The Role of Chunking and Analogy in Early Vocabulary Acquisition and Processing

A thesis submitted for the degree of Doctor of Philosophy

by

Francesco Cabiddu

School of Psychology, Cardiff University, October 2023

Thesis Summary

Chunking and analogy, learning through associations and similarities respectively, are crucial cognitive processes in a usage-based theory of language development. Assessing their roles in child naturalistic word learning has posed significant challenges. In this thesis, I offer methodological solutions to examine the developmental plausibility of these processes. Chapter 2 discusses limitations in studies of early word segmentation from naturalistic speech, affecting conclusions about the processes' developmental plausibility. I present a new chunking-based model, CLASSIC Utterance Boundary (CLASSIC-UB), to study how English infants discover words from continuous naturalistic speech. Its plausibility is assessed through new metrics focusing on child production vocabularies from large-scale conversational corpora. I show the advantages of using large word production samples and how this can improve the refinement of early word segmentation and learning theories. In Chapter 3, conclusions about CLASSIC-UB's plausibility are supported by extending this approach cross-linguistically, using Italian as a case study. Across Chapters 2 and 3, CLASSIC-UB more accurately captures child productions than other chunking and non-chunking accounts, supporting its plausibility in early word segmentation and learning. In Chapter 4, I identify methodological challenges in assessing the independent effects of chunking and analogy in child word processing. I focus on how children use sentence context to resolve ambiguous word meanings (word sense disambiguation). I present ChiSense-12, a new open-access sense-tagged corpus of child-directed speech, and describe its use in creating experimental stimuli to disentangle variables (verb-object associations and verb-event structures) that are informative about the independent role of chunking and analogy. Using this corpus, I showed - for the first time - that 4-year-old children exploit both bottom-up verb-object associations and top-down verb-event structures to resolve lexical ambiguities. Overall, this thesis makes a significant contribution to usage-based theories of language development and improves our understanding of how children acquire language in real-life contexts.

Declaration and Statements

Statement 1 This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Francesco Cabiddu, 18/10/2023

Statement 2

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently for any other degree or award (outside of any formal collaboration agreement between the University and a partner organisation)

Francesco Cabiddu, 18/10/2023

Statement 3

I hereby give consent for my thesis, if accepted, to be available in the University's Open Access repository (or, where approved, to be available in the University's library and for inter-library loan), and for the title and summary to be made available to outside organisations, subject to the expiry of a University-approved bar on access if applicable.

Francesco Cabiddu, 18/10/2023

Declaration

This thesis is the result of my own independent work, except where otherwise stated, and the views expressed are my own. Other sources are acknowledged by explicit references. The thesis has not been edited by a third party beyond what is permitted by Cardiff University's Use of Third Party Editors by Research Degree Students Procedure.

Francesco Cabiddu, 18/10/2023

Contents

Acknowledgements

I am grateful for the support I received from the School of Psychology at Cardiff University and for the help from many people before and during my PhD journey.

I would like to express my profound gratitude to all my supervisors for helping me grow as a researcher.

Chiara, I cannot thank you enough for your generosity in providing extremely prompt, clear, and detailed feedback about my work. Your encouragement to always keep an eye on both the immediate and distant future of my research has significantly influenced my approach to my PhD. Your guidance will continue to influence my work in the future.

Lewis, thank you for challenging my ideas in ways that have significantly improved my writing and the framing of my research questions. You helped me consider alternative approaches and broaden the scope of my work. I appreciate your help in progressing through my PhD. Your straightforwardness and sense of humour have boosted my morale and energized me to continue on my path.

Gary, I have been incredibly lucky to have you as a mentor. Thank you for the countless opportunities you provided for me to learn about doing research. Without your support, I would never have been able to embark on this journey. Your passion for research and your innovative ideas have inspired me. You have shown me how exciting and enjoyable research can be.

Thank you to Richard Morey, who has helped me as an independent reviewer during my PhD annual review meetings. His insights have been invaluable in improving my understanding of the contributions of computational modelling to psychological research, and in reflecting on the importance of ensuring that study designs and statistical methods are well-suited to effectively answer the research questions of interest.

I would also like to dedicate a moment to thank Bill, who instilled confidence in me at the start of my PhD journey. Though our time together was brief, Bill's expertise and attitude profoundly impacted me.

Thank you to my fellow PhD students, Kelsey and Adelina, who have been great companions on this journey.

Infinite thanks to my parents, Giuseppe and Vincenza, who supported me with all their strength and resources. They taught me to be hardworking, generous, and humble. Thanks to my brother Federico, for being a great friend and an example of strength and determination.

Thank you, Francesca. None of this would have meaning without your presence. You've taught me that it's possible to desire more from life.

Thanks to the friends who were closest to me during my PhD, Michele and Riccardo. You helped me not to take life too seriously, bringing lightness to my daily life.

Thank you to my dogs, Attila and Buddy. Your vulnerability and sweetness have made me more understanding and loving towards others.

Lastly, I would like to thank Gianmarco Altoè and Davide Massidda for being my initial guide and inspiration in academia. Thanks also to Claudio and Laura for your friendship, for inspiring me with your resourcefulness, and your curiosity in studying, with a hunger for deeply understanding what surrounds us.

Preface

Each chapter of this thesis is presented as an empirical article in line with peer-reviewed journals. Every chapter begins with a review of relevant literature, continues with method and results sections, and concludes with a discussion of the results. I adopted this format to facilitate the publication of this research in scientific journals. For consistency, I have used the same font and style throughout the thesis and provided a single References and Appendix sections at the end.

As of this writing, Chapter 2 has been published as a regular article in the journal *Language Learning* (https://onlinelibrary.wiley.com/doi/full/10.1111/lang.12559). Chapter 3 is currently being prepared for journal submission. Chapter 4 includes an empirical study, also being prepared for journal submission, and has already been published as a conference article in the *Proceedings of the Annual Meeting of the Cognitive Science Society 2022* (https://escholarship.org/uc/item/9kh29212). Additionally, within Chapter 4, I introduced a new corpus of child-directed speech named ChiSense-12, manually annotated for word senses. This corpus has been made available as an open-access tool and is published in the *Proceedings of the Thirteenth Language Resources and Evaluation Conference 2022* (https://aclanthology.org/2022.lrec-1.557).

I would like to thank Kelsey Frewin, Chara Sofocleous, Holly Martin, and Jennifer Lloyd for their contributions to the research detailed in Chapter 4. Specifically, Kelsey Frewin assisted with the recording of stimuli for the experimental task. Chara Sofocleous helped as a second independent annotator for the corpus inter-annotator agreement study, and both Holly Martin and Jennifer Lloyd assisted with child data collection.

Chapter 1
General Introduction

1.1 Introduction

This thesis examines the role of chunking (i.e., the ability to learn from associations, e.g., Gobet, 2017; Gobet et al., 2001) and analogy (i.e., the ability to learn schemas based on similarities across exemplars, e.g., Abbot-Smith & Tomasello, 2006; Bybee, 2010a; Ibbotson et al., 2012) in understanding how children leverage naturalistic speech to acquire and use their early vocabularies. I tested whether a chunking-based learning mechanism can accurately model how children progress from identifying word forms in naturalistic child-directed speech to building their early production vocabularies in both English and Italian. Additionally, I investigated the combined role of chunking and analogy by assessing whether the performance of 4-year-old English-speaking children in word sense disambiguation depends on their sensitivity to verb-object associations found in naturalistic child-directed speech, as well as their ability to apply known verbs to new objects.

The influential usage-based theory of language development assumes that children's linguistic knowledge emerges from the recurrent application of domain-general cognitive processes of chunking and analogy to the language events they encounter (Behrens, 2009; Bybee, 2010b; Tomasello, 2000, 2003, 2009). This theory posits that children can gradually attain adult-level linguistic competence by bootstrapping their linguistic knowledge from naturalistic input, using processes that are applicable across various domains (i.e., domain-general). This contrasts with an approach to language development that assumes the need for innate language knowledge (e.g., Valian, 2015). Therefore, a fundamental assumption of a usage-based theory is that children's learning is influenced by their language experiences. This has made large conversational corpora valuable resources for investigating whether the input contains sufficient information for children to acquire specific linguistic knowledge. These datasets also allow, via computational experiments, for the examination of which learning mechanisms can effectively make use of the

information available in naturalistic input to acquire linguistic knowledge. Although previous work has produced a significant body of evidence supporting the role of chunking and analogy in children's language development (e.g., Behrens, 2021; Lieven, 2016), one of the challenges for researchers has been determining whether learning mechanisms, when applied to naturalistic input, can actually capture children's outcomes in real-world settings. In Chapter 2 and 3, I aimed to address this by examining if a chunking-based learning mechanism applied to child-directed speech could capture various distributional aspects of children's production vocabularies measured from naturalistic conversations. Additionally, another challenge has been assessing whether learning mechanisms can explain the influence of naturalistic language experiences on children's word processing. To address this, in Chapter 4 I explored how chunking and analogy might independently explain how children's naturalistic language experiences affect their ability to resolve lexical ambiguities in the lab. Specifically, I investigated whether children's chunking of frequent verb-object associations from child-directed speech and their generalizations from verb knowledge might help them disambiguate words with multiple meanings.

In Study 1 (Chapter 2), I introduced a new computational model for early naturalistic word segmentation called CLASSIC Utterance Boundary (CLASSIC-UB). This model employs a chunking-based learning mechanism to process a large corpus of speech directed at English-speaking 2-year-old children. The model's performance is benchmarked against both a baseline and other influential models that implement different hypotheses about how children might identify word forms from continuous, naturalistic speech. Importantly, CLASSIC-UB's performance is assessed not only using traditional metrics that focus on the accuracy of word segmentation, but also through a new set of measures that relate the model's performance with children's production vocabularies found in the corpora. These new measures assess whether the model can capture the age at which children first produce a word, as well as word-level characteristics that account for a significant proportion of the variance in children's production vocabularies (word frequency, word length, neighbourhood density, and phonotactic probability). The aim of this study was to test whether a chunking-based learning mechanism, which has previously demonstrated high

accuracy in segmenting naturalistic speech, can also acquire a vocabulary that aligns with what children actually produce in naturalistic conversations.

In Study 2 (Chapter 3), I built upon the work of the first study by exploring the cross-linguistic applicability of CLASSIC-UB, using speech directed at Italian-speaking children aged 16 to 36 months as a case study. Key differences between English and Italian provided an opportunity to address limitations of the first study and to answer a new set of questions. Specifically, I investigated whether the findings of the first study could be generalized to Italian, a language that contains longer words than English. This helped to test whether the first study's results were dependent on English being relatively easier to segment due to its shorter average word length (which might favour certain learning mechanisms over others). The richer morphology of Italian allowed me to examine whether a chunking mechanism could also capture the emergence of morphological units alongside word forms as found for Italian children, also testing the idea that usage-based learners acquire representations at multiple levels of granularity. Lastly, Italian child-directed speech is notable for having a higher proportion of verbs compared to nouns, even though Italian children's vocabularies still contain more nouns than verbs (known as the "noun advantage"). This offered a chance to evaluate whether a chunking model operating on naturalistic speech could account for the noun advantage observed in Italian children's speech, even though verbs are more common in the child-directed input.

In Study 3 (Chapter 4), I explored whether 4-year-old children can use analogies to apply a known verb-event structure to an object previously unassociated with that structure, and ultimately disambiguate the meaning of the ambiguous noun object (e.g., "twist the [music/elastic] band"). The experimental stimuli for this study were developed after manually annotating all English child-directed utterances in the CHILDES database (MacWhinney, 2000). This annotation process allowed me to extract verb-object co-occurrences between verbs and noun meanings. The aim was to create stimuli that could disentangle the effects of frequent verb-object associations—encountered by children in naturalistic conversations—from the semantics of the verbs themselves. The ultimate goal was

to test whether children are sensitive to the associations present in naturalistic input, and more importantly whether they can generalize a known verb-event structure to a previously unassociated noun meaning.

This introductory chapter sets the stage for the subsequent empirical chapters. The next two sections summarize existing evidence and identify current gaps in research concerning the role of chunking in early naturalistic word segmentation and the use of verb-event structure analogy in early word sense disambiguation. The concluding section offers a summary of the research presented in this thesis, with an emphasis on its originality and contributions to the current body of knowledge in the field.

1.2 Chunking as a Process for Infant Naturalistic Word Segmentation

Infants' initial linguistic communication is closely linked to pre-verbal joint-attention events they share with their caregivers (Tomasello, 2009). During these events, infants learn that linguistic inputs (e.g., utterances) serve distinct communicative intentions, such as directing the interlocutor's attention to a particular referent. The efficiency of speech as a communicative tool increases as infants begin to realize that utterances are combinations of words. This understanding eventually benefits their comprehension and use of the morphosyntactic and semantic aspects of the language (Tomasello, 2003). Since speech input is presented as a continuous stream of sounds, a critical early developmental task for infants is to determine which parts of this stream correspond to individual word forms (e.g., Newman et al., 2016). Various hypotheses have been put forward regarding the learning mechanisms that might help infants tackle this word segmentation task (e.g., Daland & Pierrehumbert, 2011; French et al., 2011; Goldwater et al., 2009; Monaghan & Christiansen, 2010; Perruchet & Vinter, 1998; Saffran, Aslin, & Newport, 1996; Swingley, 2005). A prominent hypothesis suggests that infants may begin to recognize "chunks" of speech, which are sequences of sounds that appear frequently in the language and that become discernible to the child as distinct, familiar units during speech processing (e.g., French et al., 2011; Monaghan & Christiansen, 2010; Perruchet &

Vinter, 1998). Chunking is a domain-general cognitive process where events that are often associated in the environment gradually become represented as whole units, leading to more fluent processing (e.g., Gobet, 2017; Gobet et al., 2001). In word segmentation, chunking predicts the performance of both infants (e.g., French et al., 2011; Perruchet & Vinter, 1998) and adults (e.g., Endress & Langus, 2017; Frank et al., 2010) at segmenting artificial languages in laboratory settings. Various computational studies have also demonstrated that child-directed speech in naturalistic settings can be accurately segmented into words using a chunking-based learning mechanism (e.g., French et al., 2011; Monaghan & Christiansen, 2010).

However, the benchmark for models' segmentation accuracy of naturalistic speech has been based on word boundaries found in adult vocabularies (e.g., Daland & Pierrehumbert, 2011; Monaghan & Christiansen, 2010). These boundaries may not accurately reflect the segmentation patterns of infants and children. Specifically, standard evaluation metrics examine how accurately a computational model identifies the white spaces used as separators in orthographic transcriptions of speech. The underlying assumption of these metrics is that infants segment speech in a manner similar to adults. Yet, infants' early representations include not only words but also phonotactically legal nonword sequences (e.g., Ngon et al., 2013) and short multi-word combinations (e.g., Skarabela et al., 2021). This suggests that a developmentally plausible model might not necessarily be one that identifies a high percentage of word forms from the input, given that the initial lexicon comprises a diverse range of phonological sequences (e.g., Larsen et al., 2017). Assessing developmental plausibility is challenging, mainly because we do not know the proportion of word forms that infants actually segment from the input in naturalistic environments.

One approach to tackling this issue involves using word productions from naturalistic settings as an indicator of segmentation performance, supported by evidence that vocabulary acquisition is influenced by word segmentation (e.g., Estes et al., 2007; Hay et al., 2011). CLASSIC, a chunking-based computational model of vocabulary learning, has been shown to account for a significant proportion of the variance in English child production vocabularies (Jones et al., 2021). Specifically,

the model has simulated vocabulary growth trajectories based on the increased processing fluency derived from repeated exposure to phonological sequences in naturalistic input. However, a fundamental assumption in CLASSIC is that children already know where most word boundaries are in speech. Past simulations were more focused on aspects of vocabulary acquisition rather than the transition from initially identifying word forms within continuous speech to progressively establishing a vocabulary. Given the connection between word segmentation and word learning, I hypothesized that modifying CLASSIC for naturalistic word segmentation would lead to segmentation performance levels above chance. In Chapter 2, I evaluate this hypothesis by developing CLASSIC-UB–a version of CLASSIC that performs word segmentation by learning chunks that combine phonological sequences with utterance-boundary information–and comparing it against a selection of models.

All models were evaluated based on both word segmentation and word learning criteria. I pinpointed key variables that account for a significant proportion of the variance in word learning. These include the age of first production, as estimated from child productions (Grimm et al., 2017; Smolík & Filip, 2022), and specific distributional characteristics of child vocabularies, namely word frequency, word length, neighbourhood density, and phonotactic probability (e.g., Stokes, 2010, 2014; Storkel, 2009). Leveraging these measures allowed me to assess the developmental plausibility of segmentation models. Indeed, given the established relationship between word segmentation and word learning, the assumption is that a developmentally plausible model would more accurately segment words that children tend to produce earlier. Furthermore, the distribution of words segmented by a developmental plausible model based on various characteristics (e.g., word frequency) should mirror the distribution found in child production vocabularies.

