendation from a reviewer, we inspected these two infants as case studies. An inspection of IDS features demonstrated that phrase length and % of 1-w phrases were the only features having small changes with age, but only the latter would change in the same direction for both infants, increasing by 6% and 1% from the first to the second recording. A few ADS features also changed slightly with age, such as % of 1-w phrases, word length and entropy, but only phrase length changed in the same direction for both infants, decreasing by 1.18 and 1.66 phonemes.

Finally, we created a new model predicting token F-scores register (IDS or ADS) and infant age in months as fixed effects (and model and participant as random effects, as before), and their interaction. Both main effects and the interaction were significant (Age $\chi^2(1)=4.31$, $p<.05$; Register $\chi^2(1)=53.14$, $p<.5$; Age:register $\chi^2(1)=28.81$, $p<.05$). A follow-up analysis separating the registers indicated that ADS scores decreased by $.002 \pm .0005$ (standard error) with age, whereas there was no significant change with age for IDS.

Discussion

In this modeling study, we assessed whether there are informational differences affecting word segmentation between IDS and ADS drawn from the same ecological corpus. First, we investigated whether this naturalistic corpus had IDS-ADS differences in textual features that would make segmentation easier in the former than the latter. We found most features fit our predictions: Phrases were longer, there were more single-word phrases, lexical diversity was lower, and there were fewer hapaxes in IDS than ADS. No significant effect was found for word length and ambiguity. This result contributes to the growing literature documenting IDS features, with the important advantage that current work draws from fully ecological IDS and ADS.

Next, we investigated the segmentability of the corpora using a large set of both lexical and sublexical segmentation models. Although scores varied a great deal across algorithms and some algorithms showed the opposite effect, IDS was overall slightly easier to segment than ADS. The mean difference across registers (CDS minus ADS, in each algorithm separately) was 3%, ranging from –4% to 10%. This effect is smaller than that found in most previous studies, but similar to the one reported by Cristia et al. (2018), who were also drawing from a naturalistic IDS-ADS corpus. This is evidence that previously documented IDS-ADS segmentability differences (as in Table 1) are not representative of what infants actually hear. It is important to note that corpus length across registers was not matched in the present study for practical reasons, but, based on findings by Cristia et al. (2018), we suspect that controling for corpus size would have reduced the IDS advantage even further.

Next, we asked whether some of the above-mentioned textual features uniquely explained segmentability differences across registers. Phrase length, segmentation entropy, and repetitiveness explained significant variance in segmentation scores, above and beyond the effects of register. However, none of the features uniquely explained away the effect of the register, which remained significant in all cases. This means that register effects on segmentability cannot be reduced to any one of these features. Since we only had 18 children's data, we could not fit a model with all 6 features at once for fear of overfitting, but future work with higher power may be able to assess whether these features jointly explain away reg-

ister, or whether there are other textual features that we have not yet considered.

Furthermore, Canault et al. (2016b)'s corpus allowed us to address a question that has been seldom asked, namely IDS-ADS differences across infant ages. Results of correlations between textual features and age, and a regression model on token F-scores did not support our prediction that IDS would become more like ADS as children aged, and thus the IDS-ADS segmentability gap would close. On the contrary, we found that ADS scores dropped with child age. Although further work is needed, we believe this mainly reflects the lower availability of ADS in children's environment as they age. Indeed, replicating a pattern that had been documented in North American English children (Bergelson et al., 2019), we found the number of ADS utterances dropped for older, compared to younger, children.

Before closing, we would like to acknowledge some limitations of this work. Corpus size was overall small (which may lead to inconsistencies in results; Bernard et al., 2018) and, due to the work involved in collecting daylong recordings and annotating fully spontaneous speech, infant sample size was 18 infants. Moreover, data scarcity was correlated with registers and ages: While only 3 of the 18 IDS corpora contained fewer than 100 utterances, 7 did for ADS, and 4 of those belonged to infants older than 31 months. A decrease of ADS quantities with infant age in such day-long recordings has been documented in previous work on North American English (Bergelson et al., 2019), so it may not be an artifact of the current sample selection. Nonetheless, this trend may entail that if we want to control corpus size, we should over-sample ADS at later ages. However, that may not be necessary for our data, where corpus size failed to explain away the register effect, even though it accounted for some variance beyond registers.

Last, speech transcriptions were used for this study, in an attempt to look for intrinsic informational differences across registers. However, some of the most salient features of IDS are speech-related, such as prosody or intonation and acoustic properties, which might also predict ease of segmentation. Although there is a small literature looking at word segmentation from speech, including comparing IDS and ADS (Ludusan, Seidl, Dupoux, & Cristia, 2015), this task remains extremely challenging for computational modelers, with only one open source model (instantiating a single segmentation strategy) exists, which further limits the value of such a line of research.

In sum, we identified several simplicity features more prevalent in IDS than ADS drawn from an ecological French corpus. We further found a small but significant IDS segmentation advantage, contributing to a recurrent question on the learnability properties of IDS. We showed that the IDS segmentation advantage could not be explained away by any one of those simplicity features, and its size changed with infant age in unexpected directions.

Acknowledgments

References

Batchelder, E. (1997). *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. Unpublished doctoral dissertation, City University of New York.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? a large-scale cross-corpus analysis. *Developmental science*, *22*(1), e12724.

Bernard, M. (2018). Phonemizer [Computer software manual]. https://github.com/bootphon/phonemizer, doi = "http://doi.org/10.5281/zenodo.2537809".

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., ... Cristia, A. (2018). Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016a). *Lyon homebank corpus.* doi: 21415/T58P6Q

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016b). Reliability of the language environment analysis system (lena^{TM}) in european french. *Behavior research methods*, *48*(3), 1109–1124.

Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2018). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, 1–10.

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

Duddington, J. (2012). espeak text to speech [Computer software manual].

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Why is english so easy to segment? In *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (cmcl)* (pp. 1–10).

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage.

Henning, A., Striano, T., & Lieven, E. V. (2005). Maternal speech to infants at 1 and 3 months of age. *Infant behavior and development*, *28*(4), 519–536.

Johnson, E., Lahey, M., Ernestus, M., & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *The Journal of the Acoustical Society of America*, *134*(6), EL534–EL540.

Johnson, E., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS one*, *9*(1), e83546.

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).

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)* (Vol. 2, pp. 178–183).

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).

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

MacWhinney, B. (1996). The childes system. *American Journal of Speech-Language Pathology*, *5*(1), 5–14.

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.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, id rather do it myself: Some effects and non-effects of maternal speech style.

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.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science*, *20*(3), e12390.

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

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142).

10. Conclusions

With these two studies, we attempted to contribute to the literature of learnability across cultures. We used input from ecological, longitudinal recordings of diverse cultures, a Sesotho-speaking community in Lesotho and French-speaking communities in France, for corpus analysis in Chapter 8, with several confounds controlled for. As we mentioned above for cross-linguistic studies, we believe that cross-cultural comparison in future studies may ideally happen under these terms, in order to identify the main source(s) of difference between languages. We also modelled the segmentability of child- and adult-directed speech in a French-speaking community in Chapter 9.

There were significant differences in morphosyntactic and lexical features of child- and adult-directed speech, in both French and Sesotho cultures. However, two specific points drew our attention. First, there were no flagrant differences in features that could affect segmentation performance between the two registers. This might mean that, even though CDS and ADS differ, they don't differ in aspects that necessarily affect segmentability. This is also evidence that everyday adult-directed speech, like the one spoken between parents at home, may be more similar to child-directed speech than any other kind of adult-directed speech (including talk to the investigator, or in the lab).

Second, specifically in Chapter 9, the features we identified did not explain away the (small) effect of register in segmentation. One simple interpretation of the results is that register diversity, as expressed by these characteristics, is simply not a factor for segmentation. Future research should further compare data of the two registers, collected in ecological settings in order to capture real-life differences, and should consider including more features responsible for diversity, such as phonological, semantic and syntactic ones, as well as the interactions between them.

Furthermore, in Chapter 8, we compared child-directed, adult-directed, but also overheard child-directed speech, that is, speech addressed to other children in the environment. Interestingly, overheard child-directed speech is similar to child-directed speech. This result invites further research. We may hypothesize that children are more interested in speech addressed to other children, than speech addressed to adults. If this speech is also similar to speech directed to them, do children benefit from overheard other-child-directed speech more than they do for overheard adult-directed speech?

Last, in the same Chapter, input recorded in WEIRD settings of French families was mainly produced by the nuclear family, and mostly by the mother. In contrast, input recorded in non-WEIRD settings was produced by the extended family, including grandmothers, neighbors and other adults. However, the majority of input originated from other children. We further analyzed these results, comparing child-directed speech produced by other adults and other children. The two kinds of speech were also similar across a large set of variables, except for the ratio of questions. In sum, these results shed light on the vast differences in the settings of language acquisition between cultures, and invite for more comprehensive, large-scale, ecological research.

Discussion

In this overarching discussion, I will talk about some personal thoughts, future research and limitations of my studies. If we wanted to only remember a phrase from this dissertation, it would be that yes, diversity is huge, but also learning appears to be robust across most of it. When I started working with these hugely diverse input data, I expected to find much more difference both across their corpus features and across their segmentability, than what I actually found.

While working across languages and cultures, I observed the benefits of large-scale studies. Studying language acquisition at scale is a promising way to capture variability in input, measure which mechanisms and factors account for word learning, how they interact, and how much learning they explain overall. The field of language acquisition needs more quantifiable and comparable methods to study the input and its environment. As technological development in corpus based linguistics grows, so will our chances to use ecological data, in order to address acquisition and learnability issues.

Similarly, the field of language acquisition needs more data. While discussing previous studies, sometimes observations on production were my main source of information. However, production correlates with external factors, it is culturally dependent, and depends on the phonetic structure. More tests of comprehension that collect comparable data across languages and cultures may be developed and implemented, because comprehension data should be more informative on early acquisition steps, such as the one discussed here, segmentation of speech.

In future studies, I am planning to use an existing rich source of data, the MacArthur-Bates Communicative Development Inventories (CDIs) aggregated within the Wordbank Project (Frank et al., 2017). CDIs are parental reports on their children's lexical development, proven to be reliable indicators of a child's language, and they may provide valuable information on early language

comprehension. I will use these data to study the uptake of language input by children, investigating its learning factors and its computational mechanisms.

Moreover, at the beginning of this dissertation, I defined input as ambient language a child is exposed to. It is important, though, to keep in mind for future studies that language acquisition varies over time, and that input is actually interactions between children and caregiver(s), with varying communicative and coordinating goals. The input children receive is characterized by conversational dynamics, and children learn words in communicative contexts, *interacting* with speakers. Pan et al. (2005) pointed out that communication with young children is "a total package of verbal and nonverbal, linguistic and emotional interaction" (p.778). By talking about interactions, we need to accept a social-pragmatic account of acquisition (Baldwin & Meyer, 2007; Diesendruck, 2007).

Previous literature on linguistic development has rarely taken the conversational dynamics of *real-life* interactions into account, since it mostly consists in data gathered in small, controlled laboratory studies. However, since learning emerges through interactions, some learning situations may be better than others, including situations with engaging observational and social context (e.g. entities and actions observable in the scene) (He & Arunachalam, 2017).

In future studies, I am thus planning to analyze the range of interactions children participate in using large-scale, ecological studies, and to quantify whether different everyday situations (eating, playing...) have similar engaging features and predictive power in acquisition. Specifically in the domain of conversational analysis, the field of natural language processing has improved dramatically, with important industrial applications (e.g., chatbots, personal assistants). I will use these tools to analyse the nature of interactions in corpora of early child-caregiver interactions.

Furthermore, in this dissertation, I used a set of segmentation models, most of which are based on simple statistics and batch processing, with no prior biases in learning. The fact that such models manage to meet the evaluation criteria of the first chapters is actually a good sign. I suspect that a model equipped with lexical constraints (as children have from early on in development, Markman et al., 2003), parsimony bias (Frank et al., 2010) and use of more than one segmentation cue (children use several cues for segmentation from early on in development (Mersad & Nazzi, 2012), could yield results closer to what is first segmented by children. Attention and memory in children's learning are also not modeled, but probably play a role in learning. However, 'simpler' models, such as the ones implemented in Chapters 3, 4 and 8, nonetheless enabled me to address my primary question on the informativity of input across environments.

In future studies, I am planning to build a comprehensive list of models with different implementation strategies, such as learning incrementally by exposure to one utterance after another (incremental), instead of processing all information at once (batch), performing single or joint tasks and exhibiting memory limitations. I will assess the viability of these models in a comparative way, employing the plausibility tests introduced in Chapter 4.

Additionally, learnability is an issue that concerns not only segmentation, but also several other aspects of language. We could talk about learnability of syntax, semantics, or morphology. In future studies, I am planning to conduct a more comprehensive research on learnability of different aspects. It may be the case that some aspects are more affected by the input than others. For example, vocabulary could be driven by input (the particular words a child is exposed to) more than other aspects of language, which rely more on cognitive factors.

Before closing, I would like to mention two sources of diversity that were not presented in this dissertation. The first one is diversity across children. No two children learn language in an identical way (e.g. Brown, 1973; Pizzuto & Caselli, 1992). This should be taken into consideration for studies where only a handful of children were targeted for each language or culture (such as the ones described in this dissertation), and it can be framed in large-scale studies, with a considerable number of target children. The second one is multilingualism. Half of the world's children live in multilingual environments (Cenoz & Genesee, 1998). Children in multilingual communities have to learn -at least- two different languages and also appropriate code-switching between them (Loakes et al., 2013). For example, Indigenous Australian children hear and learn at least 3 languages: traditional languages, Kriol or Aboriginal English, and some level of Standard Australian English).

Last, I believe that understanding how diversity affects learning is crucial for many reasons. First, it is necessary if we want to grasp broader issues in acquisition. For example, once this diversity is taken into account, what is left can be studied for questions on cognitive biases and innateness. Second, understanding the relation between diversity and learning can contribute to promoting healthy learning environments, and productive methods of learning for all children (there is already some effort to build interventions based on previous studies, e.g. Wong et al., 2020). This knowledge can have important implications on several areas (education, parenting, psychology and even artificial intelligence -such as virtual learning companions for children).

Bibliography

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209. https://doi.org/10.1111/j.1467-7687.2005.00406.x

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning Words through Overhearing. *Child Development*, *72*(2), 416–430. https://doi.org/10.1111/1467-8624.00287

Allen, S., & Crago, M. (1996). Early passive acquisition in Inuktitut. *Journal of Child Language*, *23*(1), 129-155.

Ambridge, B., & Lieven, E. (2011). *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge University Press.

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

Bakermans-Kranenburg, M. J., IJzendoorn, M. H. van, & Kroonenberg, P. M. (2004). Differences in attachment security between African-American and white children: Ethnicity or socio-economic status? *Infant Behavior and Development*, *27*(3), 417–433. https://doi.org/10.1016/j.infbeh.2004.02.002

Baldwin, D., & Meyer, M. (2007). How inherently social is language? In *Blackwell handbook of language development* (pp. 87–106). Blackwell Publishing. https://doi.org/10.1002/9780470757833.ch5

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, *19*(3), 241–248.

Barnes, S., Gutfreund, M., Satterly, D., & Wells, G. (1983). Characteristics of adult speech which predict children's language development. *Journal of Child Language*, *10*(1), 65–84. https://doi.org/10.1017/S0305000900005146

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167–206. https://doi.org/10.1016/S0010-0277(02)00002-1

Bates, E., & MacWhinney, B. (1987). Competition, Variation and Language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 157–193). Lawrence Erlbaum.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397. https://doi.org/10.1016/j.cognition.2013.02.011

Berko, J. (1958). The Child's Learning of English Morphology. WORD, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Bertolo, S. (Ed.). (2001). *Language acquisition and learnability*. Cambridge University Press.