In Chapter 3, I examined the cross-linguistic plausibility of CLASSIC-UB. Various computational investigations have been conducted to examine how different segmentation mechanisms perform across languages (e.g., Caines et al., 2019; Fourtassi et al., 2013; Gervain & Guevara Erra, 2012; Phillips & Pearl, 2014; Saksida et al., 2016). In general, chunking has been shown to segment naturalistic speech with high accuracy across multiple languages (e.g., Caines et al., 2019). However,

performance variability has been observed based on different variables, such as average word length in a language (Caines et al., 2019) and its morphological complexity (e.g., Phillips & Pearl, 2014). One open question involves the relation between the variation in segmentation performance exhibited by chunking-based models and actual child developmental data. It remains unclear whether a model that segments with less accuracy in one language will also necessarily be less developmentally plausible (i.e., showing a worse fit to child production vocabularies), or whether lower segmentation accuracy simply reflects variations in specific language properties, which in turn influence children's segmentation and vocabulary learning. Chapter 3 fills this gap in a unique way, by investigating how cross-linguistic differences in segmentation performance relate to model developmental plausibility, using Italian as a case study. Italian child-directed speech notably differs from English, being characterized by a longer average word length (e.g., Saksida et al., 2016) and higher morphological complexity due to its more extensive inflectional paradigm system (e.g., Tardif et al., 1997). Hence, using Italian child-directed speech as a case study allowed me to explore how key language-specific variation identified in prior cross-linguistic studies relates to developmental plausibility.

1.3 Analogy as a Process for Using Verb-Event Structures in Early Word Sense Disambiguation

Analogy is a domain-general process that is ubiquitous in child learning (e.g., Christie & Gentner, 2010; Ferry et al., 2010; Gentner, 2003; Silvey et al., 2023). It involves identifying a relational structure (a category) that encodes similarities and differences between representations. Categorization can boost child learning as it allows for generalizations (e.g., Christie & Gentner, 2010; Waxman & Markow, 1995). For example, a child might attempt to dress up their dolls after the caregiver has helped them dress, because they have recognized a similar relational structure between the action of the adult dressing them and themselves dressing another entity. Such generalizations enable the child to understand and creatively engage with novel situations that share similarities with previously encountered experiences. Similarly, under a usage-based approach of language development, analogy refers to

the use of novel items in a known linguistic construction, and it is the fundamental mechanism by which the individual comes to use the language productively (Bybee, 2010a).

An example of how analogy may be used in language development involves children's grammatical generalizations centred around verbs (Tomasello, 1992, 2003, 2009). Children gradually accumulate experiences with word combinations that share common elements (e.g., "mummy kissed daddy", "mummy kissed the baby"), and these representations of short word combinations shift towards partial productivity (e.g., "mummy kissed [KISSEE]"). This productivity arises because the child creates a representational slot in the verb construction ([KISSEE]), forming a schema based on similarities across objects previously encountered in that verb construction (Ambridge & Lieven, 2015). This schema essentially allows the child to extend the construction to items previously unassociated with it, provided they fit semantically within the schema (e.g., "mummy kissed grandma").

It is important to note that this usage-based approach, which sees analogy as central to linguistic productivity, is not the only perspective. In contrast, alternative nativist approaches propose that children have some innate knowledge of the components that make up the argument structure of verbs (e.g., Gleitman & Gillette, 1995; Pinker, 1994a). For instance, a child might inherently understand that the verb "kiss" requires a noun argument, and that the verb's object is the patient of the action. The child might also possess innate knowledge about some semantic characteristics that constitute a plausible patient (e.g., a patient of "kiss" is likely an animate entity). This implies that the child may not be generalizing (at least not entirely) from previous experiences with object arguments but rather applying a known semantic rule that essentially constrains the types of novel arguments that can fill the slot. I will return to this alternative approach in the General Discussion. However, throughout the thesis, I focus on evidence supporting the role of analogy, thereby examining the extent to which the usage-based approach can explain children's early word processing without assuming innate biases.

Previous research has provided evidence for the early use of analogy (e.g., Ibbotson & Tomasello, 2009); however, the independent contributions of analogy

and chunking in early child language learning remain unclear. This issue stems from the difficulty in assessing how much of children's knowledge in experimental tasks is derived from generalizations versus rote-learned chunks stored in long-term memory as whole units. In fact, children initially start representing speech using unanalysed chunks, meaning that grammatical aspects of language are not accessed. For example, a large proportion of the word combinations that children produce in their first year are frozen phrases, where the component words have never appeared in isolation in previous child productions (e.g., Bannard et al., 2009; Lieven et al., 1992). These word combinations can be learned via chunking from the input, with frequent word combinations having stronger representations, which in turn are accessed more fluently (e.g., Bannard & Matthews, 2008). This suggests that to test for the role of generalizations in early language development, researchers need to control for the word combinations children have likely encountered in their past experiences. These rote-learned word combinations might facilitate linguistic processing without necessarily tapping into analogical reasoning for comprehension.

This challenge of disentangling the role of chunking and analogy is evident in studies of early word processing (e.g., Andreu et al., 2013; Mani et al., 2016). For instance, 2-year-olds anticipate both typical and atypical upcoming objects of verbs with similar speed in a visual setting (Mani et al., 2016). In the experiment, children were presented with images on a screen, such as a "book" and "cheese", under different conditions. In one scenario, they heard "read a…" while being shown a "book" (an appropriate-*typical* object) and "cheese" (an inappropriate object). In another, they were shown both a "letter" (appropriate-*atypical* object) and "cheese" (inappropriate object). Before the object's label was spoken, children quickly shifted their eye gaze to the appropriate referent. Notably, the speed of this anticipatory gaze was predicted by their productive vocabulary size, irrespective of the association strength (typicality) between the verb and its object. These findings suggest that children leverage their understanding of the semantics of verbs and objects (verb-event structure) to make predictions, rather than relying solely on the frequency with which specific verbs and objects appear together (frequent verb-object chunks). However, a methodological problem with this type of experiment is that verb-object associations were defined through association norms or ratings

sourced from adult participants. Such norms might not accurately reflect the associations actually available to children in their linguistic environments. In other words, in Mani et al. (2016), children might have encountered "reading a letter" in their previous linguistic environment, thereby possibly receiving facilitation from stored chunks during the experimental task.

In other research areas examining early ambiguous word processing, the roles of chunking and analogy have been entirely confounded (Hahn et al., 2015; Rabagliati et al., 2013). For instance, some studies have investigated children's abilities to use verbs to disambiguate the meanings of ambiguous words (e.g., "bat" as in an animal or a racket). Compared to unambiguous word processing, employing ambiguous words has the additional advantage of controlling for any facilitatory effects stemming from the phonological characteristics of the target word itself (i.e., the phonological word form remains constant, forcing the child to solely process the alternative meanings mapped onto the word form). However, these studies have not examined which specific mechanism—chunking or analogy—might have influenced children's disambiguation performance. For example, when verbs precede them, 4-year-old children can effectively disambiguate words with multiple meanings, as in "swing the [animal/racket] bat" (Rabagliati et al., 2013). However, the roles of chunking and analogy in this context remain unclear. This is because the frequent co-occurrence of a verb with a specific meaning might cause a sequence like "swing+the+[racket]bat" to be largely rote-learned. In such cases, a child may not be drawing from their abstract understanding of the verb (i.e., a bat being an object that can be swung, similar to swords or hammers). To give another example, in the utterance "Karl met the star", The verb "meet" is likely to co-occur more frequently with "star" in the context of a famous person rather than an astronomical object. At the same time, it is more plausible to "meet" an animate entity than an inanimate one (Hahn et al., 2015). Therefore, previous work in early word sense disambiguation has not differentiated between the role of chunking, which increases sensitivity to word associations, and analogy, which allows the individual to judge the semantic fit of verb arguments based on prior knowledge.

In Chapter 4, I selected word sense disambiguation as a case study due to the lack of evidence concerning the roles chunking and analogy might play in children's processing. I also addressed previous limitations in distinguishing the contributions of these cognitive mechanisms, as emerged in studies of both early unambiguous and ambiguous word processing. Similar to these prior studies, I focused on the influence that verb-object associations and verb-event structures might have on children's performance. One method to differentiate the roles of chunking and analogy in early word sense disambiguation involves using naturalistic corpora of conversations. These can be used to create experimental stimuli that disentangle verb-object co-occurrences from verb-event structures. A significant limitation, however, is the lack of corpora containing child-directed speech tagged for word senses. Presently, only corpora of adult-written or spoken material are accessible (e.g., Pasini & Camacho-Collados, 2020), and they are unsuitable for answering questions about child processing. This is because adult-directed input differs from child-directed input in various aspects (e.g., Saxton, 2009). Hence, in Chapter 4, I introduce the first corpus of child-directed speech tagged for word senses, named ChiSense-12. This resource can be used to answer questions about the role of naturalistic variables in early word sense disambiguation. This comprehensive corpus tags word senses for 12 ambiguous words and includes annotations for instances where these words serve as direct objects of verbs. All utterances of English child-directed speech from the CHILDES database (MacWhinney, 2000) have been tagged, making ChiSense-12 a valuable reflection of the naturalistic input variables that English-speaking children may encounter in their linguistic environment.

Leveraging ChiSense-12, I designed an experimental task to distinguish the roles of chunking and analogy in 4-year-old children's word sense disambiguation. Children participated in a web-based forced-choice task during which they listened to stories ending with a target ambiguous word. Images were also displayed, representing two alternative meanings of the ambiguous word alongside semantic distractors. The stories were constructed to isolate the impact of verb-object associations and verb-event structures. In one condition, selected verbs with a neutral verb-event structure were used (e.g., "She saw the [animal/food] chicken").

Essentially, these verbs were compatible with both potential meanings of the target word. Importantly, these verbs were frequently associated with only one of the meanings in ChiSense-12 (e.g., chicken as the animal), thereby testing for the independent effect of verb-object associations. In a different condition, verbs that had never co-occurred with either meaning in ChiSense-12 were chosen, controlling for verb-object associations. Crucially, only one meaning served as a plausible argument for the verb (e.g., "She rescued the [animal] chicken"), testing for the independent effect of verb-event structures.

In sum, Chapter 4 provides a unique opportunity to evaluate the role of naturalistic verb-object associations. This analysis sheds light on the potential influence of a chunking learning mechanism in early word sense disambiguation. Concurrently, by investigating whether children can apply known verbs to word senses that never appear alongside those verbs in child-directed speech (yet remain semantically appropriate), Chapter 4 examines the role of analogy in early word sense disambiguation.

1.4 Summary of Research, Originality, and Contributions to Knowledge

The three studies presented in this thesis are conceptually complementary, each offering a different approach to explore how naturalistic language experiences impact child development. As introduced above and discussed in detail throughout the thesis, prior research has presented various computational specifications for the learning mechanisms that might be involved in child word segmentation (Chapter 2). However, an open question is whether the accuracy of these mechanisms in segmenting words from naturalistic speech would result in developmentally plausible vocabularies, similar to those produced by children. I introduce a new approach to evaluate which computational specification most accurately captures the production vocabularies of English-speaking children in naturalistic settings. This provides a valuable method for model comparison and has the potential to significantly advance the field by assessing the developmental plausibility of the proposed mechanisms. Furthermore, I demonstrate that a method focused on capturing children's

naturalistic data can shed light on the strengths and limitations of these hypothesized learning mechanisms, particularly when augmented by cross-linguistic performance comparisons (Chapter 3).

Additionally, prior research on word sense disambiguation has been limited by the presence of confounding variables. These limitations have hindered our ability to assess the potential contribution of chunking and analogy mechanisms to child performance. I demonstrate that constructing experimental stimuli based directly on what children hear in naturalistic conversations (Chapter 4) offers a viable method to provide empirical support for the hypothesized learning mechanisms believed to operate on such naturalistic input. Creating stimuli that reflect what is found in naturalistic speech also stands as a compelling test of a usage-based theory, which suggests that language experiences shape cognitive representations.

As outlined in more detail in the General Discussion, the three studies carry significant implications for future research. The introduction of a new segmentation model, CLASSIC-UB, and new evaluation metrics could encourage the use of real-world data for model comparison in subsequent studies, thereby enhancing the ecological validity of the findings. These investigations may also contribute to refining existing theoretical models on word segmentation and acquisition. New evidence on early child word sense disambiguation expands our understanding of how chunking and analogy might function in early language development, potentially paving the way for integrated models that explain language acquisition as a synergy between these two processes. Finally, insights into the specific learning mechanisms used during development could enable more accurate mapping of developmental trajectories and individual differences. This has the potential for broad impact on society, including the enhancement of educational strategies and a better understanding of language delays.

Chapter 2

CLASSIC Utterance Boundary: A Chunking-Based Model of Early Naturalistic Word Segmentation

2.1 Abstract

Word segmentation is a crucial step in children's vocabulary learning. While computational models of word segmentation can capture infants' performance in small-scale artificial tasks, the examination of early word segmentation in naturalistic settings has been limited by the lack of measures that can relate models' performance to developmental data. Here, we extended CLASSIC (Chunking Lexical and Sublexical Sequences in Children; Jones et al., 2021), a corpus-trained chunking model that can simulate several memory and phonological and vocabulary learning phenomena to allow it to perform word segmentation using utterance boundary information, and we have named this extended version CLASSIC utterance boundary (CLASSIC-UB). Further, we compared our model to the performance of children on a wide range of new measures, capitalizing on the link between word segmentation and vocabulary learning abilities. We showed that the combination of chunking and utterance-boundary information used by CLASSIC utterance boundary allowed a better prediction of English-learning children's output vocabulary than did other models.

2.2 Introduction

Word segmentation is a fundamental process in infant language development. Phonological word forms are not given a priori but must be extracted from continuous speech input. While several computational models have captured basic word segmentation phenomena displayed by infants in small-scale artificial tasks, assessing whether models can scale up to naturalistic inputs has been hampered by limited sets of measures against which to compare performance. We present a new word segmentation model which extends CLASSIC (Chunking Lexical and Sublexical

Sequences in Children; Jones & Rowland, 2017; Jones et al., 2021; Jones, 2016; Jones, Justice, et al., 2020), a chunking model that uses naturalistic inputs to successfully simulate key developmental phenomena in memory and language. Our extended model, CLASSIC utterance boundary (CLASSIC-UB), performs unsupervised word segmentation using large-scale naturalistic inputs. Importantly, we have assessed our model against existing segmentation models using both standard evaluation metrics and novel developmental measures to provide a more comprehensive assessment of segmentation performance.

Chunking models successfully account for adult (e.g., Frank et al., 2010) and infant (e.g., French et al., 2011; Perruchet & Vinter, 1998) word segmentation in laboratory tasks by extracting and storing frequent input sequences (chunks) as candidate words that guide subsequent segmentation. This allows chunking models (e.g., Kurumada et al., 2013) to account for lexical effects in infant segmentation such as easier extraction of novel words when they are preceded by familiar words (e.g., Bortfeld et al., 2005). Lexical effects are not predicted by competing models that assume a dedicated mechanism that estimates the location of word boundaries in speech by tracking sublexical regularities, such as through forward and backward sound transitional probabilities (e.g., Cleeremans & McClelland, 1991; Saksida et al., 2016). Further, chunking also accounts for infants' sensitivity to sublexical regularities (e.g., Hay et al., 2011; Pelucchi et al., 2009; Saffran, Aslin, & Newport, 1996; Saffran et al., 1997; Saffran, Newport, & Aslin, 1996) because the component parts of a chunk are mutually linked, giving equal weight to forward and backward relations (e.g., French et al., 2011; Perruchet & Desaulty, 2008; Perruchet & Poulin-Charronnat, 2012; Perruchet & Vinter, 1998; although see McCauley & Christiansen, 2019, for a hybrid model of speech comprehension and production that forms chunks via backward transitional probability without the need to capture forward relations).

Typically, computational investigations have used artificial language tasks to assess the plausibility of learning mechanisms involved in infant (e.g., French et al., 2011; Perruchet & Vinter, 1998) and adult word segmentation (e.g., Endress & Langus, 2017; Frank et al., 2010). Although modelers have also examined scale-up

to naturalistic input (e.g., Daland & Pierrehumbert, 2011; Monaghan & Christiansen, 2010; Saksida et al., 2016), such investigations have suffered from one important limitation: The benchmark for models' segmentation accuracy has been the word boundaries present in adult vocabularies, but these word boundaries are unlikely to accurately reflect infants' and children's segmentation (e.g., Monaghan & Christiansen, 2010). In contrast, we have introduced new measures based on developmental data and specifically on the composition of children's early vocabularies. The key insight is that children's vocabularies should reflect early word segmentation processes: Word forms that are more easily discovered in the input should enter children's vocabulary earlier in development. We used these novel developmental measures alongside traditional evaluation measures to provide a much richer assessment of the developmental plausibility of word segmentation mechanisms. Specifically, we used this suite of measures to compare CLASSIC-UB to other models that have shown different strengths in modelling early naturalistic segmentation.

2.2.1 CLASSIC

CLASSIC uses a domain-general chunking mechanism (Gobet et al., 2001) to model linguistic knowledge acquisition via experience with the sequential structure of the language. It is not a model of auditory perception or production per se (as basic processes that transfer information to the learning mechanism are not modelled) but a learning model representing performance increases derived from perceptual learning and efficiency in production (Jones, Justice, et al., 2020). The accumulation of language experience is essentially represented by the chunking of adjacent items, gradually shifting the model's representations from sublexical to lexical and multiword units. A key assumption in CLASSIC is that children already know how to identify word boundaries. This has been implemented in CLASSIC because past simulations have investigated phenomena at an age where children are likely to have already learned how to segment speech into words.

We can illustrate how CLASSIC works using a simplified example in which the model repeatedly processes the phonetically transcribed utterance [d, æ, d | ɪ, z | k, ʌ, m, ɪ, ŋ][1] (i.e., *dad is coming*) where | demarcates word boundaries that, as we explained above, are given as input to the model. CLASSIC first chunks adjacent phonemes that do not cross a word boundary and forms biphone representations: [dæ, æd | ɪz | kʌ, ʌm, mɪ, ɪŋ]. Any learned chunks can subsequently be used to encode the input. For example, at the second iteration, the model would represent the utterance as [dæ, d | ɪz | kʌ, mɪ, ŋ], that is, proceeding from left to right, it uses the longest available chunks to encode each demarcated word. This way of encoding preserves the input temporal structure and represents a proxy for the increased processing efficiency derived from acquired knowledge[2]. The model then continues to join adjacent chunks; for example, the third iteration would result in the representation [dæd | ɪz | kʌmɪ, ŋ], where CLASSIC has learned two of three words in the utterance. When two adjacent chunks are words themselves, CLASSIC crosses word boundaries and learns multiword sequences (i.e., dæd|ɪz in the example); thus, at the fourth iteration, CLASSIC would encode the utterance as a two-word sequence followed by a word: [dæd|ɪz, kʌmɪŋ]. Finally, in a last iteration the model would represent the whole utterance as a single multiword chunk: [dæd|ɪz|kʌmɪŋ].

CLASSIC accounts for the role of sublexical, lexical, and multiword sequences in language development. For example, in Jones's (2016) study, incremental exposure to naturalistic speech supported CLASSIC's building up of chunks at different grain sizes, capturing 85% of variance in nonword repetition performance—a task closely related to vocabulary learning (e.g., Hoff et al., 2008)—from six studies involving 2- to 6-year-old children. CLASSIC has also simulated vocabulary learning more directly (Jones et al., 2021). Similar to the way 2–3-year-old children learn to produce words, CLASSIC gradually learns longer, more infrequent words that have a smaller number of similar words in the language (i.e., lower neighbourhood density) and higher internal predictability (i.e., higher average biphone probability or phonotactic probability). Jones et al. (2021) also showed that novel words entering children's productive vocabularies are more likely to share large phonological chunks with words that they already use, indicating a pivotal role for phonological knowledge in vocabulary learning. In sum, these studies have

shown that sublexical knowledge can be used to learn and produce pseudowords and real words (see Baayen et al., 2019; Chuang et al., 2021, for similar conclusions using linear discriminative learning).

Finally, Jones, Justice, et al. (2020) showed that phonological knowledge plays an important role in learning multiword sequences. CLASSIC captured the faster increase in children's short-term memory for digit over word sequences likely because chunks that span multiple digits are learned more quickly from random combinations of digits occurring in naturalistic speech. This study also showed how knowledge of multiword sequences facilitates lexical processing (e.g., processing of the individual items *five* and *six* becomes more efficient when the two are presented within a familiar multiword sequence *five–six*).

In sum, CLASSIC is a chunking-based model that has captured important developmental phenomena in word learning but has not yet been applied to word segmentation. We showed how CLASSIC can be extended to perform word segmentation, thus making the model more developmentally plausible: Infants must of course discover word forms before they can learn novel words and integrate them into their existing vocabulary (Newman et al., 2016).

2.2.2 CLASSIC-UB

To extend CLASSIC to perform word segmentation, we retained CLASSIC's architecture but removed word boundary information from the model input (i.e., the model was not constrained to chunk items within demarcated words). We also added utterance boundary information using positional markers (↲) that signal utterance start or end. Transcribers of the input corpora used in this study coded such positional markers based on various syntactic (e.g., utterances are centred around a main clause) and prosodic cues (e.g., pauses, intonation patterns distinguishing declarative, interrogative, or other clauses). Only written transcriptions were available for most of the input, not the original speech recordings, so it was not possible to automatically assign positional markers based on, for example, changes in phonetic features. Positional markers have been used in previous

computational work (e.g., Aslin et al., 1996; Christiansen et al., 1998; Saksida et al., 2016) as a proxy for the increased saliency that phonological units at utterance boundaries gain in child-directed speech (e.g., Fernald & Mazzie, 1991). This has been modelled via conjunctive use of utterance-boundary markers and phonological units to perform distributional learning (e.g., utterance-boundary + syllable constitutes a pair of units for which transitional probabilities can be obtained; Saksida et al., 2016). In a similar way, CLASSIC-UB treats utterance-boundary markers as additional units that can be used to form chunks (i.e., a chunk becomes longer when an utterance-boundary marker is attached to a phonological sequence).

We present a version of CLASSIC-UB that uses utterance-final markers and a version that uses both initial and final markers. Infants may privilege utterance-final words (e.g., Aslin et al., 1996; Christiansen et al., 1998) because these gain perceptual prominence from syllable lengthening (Wightman et al., 1992) and sentential accent in English (Cinque, 1993). However, some studies have suggested that infants may use both initial and final markers in segmentation (Seidl & Johnson, 2006, 2008). In fact, different cues could facilitate segmentation of utterance-initial words (e.g., exaggerated amplitude, duration, pitch, and formant structure; Cruttenden, 1986). Therefore, the presence of initial markers should provide additional facilitation over utterance-final cues. We are not aware of any computational studies assessing the relative contribution of initial and final boundaries, thus our comparing CLASSIC-UB with final markers to CLASSIC-UB with both initial and final markers could shed light on the variables that facilitate word segmentation at utterance edges.

Figure 1 illustrates how CLASSIC-UB segments input after the input has been transcribed using the *CMU Pronouncing Dictionary* (Lenzo, 2007), which contains over 134,000 phonetic transcriptions of English words and provides an automatic way to convert large orthographic input into phonetic form using alphabetic codes for phonemes rather than IPA (e.g., AE instead of æ). When encoding the utterance-final biphone AED in the first utterance, the model learns the chunk with an associated utterance-final marker (i.e., AED↵). If the chunk AED appears in later utterances, even in word-medial positions, the model will recognize that it can be

used in word-final position assuming a word boundary at this location (see the third utterance *dad is coming*). This also shows how the following phone IH is marked as "can begin a word" based on the model flagging AED as ending the preceding word DAED (bolded chunk of Figure 1). The same logic applies to utterance-initial markers. In essence, the function of the ↲ markers within chunks is akin to "this chunk can appear at the [beginning/end] of a word".

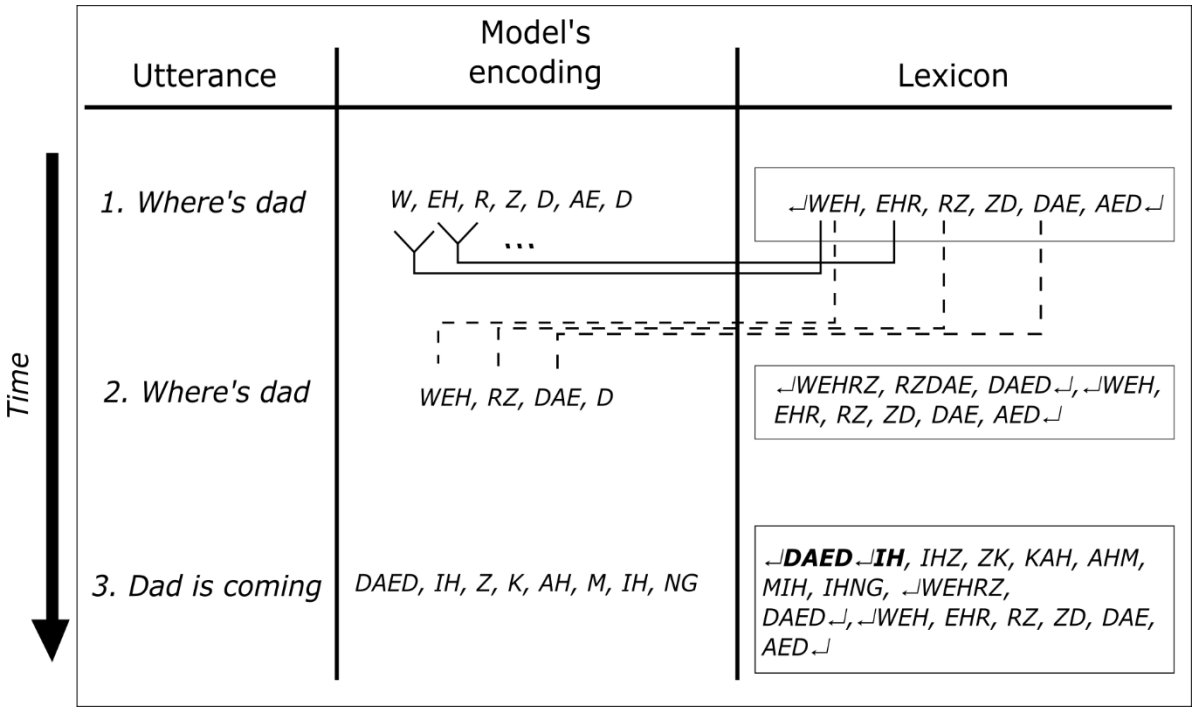


Figure 1. CLASSIC-UB generalization of utterance-boundary markers to utterance-medial position. Solid lines indicate grouping of adjacent items into single chunks and storage into the lexicon. Dashed lines indicate use of stored chunks to segment speech. Lines are only shown for the first utterance. Time indicates independent presentations of new child-directed utterances. All English phonemes are present in the lexicon but are not shown for reason of space. The transcription used is based on the CMU pronouncing dictionary (Lenzo, 2007).

Like CLASSIC, CLASSIC-UB processes phonemic input. As such, it assumes that children already know phoneme categories in line with an early phonetic category learning approach (e.g., Werker, 2018) and previous computational studies in word

segmentation (e.g., Batchelder, 2002; Daland & Pierrehumbert, 2011; Goldwater et al., 2009; but there are alternative approaches that we briefly refer to in the Discussion section). Knowledge of sound categories and co-occurrences of sounds might begin to develop at the same time or soon after infants start segmenting speech into words at around 6 months of age (Bortfeld et al., 2005). For example, between 3 and 9 months, infants discriminate between and learn new phonetic categories using distributional cues (e.g., Cristià, McGuire, et al., 2011; Maye et al., 2008; Mersad et al., 2021; Yeung et al., 2014), and they can use this information in word segmentation (e.g., Jusczyk & Aslin, 1995) and soon after in word recognition tasks (around 12 months; Mani & Plunkett, 2010) and word learning tasks (around 14 months; Fais et al., 2012). Similarly, between 4 and 9 months infants attune to native phonotactic patterns (Cristià, Seidl, & Gerken, 2011; Jusczyk et al., 1994) and can use this knowledge in word segmentation (e.g., Mattys & Jusczyk, 2001). Nevertheless, we also ran all of our simulations on syllabified input (see Method section) because infants may initially perceive syllables as basic linguistic units (e.g., Bertoncini & Mehler, 1981).

As with CLASSIC, items that co-occur often will have more opportunities to be chunked together by CLASSIC-UB. This facilitates subsequent segmentation in two ways. First, when a word is frequent in the input, its sublexical components will have more opportunities to be chunked together, reaching a whole-word representation faster. This makes the model frequency sensitive, even though frequency is not explicitly tracked (unlike in other chunking models, such as PUDDLE (Phonotactics from Utterances Determine Distributional Lexical Elements; Monaghan & Christiansen, 2010; see Appendix S1 for a detailed description of this model). Second, learning words that share phonological material with other words will be facilitated by the reuse of existing chunks (e.g., learning *just* can make the sequence *ust* available to subsequently learn *crust*). Other models, such as PUDDLE, do not include this mechanism and rely on frequency information alone.

The number and size of chunks changes as more input is processed. CLASSIC-UB processes input incrementally (i.e., one utterance at a time), as do other segmentation models (e.g., French et al., 2011; Monaghan & Christiansen,

2010; Perruchet & Vinter, 1998). As Figure 1 shows, each utterance is encoded from left to right by using existing chunks present in the model lexicon. Consistent with previous chunking models (e.g., Batchelder, 2002; French et al., 2011; Perruchet & Vinter, 1998), preference is given to encoding larger chunks over shorter ones. For example, the chunk AED ↵ that contains a boundary marker is preferred over the shorter chunk AED that does not contain a boundary marker. At the same time, new/larger chunks are stored in the model lexicon by joining adjacent encoded items together, facilitating subsequent segmentation. This makes the learning process plausible because children's learning happens incrementally as a function of their accumulating knowledge of the language (e.g., Jones et al., 2021).

Crucially, selecting larger chunks over shorter ones means that chunks formed by sublexical sequences and utterance-boundary markers are dispreferred to words, thus avoiding oversegmentation. At the same time, the presence of utterance-boundary markers prevents the model from building large undersegmented chunks. Together, these two mechanisms favour segmentation at the (intermediate) word level. However, there is no explicit rule defining when the model should stop building chunks of increasing size. In fact, at later stages, the model stores multiword chunks, which is consistent with representation of multiword sequences from 11 months of age (e.g., Jones, Cabiddu, & Avila-Varela, 2020; Skarabela et al., 2021). Notably, such longer chunks can include multiple boundary markers, which means the model can represent multiword sequences while also retaining knowledge of the individual words composing a sequence. For example, an utterance such as *I'll do it later* could be encoded using the two chunks ↵ *I'll* ↵ *do* ↵ *it* ↵ and *later* ↵. In sum, CLASSIC-UB learns chunks including both phonological and utterance-boundary information. Chunks gradually increase in size, facilitating subsequent segmentation.

2.2.3 Evaluation of Naturalistic Speech Segmentation

Corpus-based evaluations of segmentation models usually compare models' output to segmented transcriptions of child-directed speech (e.g., Monaghan & Christiansen, 2010). Precision and recall are two widely used measures. Precision is

the number of words segmented by a model divided by the total number of items segmented, including segmentation errors (i.e., how many of the items found are words). Recall is the number of words segmented by a model divided by the total number of words in the input (i.e., how many words present in the input are found). In these two measures, chunking models perform better than do models that segment speech randomly (e.g., Bernard et al., 2020; Monaghan & Christiansen, 2010), which is in line with results from computational studies capturing artificial language learning (e.g., French et al., 2011). For example, in Larsen et al.'s (2017) study, the chunking model PUDDLE showed the highest performance, reaching 82% for precision and 80% for recall. In contrast, another class of models that track sound transitional probabilities (see Appendix S1 for a detailed description) perform better than the random baseline models (e.g., Bernard et al., 2020) but less well than chunking models (e.g., 43% for precision and 51% for recall in Larsen et al.'s, 2017, study).

Although these measures capture how accurately models segment the input, they do not capture their developmental plausibility. The use of segmented input to evaluate model performance makes the implicit assumption that infants segment speech in an adult-like way. However, as discussed by Larsen et al. (2017), this assumption is likely to be wrong, given evidence that infants' initial protolexicons contain words and frequent phonotactically legal nonword sequences (e.g., Ngon et al., 2013). Addressing this problem is not straightforward because how infants segment speech in naturalistic settings is not known. Larsen et al.'s (2017) solution was to link model accuracy to word age of acquisition. For example, *dog* was understood by a higher proportion of children at 13 months of age than was *deer*, and this should be reflected by a more accurate segmentation of *dog* than *deer* (i.e., *dog* is correctly segmented on more occasions). Theoretically, the reasoning behind using word learning as a proxy for segmentation performance is that vocabulary knowledge (word–meaning mapping) is facilitated by word segmentation (e.g., Estes et al., 2007; Hay et al., 2011). For example, in Estes et al.'s (2007) study, infants were able to extract, store, and recognize word forms previously presented in fluent speech to successfully perform a label–object association task. In sum, words that are acquired early must also be accurately segmented at earlier ages[3].