Bickel, B. (2014). Linguistic diversity and universals. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The Cambridge Handbook of Linguistic Anthropology* (pp. 102–127). Cambridge University Press. https://doi.org/10.1017/CBO9781139342872.006

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721. https://doi.org/10.1037/0012-1649.29.4.711

Billman, D. (2007). *Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories*. *Language and Cognitive Processes*, *4*(2), 127-155.

Black, A., & Bergmann, C. (2017). *Quantifying Infants' Statistical Word Segmentation: A Meta-Analysis*. In *39th Annual Meeting of the Cognitive Science Society* (pp. 124-129). Cognitive Science Society.

Blanchard, D., Heinz, J., & Golinkoff, R. (2009). *Modeling the contribution of phonotactic cues to the problem of* word segmentation. *Journal of child language*, *37*(3), 487-511.

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *37*(3), 487–511. https://doi.org/10.1017/S030500090999050X

Bleses, D., Vach, W., Slott, M., Wehberg, S., & Thomsen, P. (2008). *Early vocabulary development in Danish and other languages:* A CDI-based comparison. *Journal of child language*, *35*(3), 619-650.

Bloom, P. (2004). Myths of Word Learning. In *Weaving a lexicon* (pp. 205–224). MIT Press.

Booth, A. E., & Waxman, S. R. (2009). A Horse of a Different Color: Specifying With Precision Infants' Mappings of Novel Nouns and Adjectives. *Child Development*, *80*(1), 15–22. https://doi.org/10.1111/j.1467-8624.2008.01242.x

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pecheux, M.-G., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-Linguistic Analysis of Vocabulary in Young Children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development*, *75*(4), 1115–1139. https://doi.org/10.1111/j.1467-8624.2004.00729.x

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

Brown, P., & Gaskins, S. (2014). Language acquisition and language socialization. In N. J. Enfield, P. Kockelman, & J. Sidnell (Eds.), *The Cambridge Handbook of Linguistic Anthropology* (pp. 187–226). Cambridge University Press. https://doi.org/10.1017/CBO9781139342872.010

Brown, R. (1973). *A first language: The early stages* (pp. xx, 437). Harvard U. Press.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455. https://doi.org/10.1080/01690969508407111

Callanan, M. A., & Sabbagh, M. A. (2004). Multiple Labels for Objects in Conversations With Young Children: Parents' Language and Children's Developing Expectations About Word Meanings. *Developmental Psychology*, *40*(5), 746–763. https://doi.org/10.1037/0012-1649.40.5.746

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Casillas, M., Brown, P., & Levinson, S. C. (2019). Early Language Experience in a Tseltal Mayan Village. *Child Development*. https://doi.org/10.1111/cdev.13349

Cenoz, J., & Genesee, F. (1998). *Beyond Bilingualism: Multilingualism and Multilingual Education*. Multilingual Matters.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2002). *Brief article Infants learn phonotactic regularities from brief auditory experience*.*Cognition*, *83*(1), B13-B23.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77. https://doi.org/10.1016/s0010-0277(02)00233-0

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language*, *22*(3), 497–529. https://doi.org/10.1017/S0305000900009934

Chomsky, N. (1959). A review of BF Skinner's Verbal Behavior. *Language*, *35*(1), 26–58.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(5), 489–509. https://doi.org/10.1017/S0140525X08004998

Christophe, A., Guasti, T., Nespor, M., Dupoux, E., & Van Ooyen, B. (1997). Reflections on Phonological Bootstrapping: Its Role for Lexical and Syntactic Acquisition. *Language and Cognitive Processes*, *12*(5–6), 585–612. https://doi.org/10.1080/016909697386637

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008a). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, *51*(1–2), 61–75. https://doi.org/10.1177/00238309080510010501

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008b). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, *51*(1–2), 61–75. https://doi.org/10.1177/00238309080510010501

Clark, E. (2017). Morphology in language acquisition. *The handbook of morphology*, 374-389.

Clark, E V, & Hecht, B. F. (1983). Comprehension, Production, and Language Acquisition. *Annual Review of Psychology*, *34*(1), 325–349. https://doi.org/10.1146/annurev.ps.34.020183.001545

Clark, Eve V. (2017). Morphology in Language Acquisition. In *The Handbook of Morphology* (pp. 374–389). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405166348.ch19

Connelly, M. (1984). *Basotho children's acquisition of noun morphology* (Doctoral dissertation, University of Essex)

Conway, C. M., Baurnschmidt, A., Huang, S., & Pisoni, D. B. (2010). Implicit Statistical Learning in Language Processing: Word Predictability is the Key. *Cognition*, *114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Crago, M., & Allen, S. (1998). Acquiring Inuktitut. In *Language acquisition across North America: Cross-cultural and cross-linguistic perspectives* (pp. 245-279). Singular Publishing Group, Inc.

Crago, M., Annahatak, B., & Ningiuruvik, L. (1993). Changing Patterns of Language Socialization in Inuit Homes. *Anthropology & Education Quarterly*, *24*(3), 205–223. https://doi.org/10.1525/aeq.1993.24.3.05x0966d

Craig, H. K., & Washington, J. A. (2004). Grade-Related Changes in the Production of African American English. *Journal of Speech, Language, and Hearing Research*, *47*(2), 450–463. https://doi.org/10.1044/1092-4388(2004/036)

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

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, *2*(3–4), 133–142. https://doi.org/10.1016/0885-2308(87)90004-0

Demuth, K. (1988). Noun classes and agreement in Sesotho acquisition. *Agreement in Natural Language: Approaches, Theories and Descriptions*, 305–321.

Demuth, K. (1990). Subject, topic and Sesotho passive. *Journal of Child Language*, *17*(1), 67-84.

Demuth, K. (1992). *Acquisition of Sesotho.* (Vol. 3, pp. 557–638). Lawrence Erlbaum Associates.

Demuth, K. (1998). Argument structure and the acquisition of Sesotho applicatives. *Linguistics*, *36*(4). https://doi.org/10.1515/ling.1998.36.4.781

Depaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2014). When do infants begin recognizing familiar words in sentences? *Journal of Child Language*, *41*(1), 226–239. https://doi.org/10.1017/S0305000912000566

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A

crosslinguistic study of the relationship between grammar and lexical development. *Journal of Child Language*, *32*(4), 759–786. https://doi.org/10.1017/S0305000905007105

Diesendruck, G. (2007). Mechanisms of Word Learning. In Erika Hoff & M. Shatz (Eds.), *Blackwell Handbook of Language Development* (pp. 257–276). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757833.ch13

Dilworth-Anderson, P. (1992). Extended Kin Networks in Black Families. *Generations: Journal of the American Society on Aging*, *16*(3), 29–32. JSTOR.

Dixon, R. M. W., & Aĭkhenval′d, A. I. (Eds.). (2004). *Adjective classes: A cross-linguistic typology*. Oxford University Press.

Dressler, W. U., Lettner, L. E., & Korecky-Kröll, K. (2010). First language acquisition of compounds. *Cross-Disciplinary Issues in Compounding*, 323–344.

Dunn, J., & Shatz, M. (1989). Becoming a Conversationalist despite (Or Because of) Having an Older Sibling. *Child Development*, *60*(2), 399–410. JSTOR. https://doi.org/10.2307/1130985

Elman, J. L. (Ed.). (1996). *Rethinking innateness: A connectionist perspective on development*. MIT Press.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199.

Endress, A. D., Nespor, M., & Mehler, J. (2009). Perceptual and memory constraints on language acquisition. *Trends in Cognitive Sciences*, *13*(8), 348–353. https://doi.org/10.1016/j.tics.2009.05.005

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can Infants Map Meaning to Newly Segmented Words?: Statistical Segmentation and Word Learning. *Psychological Science*, *18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Evans, N., & Levinson, S. C. (2009a). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Everett, D. L. (2005). Cultural Constraints on Grammar and Cognition in Pirahã: Another Look at the Design Features of Human Language. *Current Anthropology*, *46*(4), 621–646. https://doi.org/10.1086/431525

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. JSTOR. https://doi.org/10.2307/1166093

Fernald, A., & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*

(pp. 365–388). Lawrence Erlbaum Associates, Inc.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(03), 477–501. https://doi.org/10.1017/S0305000900010679

Finn, A. S., Lee, T., Kraus, A., & Kam, C. L. H. (2014). When It Hurts (and Helps) to Try: The Role of Effort in Language Learning. *PLOS ONE*, *9*(7), e101806. https://doi.org/10.1371/journal.pone.0101806

Fortescue, M. (1984). Learning to speak Greenlandic: A case study of a two-year-old's morphology in a polysynthetic language. *First Language*, *5*(14), 101–112. https://doi.org/10.1177/014272378400501402

Fortescue, M., & Olsen, L. L. (1992). The Acquisition of West Greenlandic. *Crosslinquistic Study of Language Acquistion*, 111–219.

Fourtassi, A., & Dupoux, E. (2014). A Rudimentary Lexicon and Semantics Help Bootstrap Phoneme Acquisition. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 191–200. https://doi.org/10.3115/v1/W14-1620

Fouts, H. N., Roopnarine, J. L., Lamb, M. E., & Evans, M. (2012). Infant Social Interactions With Multiple Caregivers: The Importance of Ethnicity and Socioeconomic Status. *Journal of Cross-Cultural Psychology*, *43*(2), 328–348. https://doi.org/10.1177/0022022110388564

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

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

Frank, Michael C., Tily, H. J., Arnon, I., & Goldwater, S. J. (2010). Beyond Transitional Probabilities: Human Learners Impose a Parsimony Bias in Statistical Word Segmentation. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 760–765.

Gambell, T., & Yang, C. (2005). Mechanisms and constraints in word segmentation. *Unpublished Manuscript, Yale University*. http://ling.umd.edu/~jlidz/Teaching/SP06Seminar/gambleyang05.pdf

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

Gelman, S. A., & Taylor, M. (1984). How Two-Year-Old Children Interpret Proper and Common Names for Unfamiliar Objects. *Child Development*, *55*(4), 1535.

https://doi.org/10.2307/1130023

Genovese, G., Spinelli, M., Romero Lauro, L. J., Aureli, T., Castelletti, G., & Fasolo, M. (2020). Infant-directed speech as a simplified but not simple register: A longitudinal study of lexical and syntactic features. *Journal of Child Language*, *47*(1), 22–44. https://doi.org/10.1017/S0305000919000643

Gerken, LouAnn. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67–B74. https://doi.org/10.1016/j.cognition.2005.03.003

Gerken, Louann, Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249–268. https://doi.org/10.1017/S0305000904006786

Gervain, J., & Mehler, J. (2010). Speech Perception and Language Acquisition in the First Year of Life. *Annual Review of Psychology*, *61*(1), 191–218. https://doi.org/10.1146/annurev.psych.093008.100408

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

Gierut, J. A. (2007). Phonological Complexity and Language Learnability. *American Journal of Speech-Language Pathology*, *16*(1), 6–17. https://doi.org/10.1044/1058-0360(2007/003)

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55.

Goldfield, B. A. (1993). Noun bias in maternal speech to one-year-olds. *Journal of Child Language*, *20*(1), 85–99. https://doi.org/10.1017/S0305000900009132

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

Gómez, R. L. (2002). Variability and Detection of Invariant Structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/10.1111/1467-9280.00476

Gomez, R. L., & Gerken, L. (1999a). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Gomez, R. L., & Gerken, L. (1999b). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old

infants. *Developmental Science*, *7*(5), 567–580. https://doi.org/10.1111/j.1467-7687.2004.00381.x

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language; Cambridge*, *35*(3), 515–531. http://dx.doi.org/10.1017/S0305000907008641

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*(3), 281–287. https://doi.org/10.1111/j.1467-7687.2007.00584.x

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*(4), 548–567. https://doi.org/10.1016/j.jml.2004.07.002

Hallé, P. A., Durand, C., & de Boysson-Bardies, B. (2008). Do 11-month-old French Infants Process Articles? *Language and Speech*, *51*(1–2), 23–44. https://doi.org/10.1177/00238309080510010301

Hamilton, A. (1981). *Nature and nurture: Aboriginal child-rearing in north-central Arnhem Land* (No. 20). Australian Institute of Aboriginal Studies

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children* (pp. xxiii, 268). Paul H Brookes Publishing.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64. https://doi.org/10.1016/S0010-0277(00)00132-3

Hayes, B., & Wilson, C. (2008).A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379-440.

He, A. X., & Arunachalam, S. (2017). Word learning mechanisms: Word learning mechanisms. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(4), e1435. https://doi.org/10.1002/wcs.1435

Hengeveld, K. (1992). *Parts of speech*. Layered structure and reference in a functional perspective, 29-55.

Hoff, E. (2006). How social contexts support and shape language development☆. *Developmental Review*, *26*(1), 55–88. https://doi.org/10.1016/j.dr.2005.11.002

Hoff, Erika, & Naigles, L. (2002). How Children Use Input to Acquire a Lexicon. *Child Development*, *73*(2), 418–433. https://doi.org/10.1111/1467-8624.00415

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology*, *22*(2), 155–163. https://doi.org/10.1037/0012-1649.22.2.155

Hoff-Ginsberg, E. (1991). Mother-Child Conversation in Different Social Classes and Communicative

Settings. *Child Development*, *62*(4), 782. https://doi.org/10.2307/1131177

Hoff-Ginsberg, E. (1998). The relation of birth order and socioeconomic status to children's language experience and language development. *Applied Psycholinguistics*, *19*(4), 603–629. https://doi.org/10.1017/S0142716400010389

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional Elements in Infants' Speech Processing: The Role of Determiners in the Syntactic Categorization of Lexical Elements. *Infancy*, *5*(3), 341–353. https://doi.org/10.1207/s15327078in0503_5

Hsieh, L., Leonard, & Swanson. (1999). Some differences between English plural noun inflections and third singular verb inflections in the input: The contributions of frequency, sentence position, and duration. *Journal of Child Language*, *26*(3), 531-543.

Hunter, M. A., Ames, E. W., & Koopman, R. (1983). Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli. *Developmental Psychology*, *19*(3), 338–352. https://doi.org/10.1037/0012-1649.19.3.338

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Jaeger, T. F., & Norcliffe, E. J. (2009). The Cross-linguistic Study of Sentence Production. *Language and Linguistics Compass*, *3*(4), 866–887. https://doi.org/10.1111/j.1749-818X.2009.00147.x

Jiang, H., Frank, M. C., Kulkarni, V., & Fourtassi, A. (2020). *Exploring patterns of stability and change in caregivers' word usage across early childhood* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fym86

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*(4), 548–567. https://doi.org/10.1006/jmla.2000.2755

Johnson, E. K., & Tyler, M. D. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345. https://doi.org/10.1111/j.1467-7687.2009.00886.x

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Junge, C. (2018). The proto-lexicon: Segmenting word-like units from the speech stream. In *Early word learning* (pp. 15–29). Routledge/Taylor & Francis Group.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Jusczyk, Peter W., & Hohne, E. A. (1997). Infants' Memory for Spoken Words. *Science*, *277*(5334), 1984–1986. https://doi.org/10.1126/science.277.5334.1984

Jusczyk, Peter W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, *61*(8), 1465–1476. https://doi.org/10.3758/BF03213111

Jusczyk, Peter 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

Kelly, B., Wigglesworth, G., Nordlinger, R., & Blythe, J. (2014). The Acquisition of Polysynthetic Languages: The Acquisition of Polysynthetic Languages. *Language and Linguistics Compass*, *8*(2), 51–64. https://doi.org/10.1111/lnc3.12062

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS One*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453. https://doi.org/10.1016/j.cognition.2013.02.002

Kusters, W., & Muysken, P. C. (2001). The complexities of arguing about complexity. *Linguistic Typology*, 182–185.