We also capitalized on the link between vocabulary knowledge and segmentation as suggested by Larsen et al. (2017), but instead of age of acquisition derived from parental reports, we used age of first production derived from child speech (Grimm et al., 2017). Looking at production rather than comprehension has drawbacks, but it also has important advantages. The words children produce are, of course, not a direct reflection of their segmentation abilities. Production involves additional variables related to recalling stored instances from the lexicon and to articulation, and, of course, what children spontaneously produce at the time of recording does not reflect the entirety of their comprehension vocabularies. Further, there are limitations inherent in estimating children's knowledge from a small number of relatively short samples of speech filtered through adult transcribers' potentially biased judgement (e.g., leading to the omission of nonlexical productions). Nevertheless, using production vocabularies has two key advantages. First, it dramatically increases the number of words examined: The British communicative development inventory (CDI; Alcock, 2020), a parent-report measure of age of acquisition, contains only 330 words[4], lacking sufficient statistical sensitivity. Second, we found that the CDI word sample has a word frequency distribution shifted toward high-frequency words not reflecting the Zipfian input that infants hear, that is, many low frequency and few high-frequency word types (Hendrickson & Perfors, 2019)[5]. Using such a sample might bias results because transitional probability models might perform well only because the distribution considered is less skewed toward low frequency words (Kurumada et al., 2013).

We have additionally proposed a new measure examining whether a model can capture word-level characteristics of child vocabularies. Previous measures did not examine whether a model capitalized on sublexical/lexical regularities (similarly to how learning is evaluated in laboratory settings). Traditional measures have focused on finding a mechanism that minimizes segmentation errors, while the age of acquisition/production measure is focused on the time course of acquisition. In contrast, with our final set of analyses, we assessed whether the characteristics of the vocabulary learned by a model reflected what children had produced in the language corpora. In other words, we assessed whether the models and children were sensitive to input characteristics in a similar way. We focused on three lexical

measures—word frequency, word length, neighbourhood density—and one sublexical measure—phonotactic probability. These characteristics have explained approximately 50% of variance in word learning (Stokes, 2010, 2014; Storkel, 2009). Finally, although word comprehension as a marker of vocabulary growth has been predominant (e.g., Fernald & Marchman, 2012), the use of evaluation measures based on early production was reasonable given both the relation between early vocalizations and vocabulary growth (McGillion et al., 2017) and the relation between early segmentation abilities and later expressive vocabularies (Newman et al., 2006, 2016).

In summary, we asked whether a novel chunking account of word segmentation could scale up to naturalistic speech in a developmentally plausible way by comparing CLASSIC-UB to PUDDLE, a model that has shown high performance in traditional measures of naturalistic segmentation, and to backward and forward transitional probability models that might account for a high proportion of variance in child word knowledge (Larsen et al., 2017). We also asked whether utterance-initial edges play a role in segmentation beyond final edges by comparing two different implementations of CLASSIC-UB. Finally, we asked whether transitional probability models could capture developmental data better than chunking accounts by comparing PUDDLE to transitional probability models to test whether we had replicated previous results (Larsen et al., 2017) using different corpora and performance measures.

2.3 Method

2.3.1 Computational Models

We compared CLASSIC-UB to forward and backward transitional probability (Saksida et al., 2016), PUDDLE (Monaghan & Christiansen, 2010), and a random baseline relying on a coin toss to place a boundary after each input unit (Lignos, 2012). A full description of these models can be found in Appendix S1. We implemented the models to process syllables or phonemes as basic units (see Appendix S2 for details). Python and R scripts for preparing the input, running the models, and

analyzing the output are available at the project's OSF page (https://doi.org/10.17605/osf.io/kbnep).

2.3.2 Corpora

We used seven English corpora following Grimm et al.'s (2017) study (see Appendix S2 for input preprocessing and characteristics). We downloaded the corpora from the CHILDES database (MacWhinney, 2000). As target input for the models, we considered only transcripts of children aged 2 years. While infants start segmenting speech much earlier than 2 years of age, our choice to focus on this age group was motivated by the much smaller size of corpora of speech directed at children of younger ages (e.g., 54,274 utterances at age 1 year vs. 604,000 utterances at age 2 years). As we show in Appendix S2, this limits the representativeness of input directed at children of younger ages. In total, the input to models contained 604,000 utterances (mean length of utterance = 4.39) from 332 different speakers, directed to 53 target children. Such input was 3 to 60 times larger than input used in previous studies (Christiansen et al., 1998; Daland & Pierrehumbert, 2011; Larsen et al., 2017; Monaghan & Christiansen, 2010; Saksida et al., 2016).

2.3.3 Measures of Model Performance

2.3.3.1 *Precision and Recall*

We compared the models' performance by looking at the pairwise differences in mean precision and recall scores (e.g., Monaghan & Christiansen, 2010). We tested the last 10,000 utterances of output because the models' performance was stable (see Figure 2) and because testing the entire output (i.e., 604,000) would have led to significant results even for trivial differences. We used a Welch's *t* test for unequal variances, with *p* values and bootstrap 95% confidence intervals corrected for multiple comparisons using Holm's correction.

2.3.3.2 *Word Age of First Production*

We used the mean length of utterance for transcripts as a proxy of word age of first production following Grimm et al.'s (2017) study (see Appendix S3 for details). Mean length of utterance is a useful estimator of child gross linguistic skills (i.e., developmental stage), controlling for the fact that children with a similar age might be far apart in their language development. The sample contained 5,480 words. We fitted linear regression models predicting word age of first production as a function of the log10 number of times a target word was correctly segmented by each algorithm (Larsen et al., 2017). We weighted the number of times a word was correctly segmented by dividing it by input word frequency before fitting the regression models as the two variables correlated highly (e.g., for a random baseline, $r = .92$). Word frequency correlates highly with the age of word acquisition (e.g., Morrison et al., 1997), therefore failing to control for its effect might have led to results that were an artifact of frequency. Indeed, input frequency tended to strongly affect models' performance; for example, for the random model, the correlation between the number of correct segmentations and age of first production dropped from .58 to .20 after we controlled for frequency. Therefore, controlling for input frequency allowed us to assess the performance of each segmentation algorithm over and above the fact that words that appear more often are acquired earlier.

Since previous studies had not used weighting by word frequency, we also included analyses for the unweighted measure in Appendix S6 to facilitate comparison. To foreshadow our findings, differences between models were consistent when we used either the weighted or unweighted measure, with only one exception pertaining to transitional probability models that we address in the Discussion section. We based comparisons between models on pairwise differences in adjusted R^2 from the regression models; we bootstrapped the 95% confidence interval of the difference between coefficients and corrected the interval using Holm's correction (Grimm et al., 2017). We concluded that two coefficients differed significantly from one another if the corrected 95% confidence interval did not include 0.

2.3.3.3 *Word-Level Measures*

We compared the distributions of unique words discovered by each model to children's actual vocabulary (i.e., the words produced by children in the corpus) for phonemic length, word frequency, neighbourhood density, and phonotactic probability. According to Jones et al. (2021), the distribution of words relative to sublexical and lexical characteristics should be similar between children and model if the model's learning mechanism is developmentally plausible. As in previous studies (e.g., Storkel, 2009; Swingley & Humphrey, 2018; Vitevitch & Luce, 1998), word length referred to the number of phonemes in a word; word frequency was the log10 frequency of a word across the input; phonotactic probability was the mean probability of a phoneme pair's appearing in a word; neighbourhood density was the raw count of phonemic words that differed from a target word by one phoneme (i.e., by deletion, insertion, or substitution). We left phonotactic probability and neighbourhood density unmarked for stress to be consistent with previous work (e.g., Storkel, 2009; Swingley & Humphrey, 2018).

We carried out a chi-square goodness of fit test to compare observed probabilities of a word's being of a certain length (in the output of a segmentation model) to the expected probabilities in children's utterances; we focused on lengths of two to eight phonemes due to the low number of words at other phonemic lengths. We defined probabilities as the proportion of types at each length. We then looked at the pairwise differences in chi-square test statistics, using bootstrap confidence intervals as we described in the previous section. In other words, this analysis first looked at how close each model was to children's performance and then used the estimates of such distance to compare models to one another.

For word frequency, neighbourhood density, and phonotactic probability, which are continuous measures, we followed a similar procedure to the one that we used for word-level measures, but we used a Kolmogorov–Smirnov test statistic. Following Piantadosi et al.'s (2012) study, we divided each of these measures by word length. Word length tends to be anticorrelated with word frequency (e.g., Zipf,

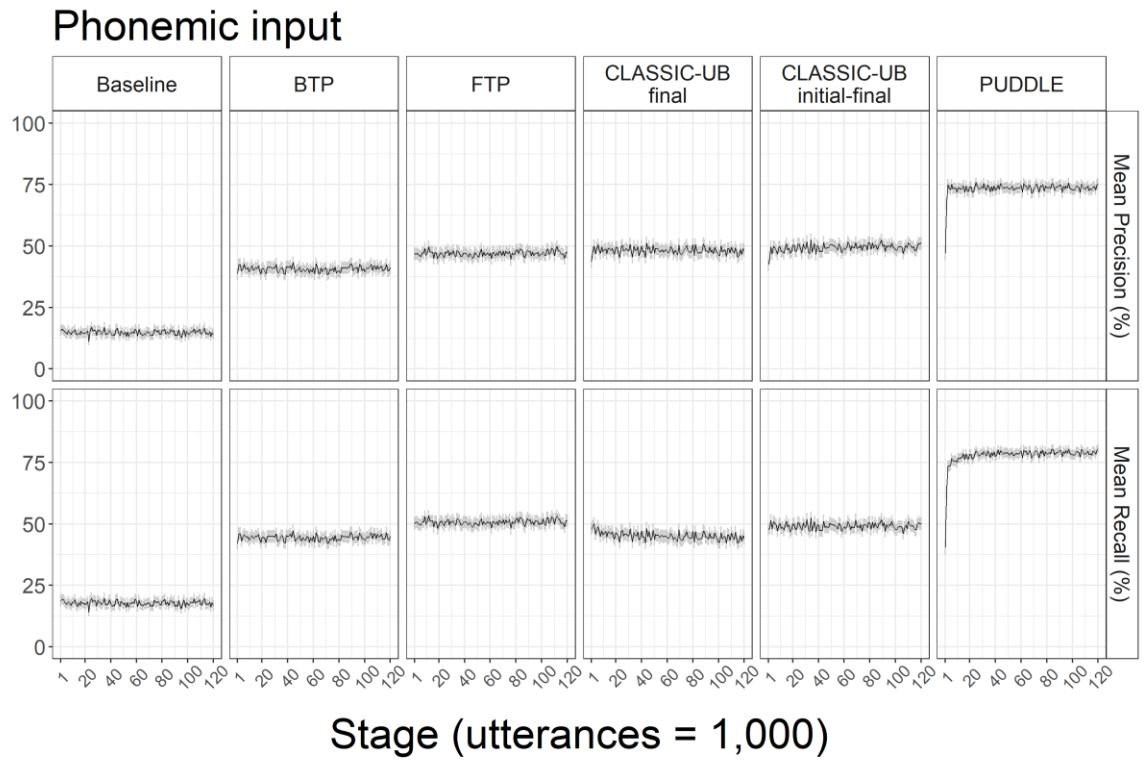
1936) and neighbourhood density (Storkel, 2004) and positively correlated with phonotactic probability (Storkel, 2004). In our dataset, the correlations varied from moderate to strong: length and frequency ($r_s = -.37$), length and neighbourhood density ($r_s = -.86$), and length and phonotactic probability ($r_s = .42$).

2.4 Results

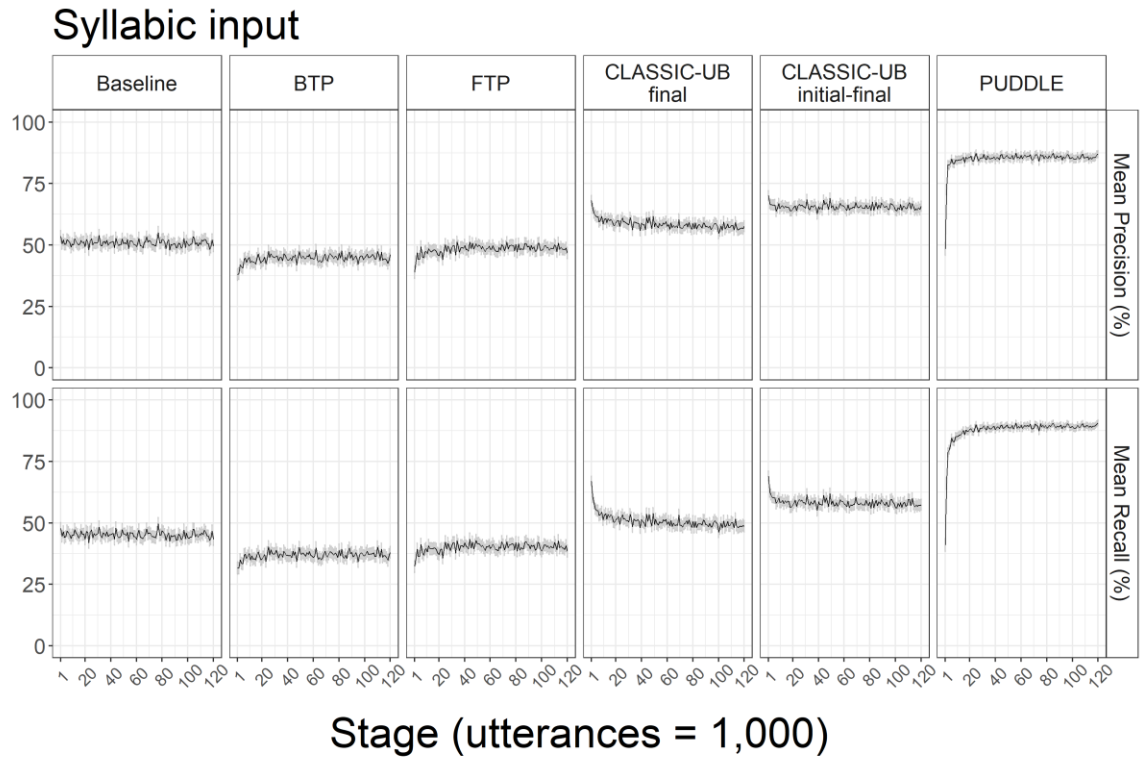
We first report results for precision/recall and age of first production and finally for word-level measures. For ease of readability, in each subsection we give only a discursive presentation of key results and point to statistical results in the appendices. We have included both CLASSIC-UB initial and CLASSIC-UB initial-final in this section; however, for reasons of space, we have provided a discursive comparison between the two models in Appendix S11.

2.4.1 Precision and Recall

All models showed rapid learning (see Figure 2), reaching a ceiling in performance after approximately 40,000 utterances and indicating that the quantity of the input did not affect their performance (consistent with Daland & Pierrehumbert, 2011). We have provided pairwise statistical comparisons for the models in Appendix S4. All models segmented the input above chance (baseline), except for the transitional probability models when the input was syllabified (see Panel B in Figure 2 and Appendix S4).



A.



B.

Figure 2. Mean precision and recall performance with phonemic (Panel A) and syllabic (Panel B) input. The figure shows the random baseline, backward transitional probability (BTP) and forward transitional probability (FTP), CLASSIC-UB

with utterance-final and initial-final markers, PUDDLE. Performance was averaged every 1,000 utterances (Stage). Only the first 120 stages are shown to better appreciate changes in performance and because the performance of the models was stable. Grey confidence bands indicate the 95% confidence interval around the mean.

In line with Larsen et al.'s (2017) findings, PUDDLE showed the best performance, outperforming the baseline, transitional probability, and CLASSIC-UB models. When we used phonemic input, PUDDLE found 73% of items were words for the precision measure and 79% of items were words for the recall measure. This model's accuracy was higher when segmenting syllabified input, reaching 85% for the precision measure and 89% for the recall measure. CLASSIC-UB's performance lay between the PUDDLE and the transitional probability models, with CLASSIC-UB initial-final reaching 50% for precision and recall with phonemic input, and 66% for precision and 58% for recall with syllabified input.

Overall, the models segmented naturalistic speech above chance. However, while traditional measures examined models' accuracy, they told us nothing regarding whether a model's segmentations reflected how infants segment speech, and we were not able to make any claim regarding the plausibility of one model compared to another. To address this issue, we turned to the next set of measures that related model performance to child data.