Ladányi, E., Kovács, Á. M., & Gervain, J. (2020). How 15‐month‐old infants process morphologically complex forms in an agglutinative language? *Infancy*, *25*(2), 190–204. https://doi.org/10.1111/infa.12324

Lany, J., & Gómez, R. L. (2008). Twelve-Month-Old Infants Benefit From Prior Experience in Statistical Learning. *Psychological Science*, *19*(12), 1247–1252. https://doi.org/10.1111/j.1467-9280.2008.02233.x

Lawrence, V. W., & Shipley, E. F. (1996). Parental speech to middle- and working-class children from two racial groups in three settings. *Applied Psycholinguistics*, *17*(2), 233–255. https://doi.org/10.1017/S0142716400007657

Leffel, K. R., & Suskind, D. L. (2013). Parent-directed approaches to enrich the early language environments of children living in poverty. *Seminars in Speech and Language*. https://doi.org/10.1055/s-0033-1353443

León, L. D. (2011). Language Socialization and Multiparty Participation Frameworks. In *The Handbook of Language Socialization* (pp. 81–111). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444342901.ch4

Levinson, S. C. (2012). The Original Sin of Cognitive Science. *Topics in Cognitive Science*, *4*(3), 396–403. https://doi.org/10.1111/j.1756-8765.2012.01195.x

Levy, Y. (1983). The acquisition of Hebrew plurals: The case of the missing gender category. *Journal of Child Language*, *10*(1), 107–121. https://doi.org/10.1017/S0305000900005171

Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In *Input and interaction in language acquisition* (pp. 56–73). Cambridge University Press. https://doi.org/10.1017/CBO9780511620690.005

Liszkowski, U., Brown, P., Callaghan, T., Takada, A., & Vos, C. de. (2012). A Prelinguistic Gestural Universal of Human Communication. *Cognitive Science*, *36*(4), 698–713. https://doi.org/10.1111/j.1551-6709.2011.01228.x

Loakes, D., Moses, K., Wigglesworth, G., Simpson, J., & Billington, R. (2013). Children's language input: A study of a remote multilingual Indigenous Australian community. *Multilingua*, *32*(5). https://doi.org/10.1515/multi-2013-0032

Low, S. M., & Stocker, C. (2005). Family Functioning and Children's Adjustment: Associations Among Parents' Depressed Mood, Marital Hostility, Parent-Child Hostility, and Children's Adjustment. *Journal of Family Psychology*, *19*(3), 394–403. https://doi.org/10.1037/0893-3200.19.3.394

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*. 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. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 93–102. https://doi.org/10.18653/v1/W15-2413

Macwhinney, B. (1976). Hungarian research on the acquisition of morphology and syntax. *Journal of Child Language*, *3*(3), 397–410. https://doi.org/10.1017/S0305000900007261

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' Recognition of the Sound Patterns of Their Own Names. *Psychological Science*, *6*(5), 314–317. https://doi.org/10.1111/j.1467-9280.1995.tb00517.x

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275. https://doi.org/10.1016/S0010-0285(03)00034-3

Mattys, S. L., & Jusczyk, P. W. (2001). *Phonotactic cues for segmentation of fluent speech by infants*. *Cognition*, *78*(9), 1-1.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, 477–500.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, *5*(2–3). https://doi.org/10.1515/lity.2001.001

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*, 303–315.

Miestamo, M., Sinnemäki, K., & Karlsson, F. (Eds.). (2008). *Language complexity: Typology, contact, change*. John Benjamins Pub. Co.

Mintz, T. H. (2013). The Segmentation of Sub-Lexical Morphemes in English-Learning 15-Month-Olds. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00024

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, *26*(4), 393–424. https://doi.org/10.1207/s15516709cog2604_1

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and Diversity: Simulating Early Word Learning Environments. *Cognitive Science*, *42*, 375–412. https://doi.org/10.1111/cogs.12592

Morgan, J. L., & Demuth, K. (1996). Signal to syntax: An overview. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, 1–22.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *20*(1), 67–85. https://doi.org/10.1016/S0022-5371(81)90312-1

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*(1), 95–120. https://doi.org/10.1017/S0305000997003358

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

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Norcliffe, E., Harris, A. C., & Jaeger, T. F. (2015). Cross-linguistic psycholinguistics and its critical role in theory development: Early beginnings and recent advances. *Language, Cognition and Neuroscience*, *30*(9), 1009–1032. https://doi.org/10.1080/23273798.2015.1080373

Ochs, E., & Schieffelin, B. B. (1984). Language acquisition and socialization. In *Culture theory: Essays on mind, self, and emotion* (pp. 276–320).

Ogura, T., Dale, P. S., Yamashita, Y., Murase, T., & Mahieu, A. (2006). The use of nouns and verbs by Japanese children and their caregivers in book-reading and toy-playing contexts. *Journal of Child Language*, *33*(1), 1–29. https://doi.org/10.1017/S0305000905007270

Oshima-Takane, Y., & Robbins, M. (2003). Linguistic Environment of Secondborn Children. *First Language*, *23*(1), 21–40. https://doi.org/10.1177/0142723703023001002

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal Correlates of Growth in Toddler Vocabulary Production in Low-Income Families. *Child Development*, *76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Pelucchi, B., Hay, J., & Saffran, J. (2009). *Statistical Learning in a Natural Language by 8‐Month‐Old Infants*. *Child development*, *80*(3), 674-685.

Penke, D. M. (2012). *On the Acquisition of inflectional morphology*. 33.

Peters, A. M., & Menn, L. (1993). False Starts and Filler Syllables: Ways to Learn Grammatical Morphemes. *Language*, *69*(4), 742–777.

Pfeiler, B. (2003). Early acquisition of the verbal complex in Yucatec Maya. In D. Bittner, W. U. Dressler, & M. Kilani-Schoch (Eds.), *Development of Verb Inflection in First Language Acquisition*. De Gruyter Mouton. https://doi.org/10.1515/9783110899832.379

Pierrehumbert, J. B. (2003a). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech*, *46*(2–3), 115–154. https://doi.org/10.1177/00238309030460020501

Pierrehumbert, J. B. (2003b). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech*, *46*(2–3), 115–154. https://doi.org/10.1177/00238309030460020501

Pizzuto, E., & Caselli, M. C. (1992). The acquisition of Italian morphology: Implications for models of language development. *Journal of Child Language*, *19*(3), 491–557. https://doi.org/10.1017/S0305000900011557

Plunkett, K. (1993). Lexical segmentation and vocabulary growth in early language acquisition. *Journal of Child Language*, *20*, 43–60. https://doi.org/10.1017/S0305000900009119

Pye, C. (1986). Quiché Mayan speech to children. *Journal of Child Language*, *13*(1), 85–100. https://doi.org/10.1017/S0305000900000313

Roberts, M. Y., & Kaiser, A. P. (2011). The Effectiveness of Parent-Implemented Language Interventions: A Meta-Analysis. *American Journal of Speech-Language Pathology*, *20*(3), 180–199. https://doi.org/10.1044/1058-0360(2011/10-0055)

Rogoff, B. (2003). *The Cultural Nature of Human Development*. Oxford University Press.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cognitive Science*, *1*(6), 906–914. https://doi.org/10.1002/wcs.78

Rothbaum, F., Weisz, J., Pott, M., Miyake, K., & Morelli, G. (2000). Attachment and culture: Security in the United States and Japan. *American Psychologist*, *55*(10), 1093–1104.

https://doi.org/10.1037/0003-066X.55.10.1093

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

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

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology*, *39*(3), 484–494. https://doi.org/10.1037/0012-1649.39.3.484

Sakai, K. L. (2005). Language Acquisition and Brain Development. *Science*, *310*(5749), 815–819. https://doi.org/10.1126/science.1113530

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, *9*(6), 565–573. https://doi.org/10.1111/j.1467-7687.2006.00534.x

Shi, R., Cutler, A., Werker, J., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America*, *119*(6), EL61–EL67. https://doi.org/10.1121/1.2198947

Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, *11*(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language*, *25*(1), 169–201. https://doi.org/10.1017/S0305000997003395

Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and Representation of Function Words in English-Learning Infants. *Infancy*, *10*(2), 187–198. https://doi.org/10.1207/s15327078in1002_5

Shneidman, L. A., & Goldin‐Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, *10*(1), 1–40. https://doi.org/10.1515/LINGTY.2006.001

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, *108*(15), 6038–6043. https://doi.org/10.1073/pnas.1017617108

Slobin, D. I. (1967). *A field manual for cross-cultural study of the acquisition of communicative competence.* ERIC Clearinghouse.

Snow, C. E. (1972). Mothers' Speech to Children Learning Language. *Child Development*, *43*(2), 549. https://doi.org/10.2307/1127555

Snow, C. E. (1977). Mothers' speech research: From input to interaction. *Talking to children: Language input and acquisition*, *3149*.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M., & Wittebolle, K. (2013). When Do Caregivers Talk? The Influences of Activity and Time of Day on Caregiver Speech and Child Vocalizations in Two Childcare Environments. *PloS one, 8(11)*.

Stephany, U., & Voeikova, M. D. (2009). *Development of Nominal Inflection in First Language Acquisition: A Cross-Linguistic Perspective*. Walter de Gruyter.

Stoll, S., & Bickel, B. (2013). Capturing diversity in language acquisition research. In B. Bickel, L. A. Grenoble, D. A. Peterson, & A. Timberlake (Eds.), *Typological Studies in Language* (Vol. 104, pp. 195–216). John Benjamins Publishing Company. https://doi.org/10.1075/tsl.104.08slo

Stoll, S., Bickel, B., Lieven, E., Paudyal, N. P., Banjade, G., Bhatta, T. N., Gaenszle, M., Pettigrew, J., Rai, I. P., Rai, M., & Rai, N. K. (2012). Nouns and verbs in Chintang: Children's usage and surrounding adult speech. *Journal of Child Language*, *39*(2), 284–321. https://doi.org/10.1017/S0305000911000080

Stoll, S., Bickel, B., & Mažara, J. (2017). *The Acquisition of Polysynthetic Verb Forms in Chintang* (M. Fortescue, M. Mithun, & N. Evans, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199683208.013.28

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

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*(2), 454–464. https://doi.org/10.1037/0012-1649.43.2.454

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3617–3632. https://doi.org/10.1098/rstb.2009.0107

Swingley, D., & Aslin, R. N. (2002). Lexical Neighborhoods and the Word-Form Representations of 14-Month-Olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development*, *89*(4), 1247–1267.

https://doi.org/10.1111/cdev.12731

Tardif, T., Gelman, S. A., & Xu, F. (1999). Putting the "Noun Bias" in Context: A Comparison of English and Mandarin. *Child Development*, *70*(3), 620–635. https://doi.org/10.1111/1467-8624.00045

Tatsumi, T., Ambridge, B., & Pine, J. M. (2018). Testing an input-based account of children's errors with inflectional morphology: An elicited production study of Japanese. *Journal of Child Language*, *45*(5), 1144–1173. https://doi.org/10.1017/S0305000918000107

Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, *10*(1), 21. https://doi.org/10.1186/1471-2202-10-21

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

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

Thiessen, E. D., & Saffran, J. R. (2009). How the Melody Facilitates the Message and Vice Versa in Infant Learning and Memory. *Annals of the New York Academy of Sciences*, *1169*(1), 225–233. https://doi.org/10.1111/j.1749-6632.2009.04547.x

Thompson, S. P., & Newport, E. L. (2007). Statistical Learning of Syntax: The Role of Transitional Probability. *Language Learning and Development*, *3*(1), 1–42. https://doi.org/10.1080/15475440709336999

Toda, S., Fogel, A., & Kawai, M. (2009). Maternal speech to three-month-old infants in the United States and Japan *Journal of Child Language*, *17*(2), 279-294.

Tomasello, M. (2003). *Constructing a language: A usage-based approach to child language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, Michael. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1–2). https://doi.org/10.1515/cogl.2001.012

Tomasello, Michael, & Mannle, S. (1985). Pragmatics of Sibling Speech to One-Year-Olds. *Child Development*, *56*(4), 911–917. JSTOR. https://doi.org/10.2307/1130103

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

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., Palma, P. D., & MacWhinney, B. (2016). HomeBank: An Online Repository of Daylong Child-Centered Audio Recordings. *Seminars in Speech and Language*, *37*(2), 128–142. https://doi.org/10.1055/s-0036-1580745

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, *45*(6), 1611–1617. https://doi.org/10.1037/a0016134

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weisner, T. S., Gallimore, R., Bacon, M. K., Barry, H., Bell, C., Novaes, S. C., Edwards, C. P., Goswami, B. B., Minturn, L., Nerlove, S. B., Koel, A., Ritchie, J. E., Rosenblatt, P. C., Singh, T. R., Sutton-Smith, B., Whiting, B. B., Wilder, W. D., & Williams, T. R. (1977). My Brother's Keeper: Child and Sibling Caretaking [and Comments and Reply]. *Current Anthropology*, *18*(2), 169–190. https://doi.org/10.1086/201883

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*(2), 265–279. https://doi.org/10.1037/0012-1649.37.2.265

Wittek, A., & Tomasello, M. (2002). German children's productivity with tense morphology: The *Perfekt* (present perfect). *Journal of Child Language*, *29*(3), 567–589. https://doi.org/10.1017/S0305000902005147

Wong, K., Thomas, C., & Boben, M. (2020). Providence talks: A citywide partnership to address early childhood language development. *Studies in Educational Evaluation*, *64*, 100818. https://doi.org/10.1016/j.stueduc.2019.100818

Woollett, A. (1986). The influence of older siblings on the language environment of young children. *British Journal of Developmental Psychology*, *4*(3), 235–245. https://doi.org/10.1111/j.2044-835X.1986.tb01015.x

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., Gagarina, N., Hrzica, G., Ketrez, F. N., Kilani-Schoch, M., Korecky-Kröll, K., Kovacˇevic´, M., Laalo, K., Palmovic´, M., Pfeiler, B., Voeikova, M. D., & Dressler, W. U. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461–479. https://doi.org/10.1177/0142723711409976

Yung Song, J., Sundara, M., & Demuth, K. (2009). Phonological Constraints on Children's Production of English Third Person Singular – *s*. *Journal of Speech, Language, and Hearing Research*, *52*(3), 623–642. https://doi.org/10.1044/1092-4388(2008/07-0258)

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*(1), 1–22.

Appendix A

Litterature Review

Here(https://docs.google.com/document/d/1JyNIyu_KvMIjxKtxfh4ImGG7iUjR6T08XVHDmjrB4h8/edit?usp=sharing) is a review of papers with computational models looking at word/morpheme segmentation from running speech represented as text transcripts. Human experiments and segmentation from audio was excluded. The papers should contain information on two or more languages within the same paper and/or be comparable to another article (i.e., exact same algorithm with the exact same parameters). Papers with a single language or incomparable were excluded. The search was done using google scholar, and the keywords were “cross-linguistic infant word segmentation computational models transitional probabilities adaptor grammar minimum description”.