2.4.2 Word Age of First Production

Table 1 shows the adjusted R^2 estimates for all linear regression models. Although the sizes of the estimates were small, they were in line with the results of Larsen et al. (2017), who, for example, showed that PUDDLE explained .067 of variance in child age of acquisition[6]. After carrying out all pairwise comparisons between adjusted R^2 estimates (see Appendix S5), we found that only CLASSIC-UB initial-final, CLASSIC-UB final, and PUDDLE—and only when we ran the models on phonemic input—outperformed the baseline at predicting word age of first

production. Surprisingly, when the models were run on syllabic input, none of them passed the baseline test (see Appendix S5). We discuss this unexpected finding in Appendix S13. Also, the results that we have reported above were based on weighting the predictor measure by frequency as we explained in the Method section. We have reported the results for the unweighted measure in Appendix S6.

Table 1 Adjusted R^2 for linear regression models predicting word age of first production as a function of weighted log10 number of times a word was correctly segmented by each model

	Phonemic input		Syllabified input	
Model	$R^2_{adjusted}$	95% CI	$R^2_{adjusted}$	95% CI
Baseline	.036	[.023, .052]	.041	[.027, 057]
Backward transitional probability	.044	[.030, .059]	.000	[.000, .002]
Forward transitional probability	.046	[.030, .060]	.013	[.007, .021]
CLASSIC-UB final	.079	[.062, .100]	.021	[.012, .030]
CLASSIC-UB initial/final	.084	[.066, .103]	.038	[.025, .051]
PUDDLE	.078	[.060, .097]	.061	[.043, .078]

Note. Heteroskedasticity-robust standard errors were computed using a HC2 estimator. The 95% confidence intervals indicate lower and upper limits of bootstrap confidence intervals around the estimate based on 1,000 iterations. Holm's correction was applied by expanding the confidence intervals.

Crucially, while CLASSIC-UB had lower precision and lower recall scores compared to PUDDLE (see Figure 2), the two models explained the same proportion of variance in child word age of first production (about 8%), suggesting that achieving lower segmentation accuracy might not necessarily lead to lower developmental plausibility. Nevertheless, age of first production did not consider the characteristics of the model's vocabulary, nor did it answer questions about whether model and

children are sensitive to similar sublexical and lexical characteristics. The following fine-grained word-level measures addressed these questions.

2.4.3 Word-Level Measures

In line with the previous analysis, the models approximated children's vocabularies better than the baseline only when we ran them on phonemic input. Therefore, in the following sections we report results for the phonemic analysis. We have included the results of the syllabic analysis in Appendices S7–S10, and we also discuss this finding in Appendix S13.

2.4.3.1 *Phonemic Length*

Qualitatively, all models learned more short than long words (see Figure 3) as children do (e.g., Storkel, 2009). However, CLASSIC-UB (both initial and initial-final) approximated the proportion of long words learned by children better than either PUDDLE or the transitional probability models did. The two CLASSIC-UB models were also the only ones to outperform the baseline (see Appendix S7). Finally, PUDDLE's performance at approximating children's vocabularies by phonemic length did not differ from forward and backward transitional probability models.

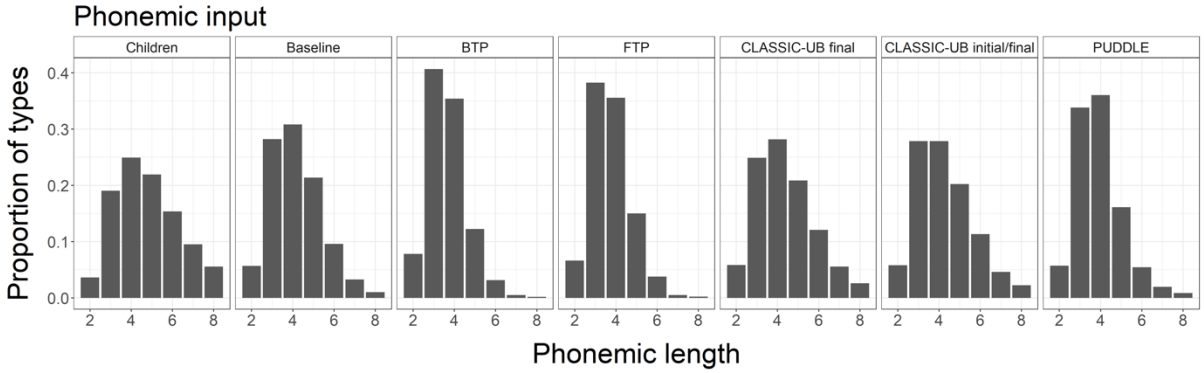


Figure 3. Proportion of word types produced by children and discovered by each model by phonemic length, when phonemic input is used.

2.4.3.2 *Word Frequency*

Children's vocabularies are Zipfian like the input that they receive (e.g., Hendrickson & Perfors, 2019), and as such their vocabularies contain more low frequency words than high frequency words. We found no significant difference between PUDDLE and CLASSIC-UB at approximating child vocabularies by word frequency (see Figure 4 and Appendix S8), but chunking models outperformed transitional probability models. This result was in line with empirical evidence showing that chunking models are better than transitional probability models at capturing lexical effects (e.g., Frank et al., 2010).

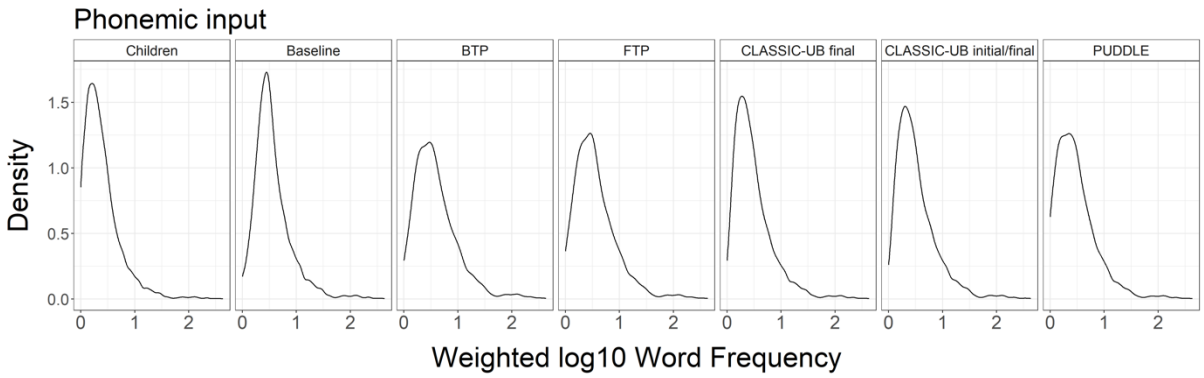


Figure 4. Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by Log10 word frequency (weighted by dividing a word frequency value by its phonemic length). Phonemic input is used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

2.4.3.3 *Neighbourhood Density*

In line with the fact that the majority of words in the language have zero or few lexical neighbours (e.g., Vitevitch, 2008), child vocabularies are populated by a high number of low-neighbourhood words. In this measure, only CLASSIC-UB final outperformed the baseline at approximating child vocabularies by neighbourhood

density, and this model performed significantly better than all other models (see Figure 5 and Appendix S9).

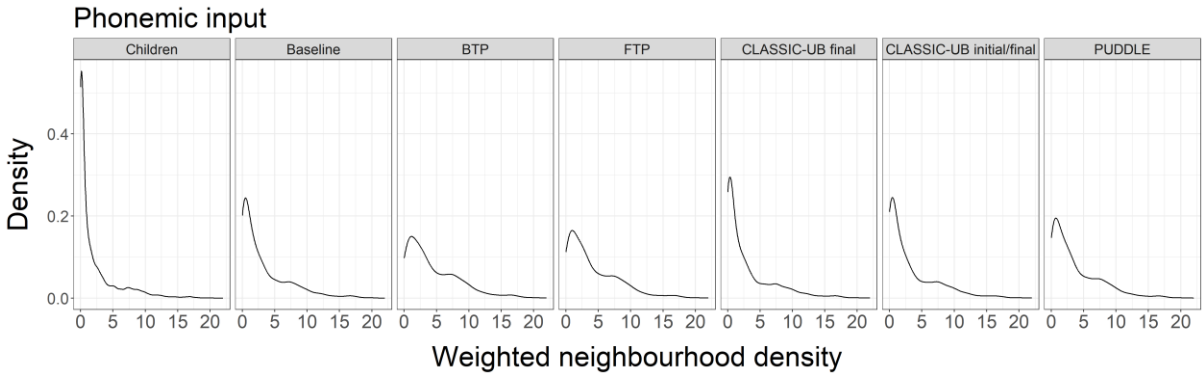


Figure 5. Distribution of unique words in child speech (Children) and discovered by each model, by neighbourhood density (weighted by dividing a word neighbourhood density value by its phonemic length). Phonemic input is used.

2.4.3.4 *Phonotactic Probability*

As Figure 6 shows, child vocabularies are populated by words with low internal predictability (e.g., Storkel, 2009). All models were equally good at approximating child vocabularies, in line with evidence showing that both chunking and transitional probability models are sensitive to sublexical regularities in the speech input. However, the models' performance did not differ statistically from the baseline model (see Appendix S10), suggesting that this measure might not have provided sufficient sensitivity for evaluating segmentation models.

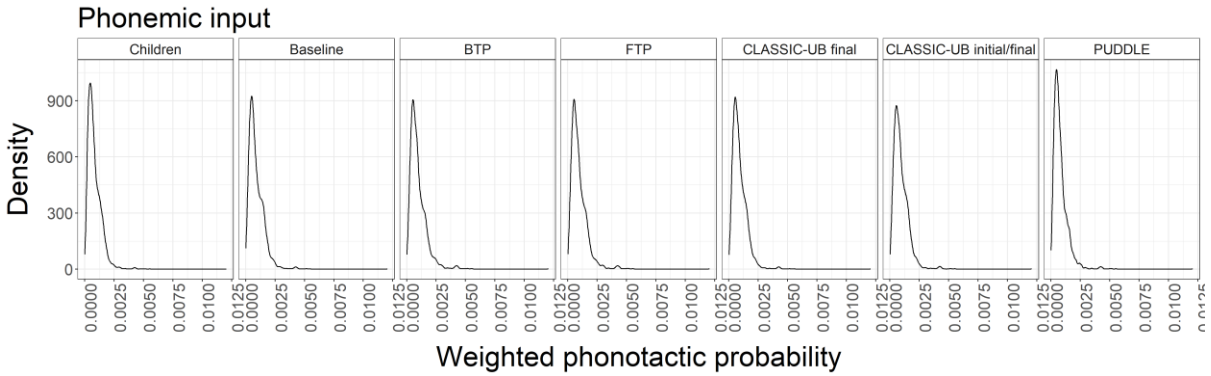


Figure 6. Distribution of unique words in child speech (Children) and discovered by each model, by phonotactic probability (weighted by dividing a word phonotactic probability value by its phonemic length). Phonemic input is used.

2.5 Discussion

We compared CLASSIC-UB, a word segmentation model that uses naturalistic input, to another chunking model (PUDDLE) as well as to nonchunking accounts of word segmentation. We broadened the assessment of model developmental plausibility by introducing new measures that related model performance to child corpus data. We found that CLASSIC-UB acquired a vocabulary that more closely captured child vocabularies than did all other models; for example, both children and CLASSIC-UB learned a higher proportion of long and low-neighbourhood words compared to other models. We discuss each of these findings in turn.

2.5.1 Measures of Developmental Plausibility

In line with Larsen et al.'s (2017) study, we found that the results of traditional evaluation measures can be inconsistent with those of measures based on child speech. In fact, overall, CLASSIC-UB performed better than PUDDLE at predicting measures based on child speech despite segmenting approximately 30% fewer word tokens. One reason for this finding might be that traditional measures represent an adult benchmark. Infants might not segment speech into the same units as adults but might, at least initially, segment and store a protolexicon made of both word and

frequent nonword units (Ngon et al., 2013). This is also consistent with different accounts (e.g., Cutler et al., 2012; Pinker, 1994b) that have predicted that learners should commit segmentation errors based on the same cues that allow them to segment speech (e.g., rhythmic structure of the language, possible-word constraint, phonotactic constraints). Although researchers still do not know which specific errors—and more importantly in which proportion—infants make when segmenting naturalistic speech over the course of development, our findings nevertheless suggest that carrying out an in-depth examination of the kind of vocabulary built by models might be a first step toward assessing models' developmental plausibility.

In Larsen et al.'s (2017) study, transitional probability models explained a higher proportion of variance in age of acquisition than did chunking models. Using our adapted production measure, we showed that this result might depend on controlling for the role of word frequency. Namely, if one controls for frequency, transitional probability models do not actually perform above chance (see transitional probability models vs. the baseline model in Appendix S5). This means that the higher performance of transitional probability models might be largely driven by input frequency. This finding is not dependent on using a production measure; in a supplementary analysis (see CDI addendum in the project's OSF profile), we examined the models' ability to predict age of acquisition based on the UK CDI (a comprehension-based measure). When the comprehension measure was not frequency-weighted, we replicated Larsen et al.'s (2017) results. But importantly, when the measure was frequency-weighted, CLASSIC-UB again performed better than the other models (consistent with the production-based analyses reported here).

We suggest that our proposed set of word-level measures might provide a richer and more nuanced method for evaluating the developmental plausibility of segmentation models. First, findings from word-level measures were in line with the age of first production results, with chunking models outperforming transitional probability and models run on syllabified input performing at chance (see Appendices S7–S10). In line with previous findings capturing in-laboratory data (e.g., French et al., 2011; Kurumada et al., 2013), word-level measures also showed

that, while both transitional probability and chunking models closely approximated child vocabularies at the sublexical level (phonotactic probability), chunking models performed better when lexical measures were considered (word length, word frequency, neighbourhood density).

Second, word-level measures provided a more detailed test of the models' lexical characteristics, highlighting performance differences that might be attributed to architectural differences across models. Indeed, CLASSIC-UB's learning mechanism facilitated the discovery of words that overlap phonologically with previously discovered words. This allowed the model to approximate a greater proportion of children's long/low-neighbourhood words than did competing models (see Figures 3 and 5). Therefore, uniquely relying on mechanisms that privilege highly probable sequences (e.g., PUDDLE, transitional probability models) makes it difficult to capture a portion of long/low-neighbourhood words that are generally more difficult to learn but that children nevertheless learn and that CLASSIC-UB can learn by exploiting phonological overlap. Interestingly, this feature of CLASSIC's learning mechanism also means that the model can account for nonword repetition effects (Jones, 2016) that are due to phonological overlap across word and nonword sequences. Similarly, it is possible that CLASSIC-UB captures additional processes of storage and recall involved in word production (i.e., going beyond aspects of segmentation) and that this sensitivity explains its superior performance in approximating the characteristics of children's productions.

Although CLASSIC-UB more accurately represented the make-up of children's early lexicons, its accuracy in segmenting words was not quite as good as that of PUDDLE (i.e., PUDDLE has a larger vocabulary). One could therefore argue that, at earlier stages in PUDDLE's learning, word-level characteristics may match those of CLASSIC-UB and that it is only the subsequent increase in PUDDLE's vocabulary that skews the distribution of the word-level characteristics. We conducted additional analyses (see Appendix S12) to evaluate this possibility. These analyses showed that differences in vocabulary size did not explain the differences in word-level measures.

Finally, to support our claim regarding the role of overlapping phonological sequences in CLASSIC-UB, we conducted an additional exploratory analysis showing

that CLASSIC-UB's ability to better approximate children's vocabulary in word length and neighbourhood density increased as word frequency increased (see Appendix S12). This is in line with recent work showing that frequent words are more likely to share phonological material with previously learned words, therefore boosting child learning compared to learning less frequent words (Jones et al., 2023). Our result was also in line with evidence showing an effect of overlapping phonological sequences on vocabulary learning at around 2 years of age (e.g., Jones et al., 2023; Stokes, 2010; Storkel, 2009) but no effect at 12–15 months (Swingley & Humphrey, 2018), suggesting that children first build a diverse repertoire of phonological chunks that later boost word learning (for a computational test of this idea using CLASSIC, see Jones & Rowland, 2017).

Overall, our results speak in favour of models that exploit phonological overlap between sequences in word segmentation (e.g., French et al., 2011; Perruchet & Vinter, 1998) and add to previous work which highlighted the significant role of the overlap between sequences in word processing and acquisition (Gathercole, 1995; Jones et al., 2021).

2.5.2 Limitations and Future Directions

We have shown that chunking might play a significant role in early word segmentation by comparing our new chunking-based segmentation model CLASSIC-UB to two other influential models: transitional probability and PUDDLE models. However, there are additional models that we did not consider. One important class of Bayesian models assumes that infants formulate hypotheses on the possible segmentations of utterances, ultimately preferring those segmentations that contain few frequent and short chunks (e.g., Goldwater et al., 2006, 2009). Another account is that infants form chunks based on both frequency and transitional probabilities (forward and backward) of syllable sequences, such as through mutual information-based clustering (Swingley, 2005). Given that these accounts are primarily driven by frequency information, future comparisons to CLASSIC-UB are important for supporting our conclusion that phonological overlap between sequences plays a role

in the segmentation process in addition to frequency. Such comparisons would also be important because one influence does not exclude the other. As we argued above, CLASSIC-UB's encoding efficiency uniquely increased when items became connected to others, that is, the more opportunities to chunk sublexical items the faster lexical representations were formed. However, once CLASSIC-UB has extracted a word representation from the input, it could further benefit from tracking its frequency in the input (e.g., see Jones, Justice, et al., 2020, for how a frequency-tracking mechanism might improve CLASSIC's performance).