Appendix B

WordSeg

Behavior Research Methods manuscript No.
(will be inserted by the editor)

WordSeg: Standardizing unsupervised word form segmentation from text

Mathieu Bernard[1,2], Roland Thiolliere[1], Amanda Saksida[3], Georgia R. Loukatou[1], Elin Larsen[12], Mark Johnson[4], Laia Fibla[15], Emmanuel Dupoux[12], Robert Daland[6], Xuan Nga Cao[1,2], Alejandrina Cristia[1]

Received: date / Accepted: date

Abstract A basic task in first language acquisition likely involves discovering the boundaries between words or morphemes in input where these basic units are not overtly segmented. A number of unsupervised learning algorithms have been proposed in the last 20 years for these purposes, some of which have been implemented computationally, but whose results remain difficult to compare across papers. We created a tool that is *open source*, enables *reproducible results*, and encourages *cumulative science* in this domain. WordSeg has a modular architecture: It combines a set of corpora description routines, multiple algorithms varying in complexity and cognitive assumptions (including several that were not publicly available, or insufficiently documented), and a rich evaluation package. In the paper, we illustrate the use of this package by analyzing a corpus of child-directed speech in various ways, which further allows us to make recommendations for experimental design of follow-up work. Supplementary materials allow readers to reproduce every result in this paper, and detailed online instructions further enable them to go beyond what we have done. Moreover, the system can be installed within container software that ensures a stable and reliable environment. Finally, by virtue of its modular architecture and transparency, WordSeg can work as an open source platform, to which other researchers can add their own segmentation algorithms.

Keywords Unsupervised word discovery · First language acquisition · Natural Language Processing · Cumulative science

A. Cristia
29, rue d'Ulm, 75005, Paris, France
Tel.: +33 1 44 32 26 23
Fax: +33 1 44 32 26 30
E-mail: alecristia@gmail.com
[1]LSCP, Département d'études cognitives, ENS, EHESS, CNRS, PSL Research University
[2]INRIA
[3]Institute for Maternal and Child Health - IRCCS "Burlo Garofolo" - Trieste
[4]Macquarie University
[5]University of East Anglia
[6]University of California, Los Angeles

1 Introduction

One of the key tasks facing the language learning infant involves finding the minimal recombinable units present in the input. Since there are no systematic silences between words or morphemes, learners may need to carve them out from the running speech, a process known as segmentation. To do this, they may use a few universal and unambiguous cues (such as lengthy pauses), as well as a host of probabilistic cues. The latter can be classified into sublexical (e.g., which sound sequences tend to be found at word edges, and seldom within words) and lexical (e.g., certain words are more likely to follow each other than expected by chance). A number of computational algorithms building on subsets of such cues have been proposed, and several have been implemented in a variety of computer languages and applied to corpora so as to model infants' word form discovery processes. Typically, these models take as input a text-based, phonological representation of the input. To mimic the word discovery process, known word or morpheme boundaries are removed, and the algorithm is applied to try to make decisions on where breaks may occur, which are then compared against the original (gold) boundaries.

These studies are informative for a host of learnability questions, such as to test the sheer feasibility of a proposed word segmentation solution [12], to compare alternative algorithms [13,33], to see whether languages differ in their intrinsic segmentability [10], or whether child-directed speech is intrinsically easier to segment than adult-directed speech [25]. Additionally, there is emergent evidence suggesting computational word segmentation results may also be relevant for infant psycholinguistics, by predicting the contents of infants' long-term vocabulary better than lexical status [32] or pure frequency [22]. These results provide initial validation to the cognitive modeling approaches to word segmentation that have enjoyed a fair amount of attention for the last several decades (e.g., [4,12,13, 15], as they reveal that the latter may be close enough to infants' segmentation to make predictions that can be validated via direct experimental or correlational tests. In this context, it becomes crucial for the field to standardize segmentation methodology, so as to better explore the phenomenon of segmentation and make empirically informed predictions for infant experimental work.

In this paper, we present WordSeg, a software package conceived to allow this field of research to do cumulative science. The last few decades have seen a surge of interest in open science methods, where researchers' choices are rendered transparent, enabling others to replicate and extend results more easily. One could imagine this is even easier for computational modeling than, say, live experimentation, since typically modeling involves the creation of scripts which can be run time and again, are blind to the person executing them, and seem more context-independent than animals. And yet, recent articles continue to alert us on the unavailability of key research materials (including code) even of modeling work [14]. The first step towards cumulative science is thus to favor open source code, that is, code that is both available publicly and tagged for public re-use. But this is not enough. Even if the source code is made publicly available, it is often not set up to run in some other machine or operating system; and it is not sufficiently documented that it can be launched by some other user in an informed fashion so as to reproduce the original results [39]. Thus, the second step towards cumulative science involves providing appropriate documentation as well as taking steps to make sure reproducibility can be achieved outside the native context. The final

ingredient is to enable other researchers to directly build on previous work in a cumulative fashion.

With all of these considerations in mind, we created a tool that has a modular architecture (see Figure 1), combining a set of corpora description routines, several algorithms varying in complexity and cognitive assumptions, and a rich evaluation package, all integrated into a seamless pipeline. We have made our package openly accessible, and complemented it with supplementary materials allowing readers to reproduce every result in the current paper, as well as detailed online instructions further enabling them to go beyond what we have done. With this, we meet the first desideratum. Additionally, the whole system can be installed using Docker, ensuring that the environment will be stable across operating systems [17] – a requirement for reproducibility. Finally, by virtue of its modular architecture (and by clearly restricting and documenting e.g., input and output formats), the suite can work as an open source platform, to which researchers can add their own segmentation algorithms. This allows algorithm developers to benchmark their results against previously available segmentation algorithms, and should greatly facilitate making their own segmentation algorithm public – thus fitting the last desideratum, cumulativity. We believe this approach is extremely novel in our field: We cannot name one tool in psycholinguistics (or in another subfield of psychology) that attempts to provide a framework for *every* researcher to integrate and test their own model against others'.

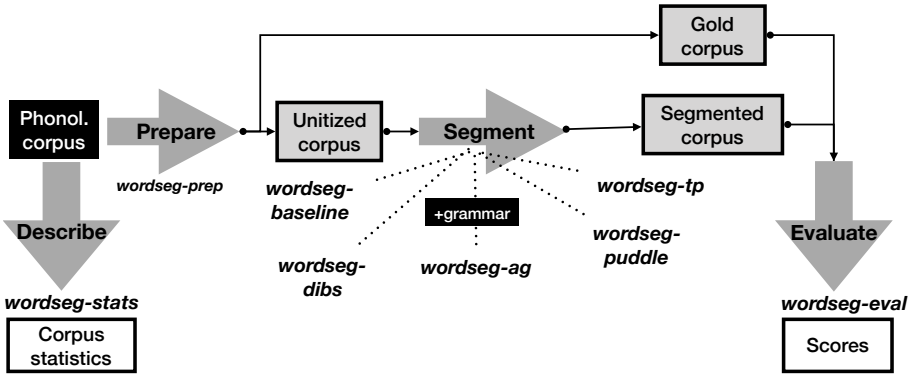

Fig. 1: **Overview of the WordSeg suite.** Black boxes represent input from the user; other boxes represent the output of a given stage; arrows represent the general description of procedures, most of which are implemented with a single command. The exception is the segmentation, where multiple segmentation processes are possible (parametric variation not shown).

We see two main use cases. The first involves fellow modelers, who are developing alternative unsupervised word segmentation algorithms. As just mentioned, our package can serve as a common platform that standardizes input and evaluation, and provides a set of alternative algorithms against which developers can benchmark their own tool. Moreover, they can then profit from the effort that has gone into making this package widely deployable by simply adapting their tool to

the WordSeg architecture and adding it as a new WordSeg module. The second set of users are linguists and other cognitive scientists interested in early language acquisition. This second group would not develop additional code, but rather make use of the standardized user interface to describe and analyze their child language corpora, or respond to specific scientific questions. For instance, a user may be curious about the ease of segmentation of social words (such as "mommy" and the baby's name) in different languages. This user could apply all segmentation algorithms, and then estimate with what frequency these words appear as such (i.e., are not obscured by under- or over-segmentation) in the segmented output. Such WordSeg uses are extremely straightforward for anyone who knows how to interact with a terminal (and for readers who do not, we recommend Software Carpentry's introduction `http://swcarpentry.github.io/shell-novice/`).

2 Previous computational modeling work

It is beyond the scope of the present article to provide a comprehensive review on computational models of infant word segmentation, and thus we refer interested readers to [6] for a fuller introduction to the basic issues surrounding computational models of infant word form segmentation, [3] for a historical classification of models, and [34] for a recent literature review on the topic. It suffices here to state that this phenomenon has garnered considerable attention, but researchers have used varying methodologies in a way that compromises comparability. Section 2.1 lays out the main approaches that are currently represented in the package. Our package sought to also systematize "irrelevant" variation, as explained in Section 2.2.

2.1 Classes of algorithms currently represented in the package

A systematic literature review[1] of 46 journal articles or theses that contained modeling results on word form segmentation published between 1993 and 2015 revealed that there are more postulated algorithms than papers, particularly when free parameters are taken into account. Thus, it was simply impossible to attempt to incorporate all previous algorithms. Our selection aimed at representing a few key dimensions of variation across open source algorithms, and it was constrained by the availability of code and quality of the documentation.

One key distinction among included models pertains to whether they rely purely on local cues for word segmentation such as transitional probabilities between sounds or syllables. We will call this class sublexical. The lexical alternative involves aiming to reparse the input stream in terms of minimal recombinable units, or, put otherwise, building the lexicon that would be ideal to generate the corpus. This conceptual distinction does not prevent the existence of models that are hybrid. For instance, one of the models included in the suite is PUDDLE [31], which uses both lexical and phonotactic cues (see Section 3.5.4).

[1] The last author performed a search with the terms *infant "word segmentation" "computational model"* in scholar.google.com on August 10, 2015. The top 220 items were extracted automatically using Zotero. They were thereafter inspected manually, excluding as off-topic 174 on the basis of title, abstract, or full-text.

Additionally, some previous work has argued strongly for algorithms that process information **incrementally**, compared to others that do so in a **batch** mode (e.g., [31]). Although we believe that, to a certain extent, the dichotomy can be ill-posed, our sampling reflects both batch and incremental learners. We return to this topic in the discussion.

Two additional classes of models are not represented in the WordSeg suite. Unsupervised segmentation models that use raw speech as input and can fully parse a corpus are uncommon in the speech technology literature [43], and not at all represented in work modeling infant word segmentation. The only exceptions we know of are closer to keyword discovery than full segmentation (e.g., [26]). Additionally, neural network type models are not represented either, mainly because this is an area of rapid technological development as neural networks are increasingly used for natural language processing in a wide range of applications including word segmentation (e.g., [5]).

2.2 Keeping other aspects constant

Most previous work uses only one or a very limited set of models, so that to decide which model performs better one often needs to compare performance across papers. However, our systematic review revealed a host of dimensions that varied across papers, and which prevent direct comparison across published work. Most saliently, it is not uncommon to observe extremely large variations in the size of the corpus used as input (e.g., [37] based on around 10,000 words versus [7] drawing on 750,000 words). Moreover, previous work investigating the effect of input quantity among Adaptor Grammars found effects that were non-linear and dependent on the grammar itself [2], making it all the more difficult to compare model performance across studies (see also [7,18,31], for further discussions of corpus size effects).

In early modeling work, it was not uncommon to use artificial corpora, and even in some current work the input consists of transcripts from broadcast speech or adult-directed speech (such as the Buckeye corpus [36]). Using such input is no longer warranted, since corpora on the CHILDES [28] repository contain hundreds of transcriptions that are child-centered. These are likely to be ecologically valid, because recordings were gathered in children's natural environments, and often with a recording device worn by the child, thus capturing both child-directed and child-overheard speech available to the child.

For studies using CHILDES corpora, there are some sources of variation whose impact has not been sufficiently considered. Although it would seem that corpora are sure to be homogeneous if drawn from the CHILDES repository, different contributors actually use different criteria to define sentences. We have noticed that some corpus contributors are probably using a "breath group" or even "conversational turn" definition, since there may be 10-20 words in a given sentence. In contrast, others are probably using a syntactically or prosodically defined sentence, with overall shorter utterances, averaging 3 words in length. Additionally, researchers studying word segmentation often mix together various corpora from children of diverse ages without controlling for the possibility that the length and complexity of sentences and the lexical diversity in them varies as a function of the

child's age. Despite the fact that they probably explain variation in segmentation performance, such characteristics are seldom thoroughly reported.

An additional source of variation relates to whether phones or syllables are the basic units at the phonological level. For example, Phillips and Pearl [35] report better performance when the basic units were syllables, rather than phones, and argued in favor of syllables on plausibility grounds. Evaluating plausibility is not within the scope of the present paper. As for performance, Larsen and colleagues fully crossed basic unit against algorithm drawing from the sublexical, lexical, and hybrid types, and although in general F-scores were higher for syllables than phones, some exceptions remained [22]. Moreover, ranking across algorithms also depended on representational unit.

Finally, nearly every research paper on computational models of infant word segmentation contains arguments for and against the range of evaluation metrics that are typically used, prioritizing precision over recall, arguing that type statistics are more interesting than token statistics – or vice versa.

All of this variation seriously impedes direct comparison across published studies, and makes it difficult for researchers to decide how to set up their preprocessing and analysis pipelines to optimize comparability with previous work.

3 The WordSeg suite

The WordSeg suite allows the use of several algorithms drawn from previous literature in a controlled environment that standardizes input and allows users to easily report the full range of input and output statistics allowing cross-paper comparison. The overall process is represented in Figure 1. Detailed instructions for use are available as online materials, which are updated as issues arise (`https://wordseg.readthedocs.io`). The version used for the current work is 0.7.1[2].

3.1 Technical characteristics

The package is distributed from `https://github.com/bootphon/wordseg`, with a GPL-3.0 re-use license, from where it can be cloned or downloaded as a zip. In all cases, WordSeg requires several additional pieces of software (e.g., Python 3) to function. Installation instructions are provided covering how to download and install this ancillary software, as well as how to install WordSeg itself. The user can install WordSeg such that it will be available anywhere within the system, or only in a virtual environment via the use of Docker™ [17]. WordSeg has been thoroughly tested in a Linux environment, and less so in UNIX and Windows. WordSeg has native support for Linux and has been thoroughly tested on MacOS and Windows. Once the system is installed, users can use WordSeg as a command line interface from a Bash terminal or as a library from Python, with both series of commands described and exemplified in the online documentation `https://wordseg.readthedocs.io`. The code contained in WordSeg is mostly Python and C++, with variability being mainly due to the included segmentation algorithms.