Moreover, it is highly likely that early naturalistic segmentation involves the use of a combination of cues. Indeed, the results of this study indicate that chunking alone might not be enough to discover items that are very long (Figure 3), occur very infrequently (Figure 4), receive no facilitation from word neighbours (Figure 5), and are made up of improbable sequences of sounds (Figure 6). This suggests that CLASSIC-UB might need to have access to additional cues to word boundary to be able to account for children's ability to learn these words. We know that infants use a wide range of cues when segmenting speech such as prosodic salience of phrase edges (Gout at al., 2004), alternative ways to pronounce specific phonemes (i.e., allophonic variation; Hohne & Jusczyk, 1994), stress patterns (Jusczyk et al., 1999), degree of coarticulation of speech sounds (Johnson & Jusczyk, 2001), and others. Such cues could be considered in future work.

An alternative (and nonmutually exclusive) possibility is that long, infrequent items with few neighbours might be learned via generalization of linguistic structures at different levels, including the syntactic level (Lippeveld & Oshima-Takane, 2020). For example, in Abend et al.'s (2017) study, an ideal Bayesian learner performed one-shot learning (i.e., formation of new word representations from a single exposure) by leveraging the mapping of words to their syntactic categories. Examining the role of syntactic categories would be important in future work as infants' development of grammatical knowledge appears to start in parallel with the acquisition of phonology and the lexicon (e.g., Marino et al., 2020).

Aside from our focus on a single word segmentation cue, another limitation is that we did not consider the models' ability to capture the role of additional variables

in word segmentation and learning. For example, Swingley and Humphrey (2018) showed that word concreteness, word frequency in isolation (i.e., frequency with which a word occurs in a single-word utterance), and syntactic category predict word learning at 12 and 15 months of age. These predictors could be included in the statistical models of age of acquisition/production alongside our word-level predictors to see how they moderate models' accuracy (i.e., number of correct word segmentations). Alternatively, our word-level evaluation measure could be extended to examine whether segmentation models can capture the distributions of these additional word-level features in children's vocabularies. We would expect models to better capture characteristics to which they are sensitive, for example, in the sense that chunking models would show sensitivity to word frequency in isolation (Kurumada et al., 2013).

Moreover, including these additional variables would be important because they differently impacted word comprehension and production in Swingley and Humphrey's (2018) study; word concreteness only predicted word comprehension, and the effect of word frequency in isolation was moderated by syntactic category type only in word comprehension. Although we have highlighted limitations in using comprehension measures to investigate how well segmentation models perform, methods that look at comprehension and production should be considered complementary. Comparing comprehension and production would also allow researchers to test the extent to which CLASSIC-UB captures processes that are uniquely involved in production (such as recall and articulation).

We would also like to highlight limitations deriving from the use of phoneme-based input adopted in our study. The models did not have to deal with the complex problem of gradually abstracting phonological categories. Under an early phonetic learning approach (e.g., Werker, 2018), infants have to learn the relations between different realizations of phonemes based on contextual variation or lexical contrast (e.g., aspirated stops and unreleased stops are allophones of the phoneme /t/). Addressing this limitation in future work is important for increasing the developmental plausibility of the investigations. Alternatively, under more recent approaches, the goal of infant speech perception may not be to learn discrete

phonetic categories but instead be to represent continuous dimensions of raw speech (e.g., spectral energy) that are relevant to the native language (i.e., perceptual space learning; Feldman et al., 2021; McMurray, 2022). This implies that future work would need to consider more gradient units of speech perception. For example, recent work by Schatz et al. (2021) showed that a distributional learner can learn to discriminate phonetic contrasts by clustering auditory features into categories that are significantly smaller and more variable than traditional phonetic categories. Finally, we acknowledge that the early phonetic learning approach used in our work was also in contrast to other accounts that do not assume phonemes as basic units of perception, for example, work that has argued for gradient units dependent on the temporal unfolding of speech (e.g., Browman & Goldstein, 1992; Bybee, 2001; Mowrey & Pagliuca, 1995; Port & Leary, 2005) or others that have argued for features or morphophonemic forms (e.g., Chomsky & Halle, 1965; Postal, 1968).

2.6 Conclusion

Our goal in this study was to test whether a chunking-based mechanism that has previously been successful in capturing early vocabulary learning might play a significant role in infant word segmentation. We then constructed CLASSIC-UB, which forms chunks of phonological and utterance-boundary material. Our simulations make three important contributions: They offer proof that (a) utterance boundaries carry useful information for word segmentation, (b) age of production and word-level measures can sensibly be used to evaluate model performance, and (c) CLASSIC can be augmented to form the segmentation model CLASSIC-UB, consistent with the hypothesis that chunking might be an important mechanism in early naturalistic word segmentation.

Chapter 3
Simulating Early Word Segmentation and Word Learning from Italian Child-Directed Speech

3.1 Abstract

Syllables are likely the initial linguistic units infants discover from the speech stream during the first months of life. Infants might use statistical regularities between syllables to discover words from child-directed speech and build their early vocabularies. However, Cabiddu et al. (2023), Chapter 2 of the present thesis, have shown that only word segmentation models applied to phonemic input explained variability in different properties of English children's vocabularies. None of the models run on syllabified input performed above the chance level, represented by a baseline model that segmented speech at random. Although these findings suggest that subsyllabic units might play a role in word learning as soon as infants begin discovering words from naturalistic speech, they might have been produced by a lack of sensitivity of the evaluation measures used. English child-directed speech includes a high proportion of monosyllabic words, which greatly simplified the segmentation task for a random model that had to correctly guess only a low number of consecutive word boundaries per utterance. To examine this potential artifact of syllabic word length, we replicated previous simulations of English on a new sample of Italian child-directed speech. We found that assuming phonemes as basic units of speech perception still provided a better account of early vocabulary learning in a language mostly containing multisyllabic words. We also showed that, to achieve sufficient sensitivity, measures that relate model performance to child data need to focus on the word token level, rather than examining word type distributions as in previous studies. Our cross-linguistic extension of previous analyses also indicated that the significant role of chunking learning mechanisms for capturing English vocabularies generalized to Italian. Moreover, to better examine how differences between English and Italian could influence models' performance, we added two new evaluation measures that examined the models' vocabularies by part of speech and number of morphological units acquired. We found that chunking

models showed emergence of morphological representations, and a noun advantage which mirrored that observed in Italian children's production vocabularies.

3.2 Introduction

Examining the interaction of different levels of linguistic representation on language development is important because children do not solve linguistic tasks in isolation in the naturalistic environment. For example, when children start learning about the phonology of their language, they are also learning which phonological sequences in the speech input correspond to word-like units (e.g., Martin et al., 2013). In this computational study, we started examining how these two variables (type of phonological representations and segmentation learning mechanisms) might interact to build children's early vocabularies.

Evidence from speech perception studies suggests that the initial linguistic unit infants perceive might be the syllable (Bertoncini et al., 1988; Bertoncini & Mehler, 1981; Bijeljac-Babic et al., 1993; Jusczyk et al., 1995; Jusczyk & Derrah, 1987). The prosodic characteristics of syllables (e.g., sonority) facilitate segmentation of continuous speech into representations structured around vowels, and thus allow the discovery of word forms by identifying at least some word onsets and offsets in the input (e.g., Räsänen et al., 2018). This behavioural evidence has typically been used by researchers to argue that syllables should be used as the basic units of speech perception in computational models of naturalistic word segmentation (e.g., Gambell & Yang, 2006; Saksida et al., 2016; Swingley, 2005). This was a plausible assumption, as such studies aimed to answer a fundamental question about word segmentation: Given unsegmented speech input, can infants use a given learning mechanism to identify a significant portion of input word forms?

However, it is unclear whether assuming that children (and models) start from syllabified input remains plausible when the aim of the investigation moves beyond this fundamental question, for example examining how segmentation performance influences subsequent vocabulary learning (e.g., Newman et al., 2006; Newman et al., 2016). In fact, from 9 months of age infants use phonemic cues in

word segmentation (e.g., Jusczyk & Aslin, 1995; Mattys & Jusczyk, 2001), word processing (e.g., Mani & Plunkett, 2010), and word learning (e.g., Fais et al., 2012). These studies suggest that integrating phonemic knowledge might be important when modelling the shift from word segmentation to vocabulary acquisition. Alternatively, phonemic knowledge might not represent a significant contributor in the early phases of word acquisition: Using models that do not segment within syllables might still be reasonable, because other studies have highlighted how acquiring phonemic knowledge is a slow process that continues throughout childhood (3 to 12 years of age, e.g., McMurray, 2022) and that might be dependent on formal education (e.g., Morais et al., 1986, 1989).

Importantly, this is currently an open question, as different computational studies have found mixed results. One study suggests that syllabic segmentation leads to better prediction of English word age of acquisition norms compared to phonemic segmentation (Larsen et al., 2017), while another study indicated that only phonemic segmentation allows segmentation models to capture when English children are likely to first produce a word (i.e., word age of first production) and different word-level properties (e.g., word length distribution; Cabiddu et al., 2023) in child vocabularies.

In this work, we built on these studies and addressed key limitations which limited their conclusions about the role of different basic units of segmentation. Specifically, we compared a number of segmentation models to random baselines (not used in Larsen et al., 2017) that allowed us to test for the unique contribution of different assumptions about the basic units of speech segmentation on vocabulary learning (i.e., controlling for the influence of the specific word segmentation mechanisms implemented by each model). Crucially, we also tested our models on Italian child-directed speech to extend the conclusions of previous studies cross-linguistically, with the main goal of addressing a potential artifact effect of word length when modelling English vocabulary learning: One likely reason why syllable representations achieve high performance in capturing child vocabulary learning even when the segmentation mechanism is random is due to the presence of many monosyllabic words in English input (Cabiddu et al., 2023), leaving less room for

more developmentally plausible segmentation mechanisms to make a difference in terms of predictive power.

In contrast, Italian child-directed speech contains a significantly lower proportion of monosyllabic words, which should increase the sensitivity of the evaluation measures used. Consider a random baseline model that places a boundary after every input unit based on a coin toss. With the unsegmented (but syllabified) utterance "I run yes ter day", the baseline only needs to correctly guess the presence of a word boundary before and after the utterance-medial word "run" (i.e., two consecutive correct guesses, with probability 0.5^2 = .25), while the task becomes more difficult when the input is in phonemic form (I r u n y e s t e r d a y) where four consecutive choices need to be made (i.e., place a boundary before "r" and after "n", and do not place a boundary between "r" and "u", and between "u" and "n", with success probability of 0.5^4 = .063). The syllabic segmentation task becomes even easier when a monosyllabic word appears at an utterance boundary, in which case the left or right edge of the word is given for free. Finally, the facilitation is maximal in syllabified one-word utterances, where monosyllabic words are discovered by a random model with a probability of 1.

Some evidence exists that using a language with higher average word length might increase the sensitivity of model evaluation measures. For example, Gervain and Guevara Erra (2012) have examined the performance of models that locate word boundaries in unsegmented speech based on transitional probabilities of adjacent syllable pairs. The study showed that segmentation models run on Italian syllabic input segmented a higher proportion of input word tokens than a random baseline, while the same models in Cabiddu et al. (2023) did not perform above chance when segmenting English input. It remains unclear, however, whether using Italian input would translate into higher sensitivity of measures that relate models' performance to aspects of child vocabulary learning.

Aside from the main goal of examining the role of phonemes versus syllable as basic units of speech perception, our work is the first to relate models' segmentation performance to child vocabulary production data in a language other than English. Cross-linguistic examinations of segmentation mechanisms have

focused on understanding which mechanisms can maintain a high segmentation accuracy across languages and which input characteristics moderate models' performance (e.g., Caines et al., 2019; Fourtassi et al., 2013; Gervain & Guevara Erra, 2012; Phillips & Pearl, 2014; Saksida et al., 2016). Here, we examined whether computational models implementing statistical learning mechanisms could capture key characteristics of Italian children's early production vocabularies.

Following Cabiddu et al. (2023), we examined two families of segmentation models: Transitional probability and chunking models. The use of transitional probability models was based on evidence of infants' reliance on forward and backward sound transitional probabilities to locate word boundaries in artificial and natural languages (e.g., Hay et al., 2011; Pelucchi et al., 2009; Saffran, Aslin, & Newport, 1996). Chunking models instead implemented the idea that familiarity with n-gram sequences (from sublexical to multiword units) might facilitate subsequent word segmentation (e.g., Bortfeld et al., 2005; Cabiddu et al., 2023; French et al., 2011; Monaghan & Christiansen, 2010; Perruchet & Vinter, 1998). In their study, Cabiddu et al. (2023) found that chunking models outperformed transitional probability models in different evaluation measures: Chunking models segmented the speech input with higher accuracy, discovering the largest number of input word tokens. Further, chunking models' segmentation accuracy explained the largest variability in children's word age of first production. Finally, the distribution of word types acquired by chunking models more closely resembled 2-year-old children's production vocabularies by different word-level properties - frequency, phonemic length, neighbourhood density (how many words sound similar to a target word in child-directed speech), and phonotactic probability (how predictable a target word is, based on the average probability of its biphone sequences in child-directed speech). Together, these properties account for 20-50% of the variance in English child word learning (e.g., Jones et al., 2023; Stokes, 2010, 2014; Storkel, 2009). In sum, Cabiddu et al.'s (2023) study highlighted the significant role of chunking in segmenting word-like units from naturalistic unsegmented speech to build children's early vocabularies.

Here, we extended this study to examine whether chunking might also play a significant role in Italian. There is evidence that, across languages including Italian, chunking models can segment naturalistic input with higher accuracy than transitional probability models (Caines et al., 2019). However, it is unclear whether higher accuracy would translate into higher developmental plausibility, as different studies have shown how model segmentation accuracy does not always lead to better prediction of aspects of children's vocabularies (Cabiddu et al., 2023; Larsen et al., 2017).

In the following section, we introduce key features of the Italian language and how they relate to English. We also present a set of predictions for this study, based on the role that these key features played in previous behavioural and computational studies.

3.3 Similarities and Differences between Italian and English Speech

3.3.1 Word Length

Words in Italian child-directed speech have a higher average length compared to English, which should decrease the performance of baseline models that segment speech at random. Consequently, when syllabic input is used, the segmentation mechanisms tested should surpass baseline models at predicting child word age of first production and different child word-level properties as tested in Cabiddu et al. (2023). This should ultimately allow us to examine the plausibility of each segmentation mechanism in a more sensitive manner.

Most words in Italian child-directed speech are multisyllabic (*Italian mean syllabic length* = 1.83; *English mean syllabic length* = 1.16; Saksida et al., 2016). Given that there are 2^{N-1} ways to segment a string of *N* phonemes or syllables, longer words produce higher ambiguity in segmentation as confirmed by different computational studies (e.g., Caines et al., 2019; Fourtassi et al., 2013; Saksida et al., 2016). In this study, we leveraged the higher ambiguity of Italian child-directed speech to increase the sensitivity of developmental measures used in previous studies (Cabiddu et al., 2023; Larsen et al., 2017). Larsen et al. (2017) used a

developmental measure that related models' segmentation accuracy to child word age of acquisition scores: They fitted linear regression models predicting the proportion of English-speaking children that at 13 months are reported to comprehend a target word - in the Communicative Development Inventory (CDI, Fenson et al., 2007) - from the number of times a model correctly segmented the target word from the input. The study did not find a clear advantage for a specific input unit. A segmentation model that tracks sound pair transitional probabilities in English (Saksida et al., 2016) explained the largest proportion of variance in word age of acquisition when processing syllabic input. Instead, a chunking model that tracks the frequency of n-grams to determine plausible English word-like units (PUDDLE, Monaghan & Christiansen, 2010) performed better on phonemic input. Importantly, the study did not include baseline models that segment speech randomly, not allowing to test how much variance was explained by assuming access to phonemic or syllabic units while controlling for the influence of specific segmentation mechanisms.