[2] `https://zenodo.org/record/1471532,https://github.com/bootphon/wordseg/releases/tag/v0.7.1`

3.2 Input selection, cleaning, and phonologization

The suite does not directly support full pre-processing and phonologization of corpora, but we provide some pointers for users. For most researchers, the starting stage will be a CHILDES style `.cha` file, which contains comments as well as transcribed content. These first stages of cleaning will be dependent on the particular corpus because they vary somewhat across CHILDES corpora, and on the research question, since researchers may want to include or exclude specific speakers or utterances. Sample scripts we have used in the past can serve as inspiration (see the `/data/cha/` section of the package). Additionally, the WordSeg suite assumes that the input has already been phonemized and syllabified. For corpora in which this has not been done, we recommend readers look into the Phonemizer package (`https://github.com/bootphon/phonemizer`), which provides tools to convert text to phonemes. Another option is the WebMaus automatic segmentation tool (`https://www.clarin-d.net/en/webmaus-basic-`), which converts text files to phonemic transcriptions based on trained statistical models. For languages with a transparent orthography, hand-crafted rules can be used to derive the phonemic representation of words. Examples are provided in the `/data/phonorules/` section. Finally, users may want to employ a syllabification routine using the Maximize Onset Principle, a rule of thumb whereby a sequence of phones will be parsed such that the onset cluster will be as heavy as the language allows. For instance, the sequence /estra/ will be broken up into /es.tra/ in Spanish and /e.stra/ in English. We have adapted perl code that does so from [35] and provide examples in the `/data/syllabification/` section and the `wordseg-syll` tool.

3.3 Preparing the input

For the rest of the processes, the package assumes that the input file contains only the transcribed utterances in phonological form, one utterance per line. Additionally, it is assumed that word boundaries and basic units are coded in the input text. The input text can have one or both of the following basic units: phones, syllables.

The `wordseg-prep` tool in the package allows users to convert the input text from the input form where syllable and word boundaries are tagged to the input to be provided to the models. This tool outputs a unitized version and a gold version of the text. A unitized version contains spaces between phones or syllables (as chosen by the user). The gold version only has spaces between words. The gold text will be used later to evaluate the output of segmentation.

3.4 Describing the corpus

The package also contains `wordseg-stats`, a tool to describe the input corpora. This description tool prints out the number of all of the following units: sentences or lines, single-word utterances, and number of tokens, types, and hapaxes (i.e., types with token frequency of exactly one) for words, syllables, and phones. Additionally, a measure of lexical diversity that controls for corpus length is extracted, namely a moving average type to token ratio similar to that available in

Acronym	Class	Processing	Key units
baseline	sublexical	batch	units
dibs	sublexical	batch	unit bigrams
tp	sublexical	batch	unit bigrams
puddle	hybrid	incremental	unit n-grams, words
ag	lexical	batch	words

Table 1: Segmentation algorithm families currently included in WordSeg. We say "families" because each has a set of parameters that allows further variation. Class indicates the main class the algorithm belongs to; Processing whether the input is processed in batch or incrementally; and Key units the crucial representations that the algorithm uses for segmentation.

the CHILDES tools [27], where a window of 10 word tokens are considered at a time, moved one token at a time. Finally, `wordseg-stats` returns a measure of entropy, i.e. the intrinsic ambiguity found in a text (see [10] for details). In a nutshell, given a set of utterances and the lexicon found in the gold segmentation, this measure of entropy assesses to what extent there are many versus few possible parses of the utterances (i.e., in a corpus with 2 sentences, "ice cream" and "icecream", both utterances are ambiguous between "ice cream" and "icecream" segmentations).

3.5 Segmenting

All of the algorithms are called with variants of `wordseg-X`, where X is the short name for the algorithm (as shown on Table 1), together with the necessary parameters and ancillary files, both of which depend on the specific algorithm. The input for all algorithms is plain text as built by `wordseg-prep`, where only unit tokens (syllables or phonemes) are available and separated by single spaces (that is, the word boundaries have been removed), but some of them additionally require a training set or a configuration file. In the rest of this section, we provide a general description of each algorithm, parametrization and required files. We have not incorporated standardized measurements of memory requirements or length of processing, because these, we believe, could largely relate to details of implementation which may not affect fundamentally the results found.

3.5.1 Baseline

Researchers might be interested in comparing baseline results to those of the word segmentation algorithms. The WordSeg package provides tools for word segmentation baselines based on the insertion of word boundaries in random positions in the text, explained for instance by Lignos [24].

The Random Baseline assigns word boundaries with a probability parameter p specified by the researcher. By default, a random segmentation consists in adding word boundaries with $p = 0.5$ to each unit token. The user can specify a random seed, to ensure reproducibility. Alternatively, the researcher can choose $p = 0$ to generate an "Utterance Baseline", considering each utterance as a single word; and $p = 1$, to insert all possible boundaries and treat each unit token (phones or

syllables) as a word. The researcher can also inspect the statistics mentioned in Section 3.4 to calculate the true p of word boundaries given the basic unit (e.g., for a corpus unitized into syllables, $p = \frac{nw}{ns}$, where nw is number of words and ns is number of syllables). This number can then be provided by the user as the p parameter, in which case, this would be an Oracle Random Baseline [24] ("oracle" because it is given the true p by the researcher; random because it will insert the correct number of boundaries to match p, without knowing where they should occur).

3.5.2 Diphone Based Segmenter (DiBS)

Daland's DiBS (short for Diphone-Based Segmentation, [7]) uses phone bigram probabilities to decide whether a specific sequence is likely to span a word boundary (typically because the phone bigram is rare) or not. A DiBS model is any model which assigns, for each phrase-medial phone bigram, a value between 0 and 1 inclusive, representing the probability the model assigns that there is a word boundary between the two phones. In practice, these probabilities are mapped to hard decisions (break or no break).

Making these decisions requires knowing the chunk-initial and chunk-final probability of each phone, as well as all phone bigram probabilities; and additionally the probability of a sentence-medial word boundary. In our package, these 4 sets of probabilities are estimated from a training corpus also provided by the user, where word boundaries are marked. Please note we say chunk-initial and chunk-final because the precise chunk depends on the type of DiBS used, as explained in the next paragraph.

Three versions of DiBS are available. DiBS-gold is supervised in that "chunks" are the gold words. It is thus supposed to represent the optimal performance possible. DIBS-phrasal uses phrases (sentences) as chunks. Finally, DIBS-lexical uses as chunks the components of a seed lexicon provided by the user (who may want to input e.g. high frequency words, or words often said in isolation, or words known by young infants).

By default, the sentence-medial probability of word boundary is calculated in the same way for all three DiBS, and it is the actual gold probability (i.e., the number of words minus number of sentences, divided by the number of phones minus number of sentences). Via a parameter, users can also provide the algorithm with a probability of word boundary calculated in some other way they feel is more intuitive.

DiBS was initially designed with phones as basic units. However, for increased flexibility we have rendered it possible to use syllables as input.

3.5.3 Transitional Probabilities (TP)

Like DiBS, the next family of algorithms attempts to distinguish between more or less *internally cohesive* phone/syllable sequences. In the implementation we have adopted [37], transitional probabilities (TPs) are calculated in one of three ways:

- Forward TPs for XY are defined as the frequency of the sequence XY divided by the frequency of X;

- Backward TPs for XY are defined as the frequency of the sequence XY divided by the frequency of Y;
- Mutual information for XY is the log (base 2) of the frequency of the sequence XY divided by the product of frequency of X and that of Y

This direction parameter is crossed with another, defining a cut-off for how low TPs must be to signal a boundary, and which also has two settings. In the first, a boundary is posited when a relative dip in TP is found. That is, given the syllable or phone sequence WXYZ, there will be a boundary posited between X and Y if the TP for XY is lower than both that for WX and that for YZ. The second setting uses the average of the TP over the whole corpus as the threshold. Notice that both of these are unsupervised: Knowledge of word boundaries is not necessary to compute any of the parameters.

TP was initially designed with syllables as basic units, but has been adapted to accept either phones or syllables as input in this package.

3.5.4 PUDDLE

PUDDLE stands for Phonotactics from Utterance Determines Distributional Lexical Elements. This algorithm was proposed by Monaghan and Christiansen [31]; the original awk rendering (shared with us by Monaghan) was reimplemented in Python for this package. PUDDLE takes the opposite strategy of algorithms such as DiBS and TPs that focus on local events to posit breaks. In contrast, PUDDLE takes in whole utterances and tries to break them apart into relatively large chunks. The system has three long-term storage units: a "lexicon", a set of onset bigrams, and a set offset bigrams. At the beginning, all three are empty. The lexicon will be fed as a function of input utterances, and the bigrams will be fed by extracting onset and offset bigrams from the lexicon. The algorithm is incremental, as follows.

The model scans each utterance, one at a time and in the order of presentation, looking for a match between every possible sequence of units in the utterance and items in the lexicon. We can view this step as a search made by the learner as he tries to retrieve from memory a word to match it against the input. If, for a considered sequence of phones, a match is found, then the model checks whether the two units preceding and following the candidate match belong to the list of ending and beginning bigrams, respectively. Imagine a target utterance like "thisisacutebaby", unitized at the phone level; a lexicon containing the item "this"; possible bigrams thus being "th" for onsets and "is" for offsets. Although "this" is found in the target utterance, the utterance will not be split because the remainder, "**is**acutebaby", does not begin with a permissible onset. It should be born in mind that this constraint is crucial for the model to avoid over-segmentation: If not applied, the model will ultimately segment the corpus to the basic unit level (e.g., phones). If a substring match is not found, then the utterance is stored in the long-term lexicon as it is, and its onset and offset bigrams will be added to the relevant buffers. Thus, in the running example, the lexicon will end up containing two items "this"and "thisisacutebaby"; the onset buffer will have the item "th" with a frequency of 2; and the offset buffer will have "is" and "by", each with a frequency of 1.

In our implementation of PUDDLE, we have rendered it more flexible by assuming that users may want to use syllables, rather than phones, as basic units.

Additionally, users may want to set the length of the onset and offset n-grams. Some may prefer to use trigrams rather than bigrams; conversely, when syllables are the basic unit, it may be more sensible to use unigrams for permissible onsets and offsets.

3.5.5 Adaptor Grammars (AG)

In the adaptor grammar framework [13,19], parsing a corpus involves inferring the probabilities with which a set of rewrite rules (a "grammar") may have been used in the generation of that corpus. The WordSeg suite natively contains the capacity to generate one grammar, the most basic and universal one. Users can also create their own and/or change extant ones to fit the characteristics of the language they are studying (see the `/data/ag/` section of the package for more examples).

The simplest grammar, automatically generated with the call `wordseg-ag`, can be conceived as having one rewrite rule to the effect that "sentences are one or more words", one rewrite rule to the effect that "words are one or more basic units", and a set of rewrite rules that spell out basic units into all of the possible terminals. Imagine a simple language with only the sounds a and b, the abstract rules would then be:

- Sentence → Word (Word)+
- Word → Sound (Sound)+
- Sound → a
- Sound → b

A key aspect of adaptor grammar is that it can also generate subrules that are stocked and re-used. For instance, imagine "ba ba abab", a corpus in the above-mentioned simple language. As usual, we remove word boundaries, resulting in "babaabab" as the input to the system. A parse of that input using the rules above might create a stored subrule "Word → ba"; or even two of them, as the system allows homophones. The balance between creating such subrules and reusing them is governed by a Pittman-Yor process, which can be controlled by the user by setting additional parameters. For instance, one of these parameters, often called "concentration," determines whether subrules are inexpensive and thus many of them are created, or whether they are costly and therefore the system will prefer reusing rules and subrules rather than creating new ones.

The process of segmenting a corpus with this algorithm will in fact contain three distinct subprocesses. The first, as described above, is to parse a corpus given a set of rules and a set of generated subrules. This will be repeated a number of times ("sweeps"), as sometimes the parse will be uneconomical or plain wrong, and therefore the first and last sweeps in a given run will be pruned, and among the rest one in a few will be stored and the rest discarded.

The second subprocess involves applying the parses that were obtained in the first subprocess onto the corpus again, which can be thought of as an actual segmentation process. Remember that in some parses of the "ba ba abab" corpus (inputted as "babaabab"), the subrule "Word → ba" might have been created 0, 1, or 2 times. Moreover, even if we ignore this source of variation, the subrules may be re-used or not, thus yielding multiple possible segmentations ("baba abab" with no subrule, "ba ba a ba b" with one "Word → ba" subrule or the same with 3 "Word → ba" subrules, etc.)

The third and final subprocess involves choosing among these alternative solutions. To this end, Minimum Bayes Risk is used to find the most common sample segmentations.

As this description shows, there are many potential free parameters, some that are conceptually crucial (concentration) and others that are closer to implementation (number of sweeps). By default, all of these parameters are set to values that were considered as reasonable for experiments (on English, Japanese, and French adult and child corpora [10,18]) running at the time the package started emerging, and that we thus thought would be a fair basis for other general users. The full list can be accessed by typing `wordseg-ag --help`. The following is a selection based on what is often reported in adaptor grammar papers:

- number of runs: 8
- number of sweeps per run: 2000
- number of sweeps that are pruned: 100 at the beginning and end, 9 in every 10 in between
- Pittman Yor a parameter: 0.0001
- Pittman Yor b parameter: 10000
- Rule probability (theta) is estimated using Dirichlet prior

3.6 Evaluation

An objective way to measure the performance of word segmentation algorithms is to compare the segmented corpus with the gold one, which corresponds to a perfect segmentation as would be done by a literate adult. This comparison can be done at different levels: word token, word type, and boundary. We provide two boundary scores, one counting utterance edges and the other not counting edges (since these will always be correct, by definition).

At a particular level, the evaluation looks at two different criteria: precision, the probability that a segmented boundary/token/type is correct; and recall, the probability a correct boundary/token/type has been segmented. Concretely, the precision P and recall R are calculated as follows:

$$P = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{1}$$

$$R = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{2}$$

The harmonic mean between precision and recall is computed to give the F1, which we will call F-score.

To take an example, imagine a corpus ‘the dog bites the dog’; the segmented output is ‘the dog bites thedog’. This will yield the following performance:

- token precision: 0.75, recall: 0.6, F-score: 0.67
- type precision: 0.75, recall: 1, F-score: 0.86
- boundary precision: 1, recall: 0.83, F-score: 0.91
- boundary no edge precision: 1, recall: 0.75, F-score: 0.86

Two additional evaluation outputs are provided at the user's request. First, users can obtain the Rand Index RI, which captures both true positives and negatives. It is calculated as follows:

$$RI = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives}} \quad (3)$$

Our evaluation actually provides the Adjusted Rand Index, where both numerator and denominator have been adjusted for chance agreement via resampling.

Second, some readers may be specifically interested in finding out which lexical items come to be correctly segmented, or else segmented incorrectly in one of these three ways: undersegmented (i.e., joined with a neighboring word); oversegmented (i.e., broken down into subparts); or plain mis-segmented. An optional parameter yields an evaluation summary file being returned which contains all words in the gold corpus and the number of times with which they were found in each of those 4 groups.

One important consideration pertains to incremental algorithms, in which performance is changing throughout the corpus. To make their evaluation comparable to that of the others, we implemented a system of corpus folding, with a default of 5 folds (which can be parametrized by the user). For the first fold, a given algorithm is run in the whole corpus. Next, the final 20% of the corpus is moved to the onset of the corpus, and the algorithm is run again, such that this time the final 20% will in fact be the utterances that start at the 60% point in the corpus and end at the 80% point. This process repeats for the remaining 3 folds (40-60%, 20-40%, 0-20%). At this point, the final 20% of the corpora outputted in each of the 5 runs is concatenated in the right order, and the whole is evaluated. Please note that this is not an instance of cross-validation, since the models may continue learning over the last 20%.

4 Examples of use

This section has three goals. First and foremost, we aim to illustrate the package and show its flexibility. This example allows users to have a benchmark when they themselves use the package. Since we expect that we and others will continue improving it, however, we recommend users check `https://github.com/alecristia/wordseg-brm-analyses` for an up-to-date version of these results as well as reproducible code. Second, we would like to inform researchers working on this domain on the impact of key methodological and conceptual decisions, such as what input and evaluation units are used. Finally, we try to assess the conditions in which performance is *stable and replicable*.