The computational study of Cabiddu et al. (2023) included phonemic and syllabic random baseline models, and further extended the evaluation measures used by Larsen et al. (2017). They not only looked at word age of acquisition, but also word age of first production (estimated from corpora of child speech), and at how models' vocabularies related to 2-year-olds' production vocabularies in terms of the following word-level characteristics: frequency, phonemic length, neighbourhood density, and phonotactic probability. Different studies have shown that these four word-level characteristics account for 20-50% variance in English child word learning (e.g., Jones et al., 2023; Stokes, 2010, 2014; Storkel, 2009). Across all developmental measures, only phonemic segmentation led to above-chance prediction of child vocabulary data, with this pattern being consistent across all transitional probability and chunking models considered.

Although this result might support the role of phonemic segmentation in early word acquisition, it requires further investigation. In fact, Cabiddu et al.'s (2023) study showed that models trained on syllabic input failed to account for child data because even the random baseline model built a larger vocabulary than children,

despite having received a comparatively small amount of child-directed input from the CHILDES database (MacWhinney, 2000): The study used a sample of 604,000 utterances, which (according to one estimate; Swingley, 2007) would correspond to the amount of input that 1-year-old children approximately receive over just a 3-week period. The baseline model also surpassed children in certain vocabulary measures (e.g., learning a higher proportion of low frequency words than are present in children production vocabularies).

As suggested above, one reason for the high performance of the random baseline model might be the high proportion of monosyllabic word tokens in English (81% in the child-directed input used in Cabiddu et al.'s study). One way of testing this hypothesis is to repeat the simulations of Cabiddu et al. (2023) using Italian child-directed speech, which contains a lower proportion of monosyllabic word tokens (43% in the corpora of our study). We expected our Italian results to differ in two ways from Cabiddu et al.'s (2023) English results. First, as discovering word forms is more difficult when most input words are multisyllabic, we expected the syllabic random baseline model to acquire a vocabulary that is smaller than children's productive vocabularies. In addition, if previous English results were an artifact of word length, the higher segmentation ambiguity of Italian speech should reduce the ability of the random baseline to predict aspects of children's vocabulary acquisition, and consequently allow non-random models to perform above-chance in this prediction task, so that we can compare them against each other.

3.3.2 Utterance Boundary Cues

In English child-directed speech, words appearing at utterance edges gain salience from exaggerated prosodic characteristics (e.g., Cinque, 1993; Cruttenden, 1986; Wightman et al., 1992) and pauses between utterances (Fernald et al., 1989). As a result, infant word segmentation is facilitated by the salience of both utterance-initial and utterance-final words (Mattys et al., 1999; Seidl & Johnson, 2006, 2008). Further, novel nouns tend to be placed in utterance-final position (e.g., Fernald & Mazzie, 1991), which may facilitate word processing (Soderstrom, 2007) and word

learning (Golinkoff & Alioto, 1995) for these nouns. Findings from studies on Italian suggest that the role of utterance-boundary cues should be similar in this language as for English. Italian infants' word segmentation might benefit from utterance-initial and utterance-final boundary cues given their sensitivity to differences in word frequency distributions at utterance edges (Gervain et al., 2008). Further, in Italian child-directed speech, presenting novel words in utterance-final position facilitates children's word acquisition (Longobardi et al., 2015).

In their recent computational study, Cabiddu et al. (2023) found results in line with the role of utterance boundary cues in English word segmentation and word learning. They used the segmentation model CLASSIC Utterance Boundary (CLASSIC-UB), which is sensitive to utterance boundary information by recursively joining adjacent sequences composed by phonological material and utterance boundary markers. The model stores these n-gram sequences and uses them to segment speech into word-like units. CLASSIC-UB was tested in two versions, one that implemented sensitivity to utterance final cues only (CLASSIC-UB final) and one with both utterance-initial and utterance-final cues (CLASSIC-UB initial/final). The study found that both utterance edges were useful to segment speech: CLASSIC-UB final discovered a larger proportion of input word tokens than random baselines, and CLASSIC-UB initial/final performed better than CLASSIC-UB final. However, when the model vocabulary was examined (i.e., the word types learned), adding sensitivity to utterance-initial cues (on top of utterance-final cues) did not improve CLASSIC-UB's ability to capture word-level distributions of English children's vocabularies by word frequency, word length, neighbourhood density, and phonotactic probability.

These contradicting results were explained by differences in word token/type ratio at utterance boundaries in the child-directed speech input. On the one hand, a lower number of different (but highly frequent) words appeared in utterance-initial position compared to utterance-final position, increasing the likelihood of segmenting these words correctly and improving the overall model segmentation (token-based) scores; it is likely that these words included function words which are known to facilitate word segmentation in models (Johnson et al., 2014) and infants (Shi et al., 2006; Shi & Lepage, 2008). On the other hand, the input corpora

contained a larger number of different words in utterance-final position than in utterance-initial position, in line with the tendency of caregivers to present novel words in utterance-final position (Fernald & Mazzie, 1991). The higher diversity of words in utterance-final position meant that a model sensitive only to utterance-final cues could already build a large (type) vocabulary, with additional sensitivity to utterance-initial cues not significantly improving CLASSIC-UB's ability to capture the age at which children first start producing a word in the transcripts and word-level characteristics of English children's vocabularies.

In this study, we used the two versions CLASSIC-UB final and CLASSIC-UB initial/final introduced by Cabiddu et al. (2023). Different studies on Italian (Gervain et al., 2008; Gervain & Guevara Erra, 2012; Longobardi et al., 2015, 2016) suggest that we should find similar results when the models are run on Italian child-directed speech.

As in English, Italian function words are repeated frequently in utterance-initial position (i.e., high token frequency), while many different content words tend to appear in utterance-final position (i.e., high type frequency). Differently from English, in Italian the subject can be omitted, therefore verbs can also appear in utterance-initial position and might attenuate the facilitatory effect of function words. Nevertheless, Italian infants may still be sensitive to highly frequent function words in the input. For example, the computational study of Gervain and Guevara Erra (2012) found that function words are traceable in Italian word segmentation, with forward transitional probability models performing better than backward transitional probability models because tracking forward relations leads to discovery of a higher proportion of utterance-initial functors (while the opposite was found for the functor-final Hungarian language). Further, Gervain et al. (2008) showed that Italian infants preferred an artificial language with a word order that respected the function/content word distribution of their native language (i.e., utterance beginning = high token frequency, utterance end = high type frequency). This study suggests that infants were sensitive to the positional saliency of utterance-initial and utterance-final words. Therefore, in this study we expected computational models to

benefit from both utterance-initial and utterance-final cues in Italian word segmentation.

Further, even if Italian child-directed speech overall contains more verbs than nouns, the early production vocabularies of Italian-learning children still contain more nouns (Longobardi et al., 2015) – just like the vocabularies of English-learning children (Bates et al., 1994). Although there could be different explanations for this noun advantage (e.g., noun concreteness facilitating noun-meaning mapping, lower morphological complexity of Italian nouns compared to verbs), the interaction between prosodic cues and word frequency might also play a role. Nouns more often appear in utterance-final position than verbs and they more often appear in utterance-final position than in other positions. Longobardi et al. (2015) found that the percentage of maternal utterances containing nouns in final position at child age 16 months correlated with the overall percentage of noun types produced by children at 20 months. In contrast, verbs in Italian child-directed speech more often appear in medial and utterance-initial positions than nouns, with overall prevalence in medial position which should make them more difficult to segment. Importantly, the frequency and positional salience of Italian nouns and verbs is useful to test the role of utterance-final cues in determining a noun advantage. In fact, in English child-directed speech noun types not only appear more often than verbs in utterance-final position, but they are also the most frequent part-of-speech category (e.g., Jones et al., 2023). This makes it difficult to examine whether the noun advantage in children's productions is influenced by the positional salience of nouns or simply by their overall higher frequency compared to verbs. Instead, in Italian, verbs are overall more frequent than nouns. Therefore, if a noun advantage emerged in a computational model that only has access to linguistic input, one could conclude that the positional salience of utterance-final nouns would likely be driving the effect.

Based on Cabiddu et al. (2023), if tracking novel words at the end of utterances explains a large variability in child world-level properties, we would only expect a significant facilitation from utterance-final cues (i.e., with variability explained by utterance-initial cues being negligible).

Further, if prosodic salience of utterance-final words plays a key role in determining a noun advantage in Italian vocabulary learning, we expected to find more nouns than verbs in the models' acquired vocabularies, despite presence of more verbs in the Italian child-directed speech input.

3.3.3 Morphology

Italian child-directed speech has a richer morphology than English, which should overall decrease models' segmentation performance due to higher oversegmentation (Johnson, 2008). However, oversegmentation of morphologically complex languages has also been shown to lead to discovery of morphological units (Loukatou et al., 2022). Therefore, it is possible that a developmentally plausible segmentation model (which captures aspects of children's vocabularies) might show learning of morphological forms alongside word units.

A key difference between English and Italian concerns their morphological characteristics. English has a simpler morphology with most words being monomorphemic, which means that morphological and word boundaries often match. Instead, Italian has a richer morphological system mostly characterised by inflectional paradigms (e.g., *casa* = house, *case* = houses) for nouns, verbs, adjectives, articles, and pronouns. Several studies have shown that segmentation models segment input corpora with lower accuracy when morphologically rich input is provided (e.g., Fourtassi et al., 2013; Johnson, 2008; Loukatou et al., 2018; 2019; 2022). For example, Loukatou et al. (2022) have shown that segmentation accuracy decreases because models present higher rates of oversegmentation when morphological complexity increases. When segmenting morphologically complex languages, oversegmentation could be useful to discover meaningful morphemes alongside word forms, in line with children's early sensitivity to morphological units (Ferry et al., 2020; Marquis & Shi, 2015). Indeed, when plausible oversegmentation errors (e.g., oversegmenting real morphemes) were considered correct segmentations, the improvement in accuracy scores became more pronounced as morphological complexity increased, measured as the degree of synthesis (i.e.,

number of morphosyntactic features that a word in a language can encode; Loukatou et al., 2022).

In sum, previous studies have identified a connection between word oversegmentation and morphological segmentation. However, it is not clear whether models that are less accurate due to oversegmentation would still capture aspects of child vocabularies successfully. It is possible that oversegmentation might lead to discovering morphological units at the expense of word forms, ultimately decreasing the models' fit to child word-level measures. Alternatively, it is possible that lower segmentation accuracy due to oversegmentation (and discovery of morphemes) might provide a better fit to child data, in line with evidence that Italian infants understand the meaning of morphological regularities from 12 months of age (e.g., Ferry et al., 2020). In this study, we examined whether models' oversegmentation led to discovery of morphemes, and whether models that better captured aspects of child vocabularies also showed learning of morphological units.

3.3.4 Summary of Research Questions and Hypotheses

We recapitulate the research questions and hypotheses of the study to facilitate understanding of the subsequent sections.

The primary focus of the study was to determine whether the short syllabic length of English child-directed speech decreased the sensitivity of evaluation measures testing the ability of segmentation models to capture early child production vocabularies. We hypothesized that using Italian child-directed speech would increase the sensitivity of the evaluation measures, thereby allowing for a comparison of the performance of various segmentation models.

For the first time, this study examines whether the advantage that chunking models have over transitional probability models, as observed in English segmentation and vocabulary learning, also applies to Italian. Given evidence of superior performance of chunking models in Italian segmentation (e.g., Caines et al., 2019), we expected that these models would also better capture aspects of Italian child vocabulary learning compared to transitional probability models.

The study also assessed the potential impact of the saliency of utterance boundaries on Italian segmentation and vocabulary learning. Given the similarities in word type and token frequency distributions at utterance boundaries between English and Italian, we expected analogous effects of utterance boundaries in Italian segmentation and vocabulary learning to those observed in English (Cabiddu et al., 2023). Specifically, cues at the beginning and end of utterances were expected to positively influence word segmentation. We also hypothesized that utterance-initial boundary cues would not explain variability in child word learning beyond that explained by utterance-final cues. Moreover, we expected that sensitivity to utterance-final cues would produce a noun bias as observed in Italian child vocabularies (Longobardi et al., 2015).

Lastly, this study aimed to examine whether oversegmenting Italian child-directed speech (due to its being a language characterized by greater morphological complexity than English) would result in the discovery of morphological units. More specifically, a greater number of discovered morphological units was expected in models that more accurately captured child vocabulary data, consistent with evidence pointing to Italian children's early knowledge of morphology (e.g., Ferry et al., 2020), occurring concurrently with their early vocabulary development.

3.4 Method

In the following sections, we present details about the segmentation models used, the preparation of input corpora, and the evaluation measures used for the analyses. The code for preparing the input corpora, running the segmentation models, and reproducing the results of the study is freely available at https://osf.io/xwp6u/?view_only=456aba900d4a47ea9ec7f0416cff2d6b.

3.4.1 Computational models

We considered the same segmentation models used by Cabiddu et al. (2023), run on either phonemic or syllabic input. A full theoretical and computational description of

each model is provided in the original article, here we provide a brief overview of the models.

Two types of models were used, transitional probability and chunking models. We used two models that identify word boundaries based on forward or backward transitional probabilities of sound pairs (e.g., Saksida et al., 2016): A word boundary between a pair of phoneme or syllable units is placed when the probability that a unit follows (forward transitional probability) or precedes another (backward transitional probability) is low compared to the surrounding pairs: For example, in the unsegmented phonetic sequence "↲wɛrzdæd↲" ("↲wheresdad↲"), a word boundary is placed between the phoneme pair "zd" when its transitional probability is lower than the transitional probabilities of "rz" and "dæ". Also, in the example, the symbol ↲ signals an utterance boundary. Typically, transitional probability models include information about utterance boundaries (e.g., Gervain & Guevara Erra, 2012; Saksida et al., 2016) that is used to compute transitional probabilities (e.g., in "↲wɛrzdæd↲", "↲w" is treated as the first pair of the sequence).

Chunking models learn a lexicon of word-like phonological sequences (chunks) that are used to facilitate subsequent segmentation. Two chunking models were used, CLASSIC-UB (Cabiddu et al., 2023), and PUDDLE (Phonotactics from Utterances Determine Distributional Lexical Elements; Monaghan & Christiansen, 2010).

CLASSIC-UB (Cabiddu et al., 2023) is a model that uses an associative learning mechanism of chunking (Gobet et al., 2001) operating on sequences comprised of phonological and utterance boundary material. It is based on the model CLASSIC (e.g., Jones et al., 2021) and represents how gaining familiarity with sound combinations at different grain sizes facilitates language processing and learning (e.g., Christiansen & Chater, 2016; Jones, 2012; 2016; Jones et al., 2020; 2021). Before receiving any input, the model is equipped with knowledge of phonemes or syllable units. When receiving its first unsegmented utterance (e.g., "wɛrzdæd"), the model encodes the input using such basic units ("w | ɛ | r | z | d | æ | d") while also learning new chunks by joining adjacent units ("↲wɛ", "ɛr", "rz", "zd", "dæ", "æd↲"). Learning new chunks allows the model to encode future input

more efficiently (i.e., using the longest available chunks to encode new utterances). For example, a second independent presentation of the same input utterance "wɛrzdæd" would now be encoded using fewer chunks ("↲wɛ | rz | dæ | d"), and further result in learning of the new chunks "↲wɛrz", "rzdæ", and "dæd↲". As shown in this example and as found for its parent architecture CLASSIC (Jones et al., 2021), Cabiddu et al. (2023) showed that a key advantage of CLASSIC-UB is its ability to reuse phonological chunks to learn new words. For example, this was found useful to capture effects of phonological neighbours: When a new word shares phonological chunks with other familiar words, the new word enters the model lexicon more quickly than words with no phonological neighbours in the language.

As shown in the example above, the model is also made sensitive to utterance boundary information by attaching utterance boundary markers ("↲") to utterance-initial ("↲wɛrz") and utterance-final chunks ("dæd↲"). This facilitates future segmentation into word-like units. For example, a third utterance like "dædɪzkʌmɪŋ" ("dadiscoming") would be segmented as "dæd↲ | ɪ | z | k | ʌ | m | ɪ | ŋ", from which the model can start learning chunks that include demarcated word boundaries (i.e., the first chunk learned separates the word "dad" from subsequent phonological material: "dæd↲ɪ"). As explained above, CLASSIC-UB progressively constructs larger and larger chunks as a proxy for the increased processing efficiency derived from acquired knowledge. For the same reason, longer chunks are preferred for encoding the input over shorter ones. Then, the function of chunks that include demarcated word boundaries is key to the model because it prevents the building up of multi-word undersegmented chunks: The model leverages utterance boundaries to build multi-word chunks that also retain knowledge of the individual words composing the sequence, ultimately facilitating segmentation at the word level.