Crucially, we would like to make it clear from the start that the goal is not to compare performance across algorithms to find the best-performing one. Best performance against orthographic standards does not mean that the algorithm represents human performance, let alone infant performance. For instance, [22] found essentially a zero correlation between algorithms' F-scores against adult segmentation and the proportion of variance explained in infant word knowledge in an English sample. Thus, we consider that, at present, there is insufficient evidence to determine which algorithm best captures human (infant and adult) performance, and that they may all be valuable and informative to the computational modeler

interested in the psychological phenomena surrounding word form segmentation. We want to provide the research community with an array of algorithms which, given the uncertainty regarding the information that infants and other learners incorporate, has a high likelihood of capturing at least some behaviors, or at the very least allows the researcher to focus on findings that are true *regardless of which algorithm is used as a proxy.*

4.1 Methods

4.1.1 Corpus

We used the Providence corpus [2,8], available from CHILDES [28] because it is commonly used and large enough to allow us to break it down into several subparts, and apply inferential statistics to assess whether certain factors truly explain significant proportions of variance. It contains transcriptions of recordings gathered from 6 American English-speaking children. Recordings started when children spoke at least 4 words according to parental report, which happened when they were around one year of age. About one hour of child-context interactions were recorded every 1-2 weeks until they were around 3 years of age. For the present study, we focus on the 74 transcripts (from 5 children) meeting the following desiderata:

- children were two years of age or younger
- there was only one adult present (which lowers the likelihood of including adult-directed speech)
- there were at least 300 utterances spoken by the adult

These transcripts were cleaned using custom bash scripts, which removed all comment lines and all sentences uttered by children. The resulting orthographic representations were phonologized using FESTIVAL [41], which yields a representation including syllable boundaries. FESTIVAL uses a dictionary look-up system, complemented with grapheme-to-phoneme conversion rules for words not in the dictionary. The following is an example of the resulting "Tags" representation, which contains spaces to mark phone boundaries, ;s for syllable boundaries, and ;w for word boundaries.

1. Orthographic: you wanna sit with mommy
2. Tags: y uw ;s ;w w aa ;s n ax ;s ;w s ih t ;s ;w w ih dh ;s ;w m aa ;s m iy ;s ;w
3. Gold: yuw waanax siht wihdh maamiy

4.1.2 Processing with WordSeg

We generated the results for all the experiments below with a single Bash script (although we could have used Python instead). The following is a version of that Bash script, simplified for ease of inspection:

```
#!/bin/bash
# segment independent transcripts
FOLDER="/Providence/"
for tag in $FOLDER/*tags.txt; do
```

```
  # compute statistics on the unitized input text
  cat $tag | wordseg-stats --json > ${tag}_stats.json
  # prepare the input for segmentation and generate the gold text
  cat $tag | wordseg-prep --unit $unit --gold gold.txt > prep.txt
  # segment the prepared text with different algorithms
  # sublexical
  cat prep.txt | wordseg-baseline --probability 0.0 > ${tag}_seg.base00.txt
  cat prep.txt | wordseg-tp --threshold relative > ${tag}_seg.tprel.txt
  cat prep.txt | wordseg-dibs --type phrasal --unit $unit $tags > ${tag}_seg.dibs.txt
  # lexical
  cat prep.txt | wordseg-ag  > ${tag}_seg.AGu.txt
  # hybrid
  cat prep.txt | wordseg-puddle --window 2 > ${tag}_seg.puddle.txt
  # evaluate against the gold file
  for segmented in ${tag}_seg.*.txt; do
    algo=$(echo $segmented | sed 's/.*seg.//' | sed 's/.txt//')
    cat $segmented | wordseg-eval gold.txt > ${tag}_out.${algo}.txt
  done
done
```

The sample script above represents the following conceptual decisions:

- All algorithms are fed with a phone-unitized version of the corpus,
- The baseline is that which segments at utterance level only,
- For TP version uses the forward TP (default), with a relative threshold,
- The version of DiBS chosen in this example is the phrasal type, using the full corpus to extract phone bigram statistics,
- For AG, since no grammar was provided, the simple one mentioned above is automatically generated,
- For PUDDLE, we used bigrams (window of 2)

The full script can be retrieved from https://github.com/alecristia/wordseg-brm-analyses/blob/master/do_prov.sh. It actually feeds all algorithms with both phone- and syllable-unitized input, contains 3 baselines (cut at utterance boundary, at every unit boundary, and at half of them); and TP is run with both an absolute and a relative threshold.

4.2 Corpus statistics

Our call to wordseg-stats allowed us to describe the analyzed transcripts. Table 2 shows means and SDs of various corpus characteristics that are calculated by the statistics package, as well as some that can be derived from the former. The most important message we would like to convey here is that the standard deviations are quite high, particularly for sentence length. This is despite the fact that we focused on a single corpus, and further restricted inclusion to transcripts collected when children were younger than 2 years of age. Nonetheless, there are sizable changes in average sentence length, which may impact segmentation performance.

Characteristics	Mean	SD
N phone tokens	11,463.22	3,414.43
N phone types	39.62	0.51
N syllable tokens	4,581.11	1,350.05
N syllable types	688.54	163.65
N words tokens	3,720.28	1,075.26
N words types	670.99	178.71
N word hapax	293.15	93.43
MATTR	0.89	0.04
Entropy	0.018	0.002
N SWU	102.81	55.33
N utts	700.27	200.38
Derived metrics		
Prop. SWU	0.14	0.04
Prop. hapax	0.43	0.04
Avg. phones/word	3.08	0.08
Avg. syllables/word	1.23	0.03
Avg. words/utt	5.38	0.93

Table 2: **Corpus characteristics of individual transcripts.** Tokens refers to unique instances, types to abstract units. Hapax stands for types that occur exactly once. MATTR stands for Moving Average Type to Token Ratio, a TTR calculated over 10 consecutive words so as to control for overall corpus size. Entropy is a measure of ambiguity in segmentability; a higher number means more ambiguity. Utt(s) stands for utterances; SWU for Single Word Utterance.

4.3 Effects of processing unit and algorithm

As mentioned above, we have analyzed each transcript within a subset of the Providence corpus separately, encoded in terms of phones and syllables, with a set of algorithms. In this section, we report on analyses aimed at assessing to what extent performance is affected by these two factors and their interaction. As shown in Supplementary Materials (`https://github.com/alecristia/wordseg-brm-analyses/blob/master/supmat.pdf`), all performance metrics are highly correlated with each other. Therefore, we focus here exclusively on token F-scores. Figure 2 shows that performance varies enormously as a function of algorithm and basic unit, with important interactions between the two. Next, we highlight aspects of these results relevant to our three goals for Section 4.

The first result that may attract readers' attention is that performance varies greatly across algorithms. For instance, as has been discussed elsewhere [12,40], excellent scores can be achieved in English infant-directed speech samples like this one by simply segmenting every syllable, our Baseline with $p = 1$ algorithm. Above and beyond the specific explanation, this observation highlights the usefulness of WordSeg's included baseline algorithms.

A second conclusion is that algorithm and unit interact. The reason is obvious for two cases: TP (absolute versus relative), and PUDDLE. For TP, performance is higher for syllable-as-unit than phone-as-unit when using an absolute threshold, but the opposite for a relative threshold. The reason is probably that the relative threshold algorithm requires at least 4 units in a row to be able to find a local dip [12]. Therefore, no boundary can be postulated in short sentences, with fewer than 4 syllables. In contrast, a boundary can be postulated in short sentences

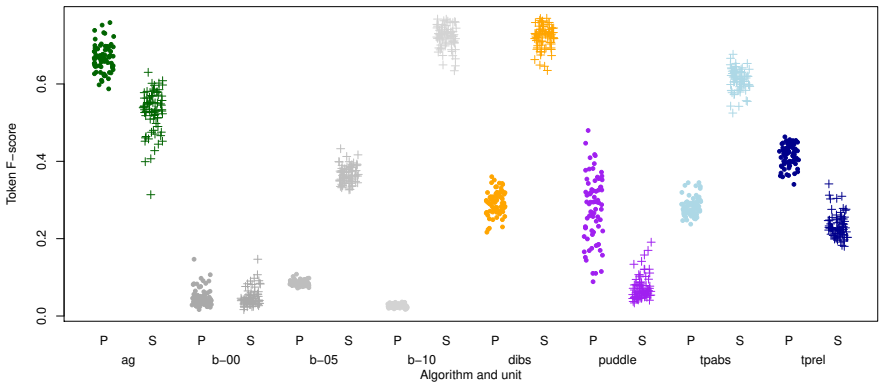

Fig. 2: **Token F-scores as a function of unit and algorithm.** Each point is the performance of a segmentation experiment on one of the 74 transcripts, using either phones (circles, left) or syllables (crosses, right) in combination with one of the 8 algorithms (distinguished by position on the x-axis as well as color). In baselines, b-00 stands for $p = 0$; b-05 for $p = 0.5$; b-10 for $p = 1$.

when these are represented in phones, because a local dip can be established when there are few syllables (provided these contain at least 4 phones).

A similar conclusion can be drawn from the PUDDLE performance, which was higher for phones than syllables. By setting the window for onset and offset buffers uniformly at 2, we effectively prevented the algorithm from breaking up more utterances when unitizing with syllables than with phones.

A third conclusion is that performance is enormously affected by unit and algorithm. To investigate this more precisely, we fit a regression with token F-scores as dependent measure, unit and algorithm as well as their interaction as fixed effects, and transcript identity as blocking factor.[3] This model explained 98% of the variance in performance, with both main effects and their interaction being highly significant.

4.4 Effects of corpus length

Although the analysis in the previous section showed that nearly all the variance in performance across transcripts was explained by algorithm, unit, and their interaction, it remains possible that transcript characteristics do affect word segmentation performance. As discussed in Section 2.2, a good candidate for a factor that would affect performance is corpus length. Preliminary analyses revealed that PUDDLE's performance was changing as a function of corpus length within the sample studied in the previous subsection. Therefore, we carried out an additional experiment to extend the length coverage. We followed previous work [2,7,31] by

[3] This regression was preferred over a mixed model because there is disagreement as to how to estimate proportion of variance explained in the latter.

submitting concatenated versions of the transcripts to our segmentation procedure. That is, we first analyzed the first transcript; then, we concatenated the first two by pasting the second transcript after the first and analyzed the resulting combined corpus; and proceeded in this manner until all included transcripts had been concatenated. Children vary in the number of included transcripts both because some were visited more regularly and from an earlier age (e.g., Naima), and because a different proportion of transcripts were excluded (due to being too short or containing more than one adult, see 4.1.1; e.g., only 4 out of 40 transcripts for William are included here).

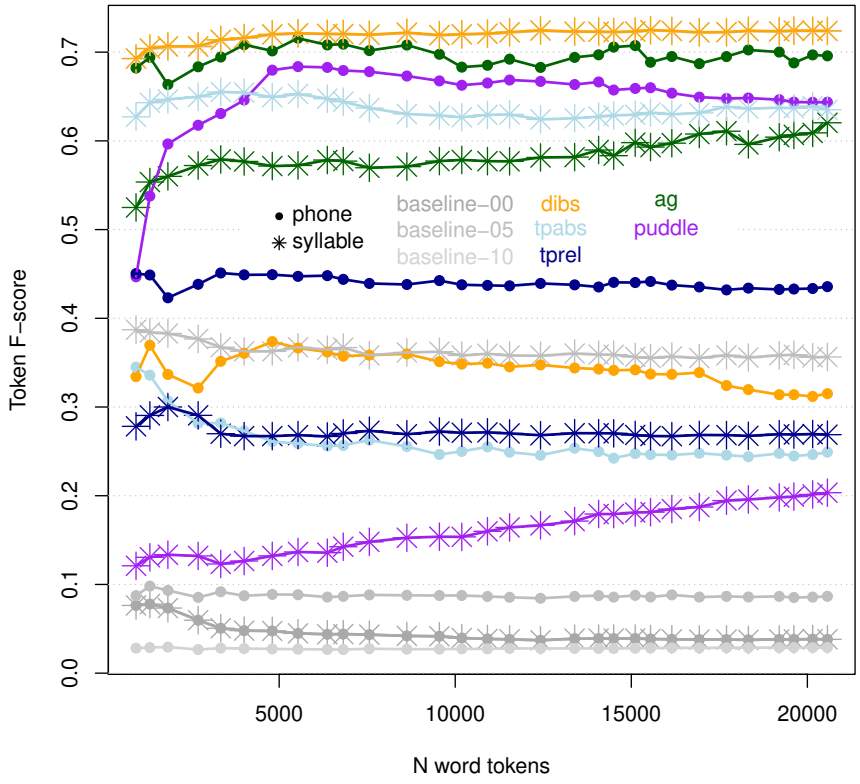

Fig. 3: **Token F-scores as a function of unit and algorithm in cumulative transcripts.** Each point is the performance of a segmentation experiment on a transcript that is the result of the concatenation of a given transcript and all preceding transcripts for a given child. Since variation across children was low, only Naima's data are shown here; see supplementary materials for the other curves.

Figure 3 portrays performance on Naima's transcripts. Since variation across children's data was very low, we have only included there results from one child to facilitate readers' visual inspection of results (see full figure in the online supplementary materials). All algorithms exhibit strong changes (upwards or downwards) in the 0-3k region, which may be associated to peculiarities of some of these transcripts, since they are visible even in the baselines. Afterwards, most algorithms remain fairly stable with very slight linear changes (if any), with one exception: PUDDLE. Indeed, we notice that PUDDLE-phones exhibits a non-linear pattern, with performance increasing rapidly between 1k and 4k sentences; peaking at 5-7k sentences; and slowly dropping (by little) thereafter. PUDDLE-syllables increases slowly and linearly throughout the range. In general terms, then, performance is stable for all algorithm-unit combinations (except for PUDDLE-syllables) in the 5k-15k region.

We investigated the effects of corpus size more precisely by fitting a linear regression taking the last data point for each child (i.e., the concatenation of all the transcripts associated with that child). As before, the dependent measure was token F-score and the predictors were the algorithm in interaction with the unit (blocked within child). This regression explained 99.2% of the variance; addition of number of words in interaction with algorithm increased this to 99.3% (which was not significant in a chi-square test).

In short, we have found that for most algorithm-unit combinations, performance is stable across a wide range of corpora sizes (roughly between 5,000 and 15,000 word tokens), and furthermore that corpora size affected performance very minimally once algorithm-unit effects were taken into account.

5 Discussion

This paper presents a package that allows the systematization of several key steps in the study of word form segmentation by infants and other agents. One of the strengths of our package is that it contains a tool to describe the input. Our analyses of a CHILDES corpus demonstrates that there is wide variability in the input to children even within 1-2 years of age in terms of sentence length and lexical properties which may impact segmentation performance. The package also provides all basic performance measures. Our analyses suggest that these are by and large correlated.

Another key strength of the package is the presence of a tool to unitize this input into phones or syllables as basic phonological units, and a third is that the package contains a range of conceptually diverse algorithms. Our analyses demonstrate that the crossing of these two factors (basic representational unit, and algorithm) has enormous effects on segmentation performance. In contrast, segmentation performance was rather stable across a wide range of corpus sizes, particularly for batch algorithms.

5.1 Limitations and future directions

The first direction in which we think the WordSeg suite should be improved is by providing users with solutions for phonologizing their texts, and facilitating in-

formed choices for data selection from CHILDES. Previous researchers have used a range of pre-processing pipelines, making choices that could affect segmentation results. Some researchers remove repetition or mumbling within sentences, which obscures any dependence which may have been present previously. For instance, "she xxx baby girl" would become "she baby girl" (since xxx indicates untranscribed spoken material in the CHAT format), which misrepresents the sequence of words produced by the speaker. Sometimes material tagged as non-lexical or onomatopeic (with the CHAT tags &hey and choochoo@o, respectively) are similarly deleted from the input. Some go so far as dropping words that are not part of the finite dictionary being used. Since child-directed speech will often contain onomatopeia and other forms of non-standard words, such an analytic decision unduly simplifies the task of the word segmenter. The latter problem can be removed by using a text-to-speech system (or grapheme to phoneme conversion rules in languages with transparent orthography) on all potential child input. Such systems may also help make some strides towards making the phonologized input more realistic via the application of phonological processes of e.g. assimilation and reduction.

Although not illustrated in the examples above, the package is flexible enough to allow evaluation of segmentation at linguistic levels other than the word level. For instance, some users may desire to evaluate on morphemes rather than words [2,35]. It has been previously discussed [35] that error evaluation based on the gold word standard might not be optimal when modeling infant segmentation of useful linguistic units. Evaluation on the morpheme level should also be considered, since segmenting out the constituent morphemes of a word could actually help infants acquire more lexical elements of their language [20,38]. Similarly, one can imagine extensions assessing segmentations of yet other levels of the prosodic hierarchy, such as syllables, or syntactic units, such as phrases. A somewhat related issue is how to deal with plausible segmentation errors due to undersegmentation of sequences of words that are often produced together (collocations). To avoid penalizing for these, the user can simply create a version of the gold where word boundaries are removed in high frequency phrases. These extensions are all possible and easy to implement in WordSeg, since both the preparation and the evaluation steps allow the user to provide the code used in their text as separators. However, they all require that the user has exhaustively tagged morpheme boundaries (or whatever other unit they want to evaluate). Future developments could integrate a morphological parser to help users who lack this level of annotation, perhaps building on extant open source, multilingual tools (e.g., CLAN, [28]).

All this said, most readers will agree with us that performance against the gold standard is not necessarily the ultimate goal of research on infant word segmentation. We have begun to investigate how the output of word segmentation algorithms may be related to human performance more directly. Specifically, we have been using parental reports of infant word comprehension as the variable to be predicted [22]. This code, although available from [21], has not been prepared for public re-use as extensively as the WordSeg code has. Additionally, there is considerable conceptual and methodological work needed to extrapolate the method to corpora of other languages (see [1] for a first attempt). We hope others will find ways of employing the WordSeg package output to relate word segmentation results from computational models to human performance, and similarly document and share their code.

Another conceptual development we foresee involves breaking down the currently incorporated algorithms into recombinable modules. We have opted to reuse extant algorithms to allow users to connect with previous literature. Nonetheless, as word segmentation research advances, it would be ideal to reflect on the fact that some extant algorithms represent a set of conceptual choices, each of which is potentially combinable with others. For example, PUDDLE [31] incorporates a strategy that profits from single-word utterances or chunks. In that model, utterances that have not been segmented are encoded directly into long-term memory, and later used to break up new utterances. We could imagine a model that encodes phonotactics like DiBS does (i.e., not with a list of permissible phone bigrams but rather as a probability distribution of the transition) together with a chunk memorization module as found in PUDDLE. It would also be interesting to explore parameters that have similarly been confounded with other design options, such as whether the model should treat differently phenomena occurring at utterance edges than utterance middles [42], or saliently whether the processing is batch or incremental.

Finally, the modular architecture of WordSeg as well as the fact that it is open source should facilitate its integration with other systems focusing on unsupervised learning of language structure at other levels. Recent research has begun to investigate word segmentation from raw speech [43], an interesting development given infant psycholinguistic research strongly suggesting young infants may build their earliest proto-lexicon using acoustic representations (e.g., [16]). Although there are very few public corpora of child-directed speech with phonological transcriptions that are aligned well enough to be usable for this process, some recent work has made great strides towards standardizing and facilitating forced alignment [29], including on CHILDES corpora [9,11]. As to the integration of systems working on other levels of acquisition, it would be worthwhile to explore parsers allowing the discovery of morphological structure within words (such as the open source Linguistica, see [23], section 5.2) as well as others that succeed in acquiring multi-word dependencies (and thus a form of shallow syntax, e.g., [30]).

It is not feasible for us to promise to implement all such developments. Fortunately, having opted for a modular, open-source structure makes it easy for *others* to contribute these and other algorithms. As more and more cognitive scientists and psychologists use computational modeling, more and more students and researchers will have the necessary computer skills to make contributions via the GitHub system. These users would fork our repository from `github.com/bootphon/wordseg`, add their tool in the wordseg/algos section, and then either keep this improved version in their own repositories, or do a pull request so that the standard WordSeg comes to include their tool. Notice incidentally that the use of `readthedocs.com` allows us to harvest help sections from within python code, thus inviting tool developers to include statements of use that directly become available to WordSeg users. For readers who find this idea appealing but do not have previous experience with git, we recommend the excellent introduction to git offered by Software Carpentry (`https://swcarpentry.github.io/git-novice/`, followed by GitHub's tutorials for forking (`https://help.github.com/articles/fork-a-repo/` and creating pull requests (`https://help.github.com/articles/creating-a-pull-request-from-a-fork/`). We provide further information in a dedicated section of our documentation `https://wordseg.readthedocs.io/en/latest/contributing.html#contributing-to-the-code`.

In conclusion, the present version of WordSeg greatly facilitates research on unsupervised wordform segmentation by integrating multiple previous contributions into a modular architecture. We look forward to further improvements, inviting feedback and development.

Acknowledgements This work was directly supported by the Agence Nationale de la Recherche (ANR-14-CE30-0003 MechELex) and the European Research Council (ERC-2011-AdG-295810 BOOTPHON). We also acknowledge funding from the Fondation de France, the Ecole de Neurosciences de Paris, the Region Ile de France (DIM cerveau et pensee); and the institutional support of Agence Nationale de la Recherche (ANR-10-IDEX-0001-02 PSL*, and ANR-10-LABX-0087 IEC). We benefited from code shared directly with AC by Padraic Monaghan. We also reused code stored on github by Lawrence Phillips and originally written by Sharon Goldwater, Lisa Pearl, and/or Mark Johnson. We are grateful to Melanie Soderstrom for help with the systematic review mentioned in Section 2; as well as to her and Nan Bernstein for comments on a previous version of this manuscript. Finally, we are grateful to members of the LAAC, CoML, and Language teams at the LSCP for helpful discussion.

References

1. Baudet, G.: Xlingcorrelation. https://github.com/bootphon/XLingCorrelation (2018)
2. Börschinger, B., Demuth, K., Johnson, M.: Studying the effect of input size for Bayesian word segmentation on the Providence corpus. In: Proceedings of COLING, pp. 325–340 (2012)
3. Brent, M.R.: Speech segmentation and word discovery: A computational perspective. Trends in Cognitive Sciences **3**(8) (1999)
4. Brent, M.R., Cartwright, T.A.: Distributional regularity and phonotactic constraints are useful for segmentation. Cognition **61**(1), 93–125 (1996)
5. Cai, D., Zhao, H., Zhang, Z., Xin, Y., Wu, Y., Huang, F.: Fast and accurate neural word segmentation for Chinese. arXiv preprint arXiv:1704.07047 (2017)
6. Daland, R.: Word segmentation, word recognition, and word learning: A computational model of first language acquisition. PhD, Northwestern University (2009)
7. Daland, R., Pierrehumbert, J.B.: Learning diphone-based segmentation. Cognitive Science **35**(1), 119–155 (2011)
8. Demuth, K., Culbertson, J., Alter, J.: Word-minimality, epenthesis and coda licensing in the early acquisition of English. Language and Speech **49**(2), 137–173 (2006)
9. Elsner, M., Ito, K.: An automatically aligned corpus of child-directed speech. Proceedings of Interspeech pp. 1736–1740 (2017)
10. Fourtassi, A., Börschinger, B., Johnson, M., Dupoux, E.: Whyisenglishsoeasytosegment. Proceedings of CMCL pp. 1–10 (2013)
11. Frermann, L., Frank, M.C.: Prosodic features from large corpora of child-directed speech as predictors of the age of acquisition of words. arXiv preprint arXiv:1709.09443 (2017)
12. Gambell, T., Yang, C.: Word segmentation: Quick but not dirty. Unpublished manuscript (2005)
13. Goldwater, S., Griffiths, T.L., Johnson, M.: A Bayesian framework for word segmentation: Exploring the effects of context. Cognition **112**(1), 21–54 (2009)
14. Gundersen, O.E., Kjensmo, S.: State of the art: Reproducibility in artificial intelligence. Thirty-Second AAAI Conference on Artificial Intelligence p. 17248 (2018)
15. Harris, Z.S.: From phoneme to morpheme. Language **31**(2), 190–222 (1955)
16. Houston, D.M., Jusczyk, P.W.: The role of talker-specific information in word segmentation by infants. Journal of Experimental Psychology: Human Perception and Performance **26**(5), 1570 (2000)
17. Hung, L.H., Kristiyanto, D., Lee, S.B., Yeung, K.Y.: Guidock: using docker containers with a common graphics user interface to address the reproducibility of research. PloS one **11**(4), e0152,686 (2016)
18. Johnson, M., Christophe, A., Dupoux, E., Demuth, K.: Modelling function words improves unsupervised word segmentation. In: ACL, pp. 282–292 (2014)

19. Johnson, M., Goldwater, S.: Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 317–325. Association for Computational Linguistics (2009)
20. Kim, Y.J.: 6-month-olds' segmentation and representation of morphologically complex words. University of California, Los Angeles (2015)
21. Larsen, E.: Wordseg comprehension. https://github.com/elinlarsen/WordSegComprehension (2018)
22. Larsen, E., Cristia, A., Dupoux, E.: Relating unsupervised word segmentation to reported vocabulary acquisition. Interspeech pp. 2198–2202 (2017)
23. Lee, J.L., Goldsmith, J.A.: Linguistica 5: Unsupervised learning of linguistic structure. Proceedings of NAACL-HLT 2016 (Demonstrations) pp. 22–26 (2016)
24. Lignos, C.: Infant word segmentation: An incremental, integrated model. In: Proceedings of the West Coast Conference on Formal Linguistics, vol. 30, pp. 13–15 (2012)
25. Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., Dupoux, E.: The role of prosody and speech register in word segmentation: A computational modelling perspective. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics **2**, 178–183 (2017)
26. Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.N., Johnson, M., Dupoux, E.: Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. Proceedings of LREC pp. 560–576 (2014)
27. MacWhinney, B.: The CHILDES Project part 1: The CHAT transcription format. Psychology Press (2009)
28. MacWhinney, B.: The CHILDES Project part 2: The database. Psychology Press (2009)
29. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. Proceedings of interspeech pp. 498–502 (2017)
30. McCauley, S.M., Christiansen, M.H.: Computational investigations of multiword chunks in language learning. Topics in cognitive science **9**(3), 637–652 (2017)
31. Monaghan, P., Christiansen, M.H.: Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. Journal of Child Language **37**(03), 545–564 (2010)
32. Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., Peperkamp, S.: (non) words, (non) words, (non) words: Evidence for a protolexicon during the first year of life. Developmental Science **16**(1), 24–34 (2013)
33. Pearl, L., Goldwater, S., Steyvers, M.: Online learning mechanisms for Bayesian models of word segmentation. Research on Language and Computation **8**(2-3), 107–132 (2010)
34. Phillips, L.: The role of empirical evidence in modeling speech segmentation. PhD, University of California, Irvine (2015)
35. Phillips, L., Pearl, L.: The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. Cognitive Science **39**(8), 1824–1854 (2015)
36. Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W.: The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. Speech Communication **45**(1), 89–95 (2005)
37. Saksida, A., Langus, A., Nespor, M.: Co-occurrence statistics as a language-dependent cue for speech segmentation. Developmental Science **20**(3), e12,390 (2017)
38. Shi, R., Werker, J.F., Cutler, A.: Recognition and representation of function words in English-learning infants. Infancy **10**(2), 187–198 (2006)
39. Stodden, V., Seiler, J., Ma, Z.: An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences **115**(11), 2584–2589 (2018)
40. Swingley, D.: Statistical clustering and the contents of the infant vocabulary. Cognitive Psychology **50**(1), 86–132 (2005)
41. Taylor, P., Black, A.W., Caley, R.: The architecture of the FESTIVAL speech synthesis system. Proc. 3rd ESCA Workshop on Speech Synthesis, pp. 147–151 (1998)
42. Venkataraman, A.: A statistical model for word discovery in transcribed speech. Computational Linguistics **27**(3), 351–372 (2001)
43. Versteegh, M., Thiolliere, R., Schatz, T., Cao, X.N., Anguera, X., Jansen, A., Dupoux, E.: The Zero Resource Speech Challenge 2015. Sixteenth Annual Conference of the International Speech Communication Association (2015)

Appendix C

Artificial language experiments - Complementary study to Chapter 3

This study is part of the Appendix because, even though relevant to the thesis project in Chapter 3, it was finalized after the date of the defense.

Follow-up studies for Chapter 3

In Chapter 3, we compared two languages, Chintang and Japanese, represented by two naturalistic corpora that may differ across several factors. For example, Chintang parents could by chance produce longer utterances, and this would affect our results. It is impossible to control for this in naturalistic corpora. We cannot be certain that the differences we observe are due to the specific factor that led us to choose these two languages or to any extraneous factor in the corpora and/or to other uncontrolled characteristics of the languages. We considered extending this approach to other natural child-centered corpora, for instance by looking at 10-20 corpora of languages varying in morphological complexity. This turned out not to be feasible, since morphological segmentation is typically not available in child-centered corpora. Moreover, there would always have been the possibility that uncontrolled differences caused, or obscured, any result that we were to find.

We have thus performed additional experiments with artificial languages, to study the effects of morphological complexity on segmentability in a more controlled fashion. Artificial languages allow us to study specific properties of languages and their effect under tightly controlled conditions. Once everything else was controlled for via artificial languages, the effect of morphological complexity was clear and could not be attributed to any confounds.

5 Experiment 2: Artificial languages varying in the number of affixes

In this experiment, we study whether languages varying in morphological complexity differ in segmentability by assessing segmentability of five morphologically diverse artificial languages, which exhibit a gradual range of morphological complexity. We make sure that the number of words per sentence are matched across corpora, and the languages only differ on this specific aspect of morphological complexity.

We focus on the factor of morphological synthesis, keeping all other variables stable. In order to study the effects of morphological synthesis on segmentability, we track changes in segmentation while modifying the ratio of morphemes per word. In

Experiment 1, the mean number of grammatical affixes accompanying the stem was about 1.69 for Chintang and 1.07 for Japanese. In Experiment 2, we increase the variability of this feature, ranging from 0 to 4.

The same questions asked above are revisited in this controlled experiment. First, do languages varying in morphological complexity differ in segmentability? Based on the key predictions above, languages with a smaller number of morphemes within words should be easier to segment than languages where words have multiple morphemes. Also, based on the key predictions above, both lexical and sublexical algorithms should yield lower word segmentation scores for more complex languages, as they are more likely to break up the stream at morpheme boundaries.

Second, how large is this effect, compared to differences across algorithms and evaluation level? We inquire whether performance varies as a function of algorithm (the specific algorithm employed during segmentation) and the level of linguistic representation on which segmentation is evaluated (words or morphemes). We concluded in Experiment 1 that morphology-related differences across languages were relatively small, but this conclusion could be curtailed by the fact that the range of variation covered with these two natural languages may be small. Experiment 2 allows us to better measure these effects by studying them in isolation, and increasing the range of linguistic variation covered.

5.1 Methods

5.1.1 Languages. Languages were created using a script in R. First, a set of consonant-vowel syllables were composed through every combination of the consonants "z", "r", "t", "y", "p", "q", "s", "d", "f", "g", and "h" and the vowels "a","e","i","o",and "u". We then composed a lexicon of 1,000 words. Function words constituted 1% of this lexicon, and they were always one syllable in length. The rest of the lexicon were content word stems, which varied in length between 1 and 4 syllables. Content words were randomly split into two classes, A and B (which may be thought of as nouns and verbs), and which selected affixes from two different paradigms. All of these aspects

were fixed across languages.

Languages varying in complexity thus differed only on the next step. The base language (0) had no affixes; the next language (1) had one affix per content word (with different affixes for class A and B stems); and so on, for up to 4 affixes (4). All affixes were one syllable long.

Mean subset stats	0	1	2	3	4
# utt	5000	5000	5000	5000	5000
# wtokens	12528 (55)	12528 (55)	12528 (55)	12528 (55)	12528 (55)
# wtypes	791 (8)	4695 (36)	7125 (31)	7482 (23)	7520 (21)
# whapaxes	1.5 (0.71)	2807 (43)	6733 (38)	7428 (26)	7506 (20)
# phtokens	50535 (460)	65515 (376)	80581 (525)	95570 (518)	110556 (490)
# mortokens	12528 (55)	20044 (75)	27560 (93)	35073 (114)	42582 (129)

Table 7

Corpus features: Means (and standard deviation) across the ten subsets of artificial languages 0, 1, 2, 3, and 4 (see main text for explanation). # stands for number, "utt" stands for utterance. "wtokens", "wtypes", "whapaxes" stand for word tokens, word types and word hapaxes. "phtokens" stands for phoneme tokens and "mortokens" stands for morpheme tokens.

The final step was also in common across languages, and consisted in creating a corpus of 5000 sentences that were between 1 and 4 words in length. Previous methodological work suggests algorithms' performance is stable by about 5000 sentences (Bernard et al., 2018). Sentence lengths of 1-4 seem reasonable for child-directed speech, according to previous descriptive studies (Loukatou, Le Normand et al., 2019). Sentences one word in length had only a stem (and, for more complex languages, its affixes); sentences with two words had a function word and a stem (and affixes); three-word sentences had a function word and two stems (and their affixes); and four-word sentences had a function word, a stem (and its affixes), a function word, and a stem (and its affixes). For clarity in the code, each sentence sampled from the lexicon for each language separately.

To make this more concrete, here is the first sentence in the five languages'

corpora in one run, containing always three words (a function word followed by two stems with their eventual affixes, depending on the language); words are separated by spaces, morphemes by dashes:

- **0:** "pi rotu rodezira"
- **1:** "yu so-se qofeharu-se"
- **2:** "tu yosoreda-ga-yi foyo-gi-su"
- **3:** "pi ruza-to-re-pu gori-di-re-ra"
- **4:** "fe zi-pa-yo-ye-gi ho-fa-ge-ye-te"

And the following are sample sentences containing the stem "rodezira", which was one of the stems in the lexicon in that run, appearing in sentences of the same word length across the five languages:

- **0:** "pi rotu rodezira" (3 words, 3 morphemes)
- **1:** "di rodezira-ge reyoha-qi" (3 words, 5 morphemes)
- **2:** "yi tohegipu-ga-ga rodezira-sa-yu" (3 words, 7 morphemes)
- **3:** "tu rodezira-de-ro-pa gitopide-pe-re-qu" (3 words, 9 morphemes)
- **4:** "gu rodezira-fa-ho-fu-qo deguqaso-ge-ri-re-hu" (3 words, 11 morphemes)

This whole process was repeated 10 times, to create 10 corpora, each 5,000 sentences in length, for each of the five different languages. Table 3 shows some basic statistics of these languages.

5.1.2 Segmentation. The same procedures were used as in Experiment 1.

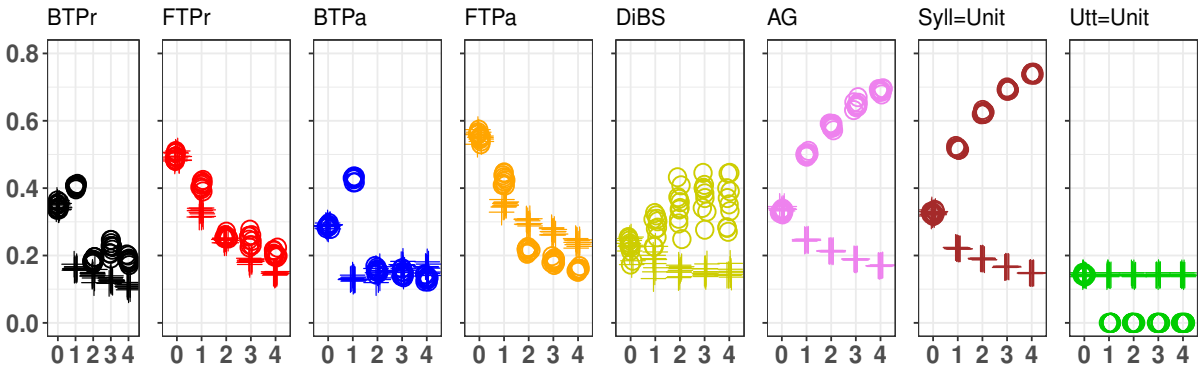

Figure 3. The y axis shows token F-scores across languages. The languages in the x axis are 0 (stems take no affixes), 1 (stems take one affix), 2 (stems take two affixes), 3 (stems take three affixes) and 4 (stems take four affixes). Evaluation levels are marked by shape (open circle for morphemes, cross for words). Color reflects algorithms (which are also used to group the data into boxes, see main title of each box).

5.2 Results and discussion

Results for Experiment 2 are shown in Fig. 3. Our first goal was to answer whether languages varying in morphological complexity, as defined by the number of morphemes per word, differ in segmentability. A regression predicting F-scores from language, level, algorithm, and their interactions accounted for most variance in the data, $R^2 = .99$ ($F(63, 568) = 720$, $p < .001$).[11] Even though the presence of significant interactions precluded a direct interpretation of the main effects, the regression confirmed a disadvantage for the most morphologically complex languages, with negative coefficients estimating the language effect (-0.05 for language 4, -0.04 for 3, -0.02 for 2). Thus, the answer to our first research question is that there are significant effects on segmentability of varying complexity across languages.

In response to our second research goal, how large the language effect is compared to differences across algorithms and evaluation level, we observed that language effects appeared to be relatively small, since the F-value for language is several times smaller than that of level and of algorithm . More detailed outcomes are provided in the online

[11] The function was: lm(token fscores ∼ language ∗ level ∗ algorithm + (1/file), subsets). Token F-scores are the F-scores to be predicted by language, level, and algorithm as fixed effects, and subset as random factor. The data frame contains 640 observations (4 languages x 2 levels x 10 subsets x 8 algorithms).

supplementary material, but two aspects of the results apparent in Fig. 3 are worth pointing out. Morpheme level scores for languages where words contain two, three, four and five morphemes, were in general higher than word level results for the same language. We also observed some interactions. For example, AG and DiBS morpheme-level results increased with language complexity, reaching and even surpassing the results for the language with no affixes.

In sum, similarly to what was found in Experiment 1, we observed the expected differences in performance as a function of language, level, and algorithm type. The results of Experiment 2 support our conclusions from Experiment 1: Languages varying in morphological complexity vary in segmentability. Overall *word* segmentation performance for the simplest language where words and morphemes coincide was better than performance for the other languages, which contain 1-4 affixes per stem. However, the strength of the language effect varied across algorithms, and was even reversed in some conditions, exactly as we observed in Experiment 1. The language effect was again smaller than the effects found for the other two factors, namely level and algorithm type, even though the range of variance here was huge, much larger than that found in Experiment 1's natural language corpora.

6 Experiment 3: Artificial languages varying in the distribution of affix number

We implemented one more set of artificial languages in order to observe the effect of morphological complexity in controlled environments. One limitation of Experiment 2 is that languages were more internally homogeneous in terms of complexity than human languages typically are: There is no human language in which each and every content word in the language must always have exactly three affixes. We relaxed this assumption while maintaining differences in complexity in our Experiment 3. Specifically, the languages in this experiment were created to differ in the distribution of affix numbers, with all artificial languages having words that contain between zero and four affixes but varying in how frequent different affix numbers were. In our baseline

language, the probability distribution was flat, with 20% probability for each option (zero to four affixes). In a simpler language, more mass was allocated to lower number of affixes. Finally, a more complex language was created with more mass allocated to higher number of affixes.

The same questions and predictions given in Experiments 1 and 2 are revisited in this experiment: We ask whether languages varying in morphological complexity differ in segmentability, and how large this effect is when compared to differences across algorithms and evaluation level.

6.1 Methods

Mean subset stats	S	B	C
# utterances	5000	5000	5000
# word tokens	12470 (106)	12518 (69)	12532 (64)
# word types	5336 (47)	6481 (43)	7174 (25)
# word hapaxes	4467 (51)	5903 (65)	6922 (27)
# phoneme tokens	70240 (676)	80574 (809)	90452 (645)
# morpheme tokens	22576 (141)	27598 (233)	32452 (195)

Table 8

Corpus features: Means (and standard deviation) across the ten subsets of artificial languages S (simpler), B (base) and C (more complex). # stands for number.

6.1.1 Languages. As in Experiment 2, we created languages with a lexicon of 1,000 items, of which 1% were one-syllable long function words, and the remaining were stems one- to four-syllables in length, randomly split into two types that selected different affix paradigms. The syllable inventory, distribution of sentence length (1-4 words), length of corpora (5,000 sentences), were also kept constant, and 10 subsets were generated for each language.

Unlike in Experiment 2, however, all languages had some affixes, meaning that stems could take between 0 and 4 affixes. The three languages we created varied in terms of the *distribution* of the number of affixes a stem took. In the base language (B),

it was equally likely for stems to have 0 to 4 affixes (i.e., 20% of chances for each). In the simpler language (S), the distribution was tilted towards fewer affixes: 35% likelihood of having 0 affixes, 25% of having 1, 20% of having 2, 10% of having 3, and 10% of having 4 affixes. The more complex language (C) had the opposite trend: 10% likelihood of having 0 affixes, 10% of having 1, 20% of having 2, 25% of having 3, and 35% of having 4 affixes. Table 4 shows some basic statistics of the languages.

6.1.2 Segmentation. The same procedures were used as in Experiment 1.

6.2 Results and discussion

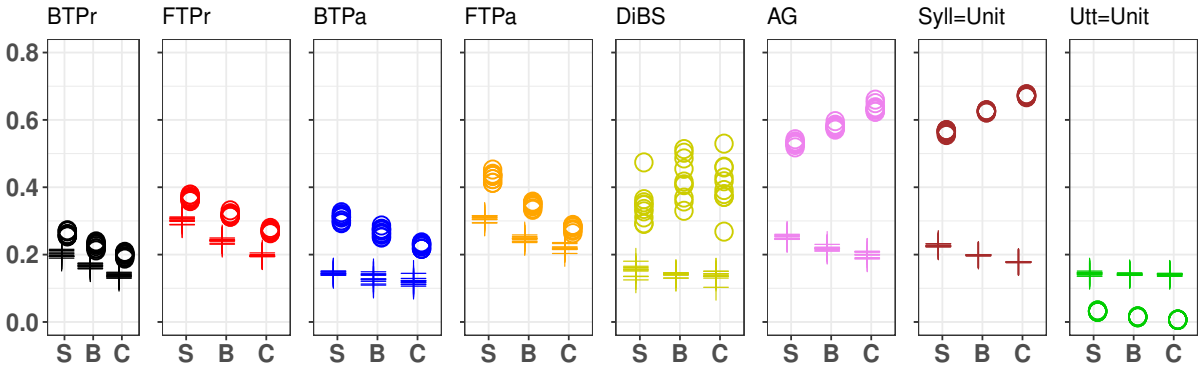

Figure 4. The y axis contains token F-scores across languages. The languages in the x axis are S(imple), B(ase) and C(omplex). Evaluation levels are marked by shape (open circle for morphemes, cross for words). Color reflects algorithms (which are also used to group the data into boxes, see main title of each box).

Results for Experiment 3 are shown in Fig. 4. In general, more similarities in segmentability across languages were found here than in Experiment 2. This may be due to the fact that the languages were more similar to each other here than in Experiment 2.

Bearing on our first research question, a regression predicting F-scores from language (S, B and C), level, algorithm, and their interactions accounted for most variance in the data, $R^2 = .99$ ($F(47, 945) = 432$, $p < .001$).[12] Even though the presence

[12] The function was: lm(token fscores ~ language * level * algorithm + (1/file), subsets). Token F-scores are the F-scores to be predicted by language, level, and algorithm as fixed effects, and subset as random factor. The data frame contains 480 observations (3 languages x 2 levels x 10 subsets x 8 algorithms).

of significant interactions precluded a direct interpretation of the main effects, the regression confirmed a disadvantage for the more morphologically complex languages, with negative coefficients estimating the language effect (-0.06 for C and -0.03 for B).

Regarding our second research goal, we observed that language effects are relatively small, since the F-value for language is half the size of level and a quarter of that for algorithm . More detailed outcomes are provided in the online supplementary material, but two further aspects of the results are worth pointing out. As in Experiment 2, scores for all three languages improved when evaluating on the *morpheme* level across algorithms. Also, similarly to Experiment 2, AG and DiBS morpheme scores increased with language complexity.

7 Summary

In the context of our artificial languages, where languages differ maximally in morphological complexity (more than what any human natural languages could differ), the effect of morphological complexity remained small, and certainly smaller than that of level and algorithm. The artificial language results were also informative on the general performance of the algorithms; the performance range of the algorithms was similar across all three experiments, highlighting the relevance of our artificial language results for broader generalization to natural languages. Additionally, this suggested that differences across algorithms are massive – even when corpora are perfectly controlled.

RÉSUMÉ

La langue est acquise par les enfants du monde entier, mais en fonction de l'input fourni, l'acquisition a probablement des différentes voies de développement, des rythmes différents et des résultats variables. Dans ce manuscrit, nous examinons de plus près l'étonnante diversité d'input que les enfants grandissent en entendant, et nous demandons en quoi cette diversité est importante pour l'acquisition du langage. Pour cela, nous utilisons des méthodes hautement interdisciplinaires. Nous considérons le type de langue et de culture comme deux sources principales de diversité, et nous les étudierons dans deux parties distinctes de cette thèse.

MOTS CLÉS

Diversité ; capacité d'apprentissage ; acquisition précoce des enfants ; langue ; culture

ABSTRACT

Language is acquired by children all around the globe, but probably along different developmental paths, at varying rates and with varying outcomes, depending on the input provided. In this dissertation, we take a closer look at the astonishing diversity of input children grow up hearing, and we ask how this diversity matters to language acquisition. For this, we employ highly interdisciplinary methods and involve several projects. We consider the type of language and culture as two principal sources of diversity, and we will investigate them in two distinct parts of this dissertation.

KEYWORDS

Diversity ; learnability ; language acquisition ; cross-linguistic ; cross-cultural