PUDDLE (Monaghan & Christiansen, 2010) is a chunking model that focuses on the role that lexical frames have in language learning, from segmentation to grammatical categorization. The model starts from the assumption that infants' early lexicons comprise sound sequences that occur frequently and that might not be

internally specified (e.g., Arnon, 2021). These lexical frames could comprise words (appearing in one-word utterances) but also multi-word sequences, and are initially extracted from the speech stream as whole unanalysed units using cues at frame edges (e.g., utterance-boundary cues). Once a diverse vocabulary of frames is acquired, the infant might start noticing similarities across frames and thus discover word boundaries within them (e.g., having encountered the one-word utterance "hello" might be useful to discover the word "baby" in "hellobaby"). Implementing these ideas, PUDDLE begins by representing whole utterances as single unanalysed chunks. For each chunk stored in the lexicon, information about its frequency of occurrence in the input is recorded (i.e., level of memory activation) and used to privilege extraction of frequent chunks in subsequent segmentations. Further, the model tracks which biphone sequences appear at chunk edges, and uses this information to constrain future segmentations (i.e., a chunk is identified within an utterance only if it is surrounded by sequences that previously began or ended other stored chunks). Given the use of whole-utterance frames at the beginning of the model learning, this tracking of biphone sequences at chunk edges essentially leverages knowledge of sounds that appear at utterance boundaries.

Finally, we used two random baseline models that processed the input as strings of phonemes or syllables, respectively. For each input utterance, the models randomly placed a word boundary after each input unit based on a coin toss. These baselines are informative as they tell us how much of the input vocabulary could be segmented and learned by chance if the infant made random guesses about word boundaries. The only information that constrained the random models was the type of input unit (phonemic or syllabic) and the utterance boundaries (that are given for free as the models processed one utterance at a time). Therefore, comparing segmentation models with random models can tell us how much variability in word segmentation and word learning is additionally captured by the transitional probability and chunking mechanisms of interest.

3.4.2 Input Corpora

As models' input, we used Italian utterances directed to children of up to 2 years of age available in the CHILDES database (MacWhinney, 2000). The corpora were Klammler (Klammler & Schneider, 2011), Antelmi (Antelmi & Morlacchi, 2005), Roma (Volterra, 1984), D'Onorico (D'Odorico & Carubbi, 2003), and Tonelli (Tonelli et al., 1998). The characteristics of the aggregated input are shown in Table 2. In terms of key characteristics relevant for our study, the average token syllabic length (1.78) was consistent with other studies that used Italian input (e.g., 1.83 in Saksida et al., 2016, 1.80 in Gervain & Guevara Erra, 2012). This average length meant that 57% of word tokens and 97% of word types in the input were multisyllabic. The input contained more verb tokens than noun tokens, confirming the verb dominance in Italian child-directed input. However, we found a similar proportion of verb and noun types, which contrasts with other studies that have found roughly twice as many verb types as there were nouns (Longobardi et al., 2015; 2016). It is unclear what determined this difference. It might be that previous studies have focused on an age that is at the low end of the range considered here (16 months), at which the input might contain more verb types. It is also possible that other studies have overestimated the proportion of verb types as they considered a word sample (*N range* = 340 - 407) that is 11 times smaller than ours (*N* = 4,408). Despite this difference, the balance between nouns and verbs in our sample still indicates that if a computational exhibited a noun advantage, this could not be attributed to noun frequency. Moreover, the proportion of noun and verb types is still quite different from what found in English (Jones et al., 2023), where at a similar child age range the proportion of input noun types is twice that of verbs.

Finally, our input confirmed that nouns appeared in utterance-final position more frequently than verbs, that most verbs appeared in medial position, with some also appearing in utterance-initial position.

Table 2 The table displays the total number of input utterances used after being transcribed phonetically (Utterances); The mean length of utterance in number of

words (MLU); The number of words including repetitions (Tokens); The number of unique words (Types); The target child age range in months (Age); The percentage of noun/verb tokens and types; The average length of a word, noun, or verb in number of phonemes or syllables (All words, Nouns, and Verbs) considering either tokens or types; The percentage of utterances in which a noun or verb appeared in utterance-initial, medial, final position, or in a one-word utterance.

	Utterances	**MLU**	**Tokens/Types**	**Age**
	22,190	4.24	94,146 / 4,408	16 - 36

	Tokens (%)	**Types (%)**
Nouns	15	41
Verbs	26	42

Mean length	**Phonemes (Tokens/Types)**	**Syllables (Tokens/Types)**
All	3.89 / 6.64	1.78 / 2.82
Nouns	5.90 / 6.73	2.54 / 2.86
Verbs	4.35 / 6.91	2.08 / 2.94

Utterances (%)	**Initial**	**Medial**	**Final**	**Isolated**
Noun	0.6	7	7.2	0.6
Verb	4.3	15.2	5.4	0.7

The procedure for preparing the input was the same as described for English in Cabiddu et al. (2023). We used the childesr package (Braginsky, Sanchez, et al., 2019), which provides the CHILDES utterances in orthographic form using a standardized procedure to treat special codes across corpora (e.g., prosodic, discourse markers). Then, the utterances were transcribed phonetically using the

PhonItalia lexicon (Goslin et al., 2014), which contains phonological and syllabic forms for 120,000 Italian words. Information about word stress was excluded from phonetic transcriptions. We only retained utterances for which all words had a correspondent phonetic form in the reference lexicon. As a final step, we randomly shuffled the utterances to control for differences in mean length of utterance across transcripts.

We also coded word forms into morphemes using the Italian section of MorphyNet (Batsuren et al., 2021) and SIGMORPHON 2022 (Shared Task on Morpheme Segmentation, Batsuren et al., 2022), which are large multi-lingual databases of root words, inflectional, derivational, and compound morphology. Note that the models only processed phonemic or syllabic input, therefore morpheme conversion only served later morphological analyses (i.e., to examine whether an input word was segmented into morphemes by a model). We discarded 15% of input word types (N = 680/4,408) for not having a corresponding morpheme entry in the databases. Given that 39% of CHILDES utterances did not include part-of-speech tags and 81% of input word types had only one possible morpheme segmentation, we used a non-contextual method for morpheme conversion. In cases where multiple morpheme segmentations were possible, we considered all alternatives as correct: For example, the form "acceso" (lit, turned on) can be segmented into different sets of morphemes depending on its role in a sentence. In "il caminetto acceso (the lit fireplace)" the word is correctly segmented as the monomorphemic noninflected "acceso" (masculine singular adjective), while in "lei ha acceso il caminetto (she has lit the fireplace)" the word is considered correctly segmented as the bimorphemic "acce | so" (verb past participle). Given our non-contextual method of conversion, in "il caminetto acceso", the word "acceso" was considered correctly segmented as either "acceso" or "acce | so". In Appendix S14, we show that the results of the morphological analyses did not change when excluding words with alternative morphological segmentations.

3.4.3 Model Evaluation Measures

3.4.3.1 *Pairwise Model Comparisons*

We used the set of evaluation measures introduced by Cabiddu et al. (2023), with the addition of two new measures looking at models' vocabularies by part of speech and number of morphological units acquired. We measured models' segmentation accuracy by first calculating precision (Words segmented in an utterance / Sequences segmented in an utterance) and recall scores (Words segmented / Words in the input). These measures assessed the accuracy of the models in discovering input words. As for English, we carried out pairwise Welch's t test comparisons taking the last 10,000 utterances as the target sample, at which the models' performance stabilizes (see Figure 7). We corrected p values and 95% bootstrap confidence intervals using Holm's correction.

We used multiple measures to assess model developmental plausibility. We related models' segmentation accuracy to children's word age of acquisition and first production. Models' segmentation accuracy was computed as the number of times a word type was correctly segmented from the input, divided by the frequency of the word type in the input. Child word age of acquisition was computed as the proportion of children that at 13 months of age were reported to understand a target word according to their caregivers. Reported comprehension scores for 436 word types were taken from the Italian Communicative Development Inventory norms (CDI, Caselli et al., 2012). The scores were downloaded from the Wordbank repository (Frank et al., 2017). This analysis considered a final sample of 289 word types, after filtering out those types that the models could not learn as they were not present in the child-directed input corpora.

Child word age of first production was instead estimated using the children's utterances (N = 10,372) available in the corpora of our study. The lowest mean length of utterance (MLU) of a transcript in which a target word type appeared was taken as the word stage of first production. Mean length of utterance was computed using a bootstrapping procedure described in Cabiddu et al. (2023), which controlled for differences in number of utterances across transcripts. Using MLU as the age of

first production is useful because it provides information about child gross linguistic skills, controlling for the fact that children of the same age can be far apart in their developmental stage. This analysis considered a final sample of 1,653 word types.

Given the focus of the age of first production estimation on linguistic competence rather than age, the nature of this corpora-based measure is different from the CDI-based one. To also include a measure of age of first production based on age, we repeated the estimation procedure described in Cabiddu et al. (2023) but using age in months rather than MLU as the stage of first production. This analysis, included in Appendix S15, returned results consistent with the MLU-based measure.

To compare models' performance in word age of acquisition and age of first production, we fitted separate linear regression models predicting each of these two outcome variables as a function of model segmentation accuracy. We then computed pairwise differences in models' adjusted $R^2$to compare how much variance in the outcome was explained by each segmentation model. We bootstrapped the 95% confidence interval of the difference between each pair of adjusted R^2, and corrected the intervals using Holm's correction. We concluded that two segmentation models did not differ in the amount of variance explained if the confidence interval of their comparison included 0.

We also compared the word types learned by each model with the ones produced by Italian children in the corpora. We compared models and children's distributions of word types by four word-level characteristics: Phonemic length, frequency, neighbourhood density, and phonotactic probability. Some evidence exists that Italian word acquisition is related to characteristics of word frequency and word length (Braginsky, Yurovsky, et al., 2019), with more frequent and shorter words being produced at an earlier age by Italian children. However, no studies have investigated whether Italian children are more likely to learn dense neighbourhood words and words with high phonotactic probability (when one controls for word length) as found for English (Jones et al., 2021). Some studies on the effect of these variables in Italian adults' nonword repetition suggest that similar effects might be found in children (Arduino & Burani, 2004; Bracco et al., 2015). In fact, Italian adults are faster at repeating dense neighbourhood and high phonotactic probability words

(e.g., Arduino & Burani, 2004). This facilitation in processing might also influence early word acquisition.

Word length was computed as the number of phonemes in a target word. Word frequency was the log10 frequency of a target word in the child-directed input. Neighbourhood density was the number of words in the input that differed from a target word by deletion, substitution, or addition of a single phoneme. Phonotactic probability was the mean probability of a target word's phoneme pair to appear in the child-directed input. In Cabiddu et al. (2023), word frequency, neighbourhood density, and phonotactic probability were correlated with word length. We found correlations in the same direction in Italian child-directed speech (r_s word length, word frequency= -.23 [-.28, -.19]; r_s word length, neighborhood density = -.67 [-.69, -.63]; r_s word length, phonotactic probability = .26 [.21, .31]), although weaker, with a difference of approximately |.1| - |.2| compared to their English counterparts. Therefore, as in Cabiddu et al. (2023), we controlled for the effect of word length by dividing a target word frequency, neighbourhood density, or phonotactic probability value by its length.

We conducted a chi-square goodness of fit test to compare the observed probabilities of encountering a word type at each phonemic length (in a model output) to the expected probabilities in children's productions. Next, we examined the pairwise differences in chi-square test statistics, using bootstrap confidence intervals as previously described. This analysis first examined how closely each model matched children's performance, to then use these distance estimates to compare the models to one another.

To compare models' fit to children in the continuous measures of word frequency, neighbourhood density, and phonotactic probability, we followed a similar procedure as with the word length measure, but we used a Kolmogorov-Smirnov test statistic.

We also carried out additional analyses beyond those proposed by Cabiddu et al. (2023). To answer the question whether sensitivity to utterance-final cues could determine a noun advantage in Italian vocabulary learning, we compared the

models' vocabularies by part of speech categories. We took all token occurrences of a target word in the input and chose the most frequent tag as the part of speech category for that target word type. The average coverage of the most frequent tag was 98% (*SD* = 7%). Further, to compare the noun advantage in models and children, we first computed the difference between the proportion of noun and verb types in children and models' distributions. This first step gave us a measure of noun advantage over verbs. For example, the noun advantage in children was computed as P = 47% (nouns) – 31% (verbs) = 16%. Then, we looked at the difference in noun advantage between children and each model (ΔP), to examine whether they differ in the size of the advantage. We bootstrap the corrected 95% confidence interval for ΔP and concluded that the noun advantage in children and model was significantly different if the interval did not include 0.

Finally, to examine if models that oversegment the input discovered morphological units and if models that captured child vocabularies acquired morphological units alongside word forms, we calculated how many morpheme tokens and types were discovered by each model.

3.4.3.2 *Comparing by input type*

We examined whether a model performed better when run on phonemic or syllabic input. First, we calculated how much variability in a certain measure a model explained beyond what could be explained by chance (baseline). This initial step ensured that differences between phonemic and syllabic model versions could not be attributed to the fact that syllabic input is easier to segment due to presence of a lower number of boundaries to estimate. Second, for each model, we carried out a comparison between input types.

For example, for the accuracy measure of precision, we took the t values that referred to the comparison between precision scores of a phonemic or syllabic model and each correspondent random baseline (e.g., t value for phonemic CLASSIC-UB final vs. phonemic random baseline, and t value for syllabic CLASSIC-UB final vs. syllabic random baseline). Then, we took the difference between phonemic and

syllabic t values (Δt), which gave us a measure of whether a model explained more variability when run on phonemic or syllabic input, while also controlling for chance levels within each input type. We computed the corrected confidence interval for Δt. If the interval did not include 0, we concluded either that CLASSIC-UB final explained more variability when run on phonemic input (positive Δt) or that the model explained more variability when run on syllabic input (negative Δt).

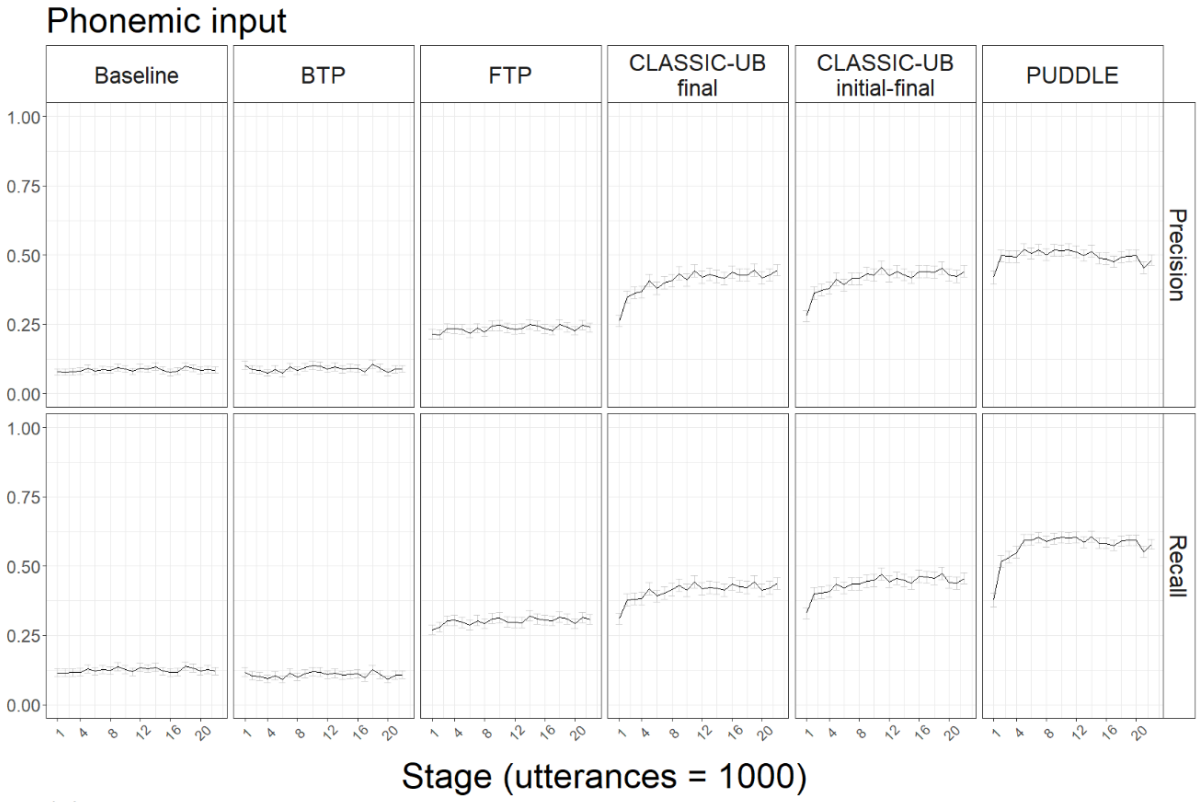
We applied the same logic to the other measures but using their reference statistics (i.e., adjR^2 for age of acquisition/first production, X^2 for phonemic length, and Kolmogorov-Smirnov *D* for word frequency, neighbourhood density, and phonotactic probability.

3.5 Results

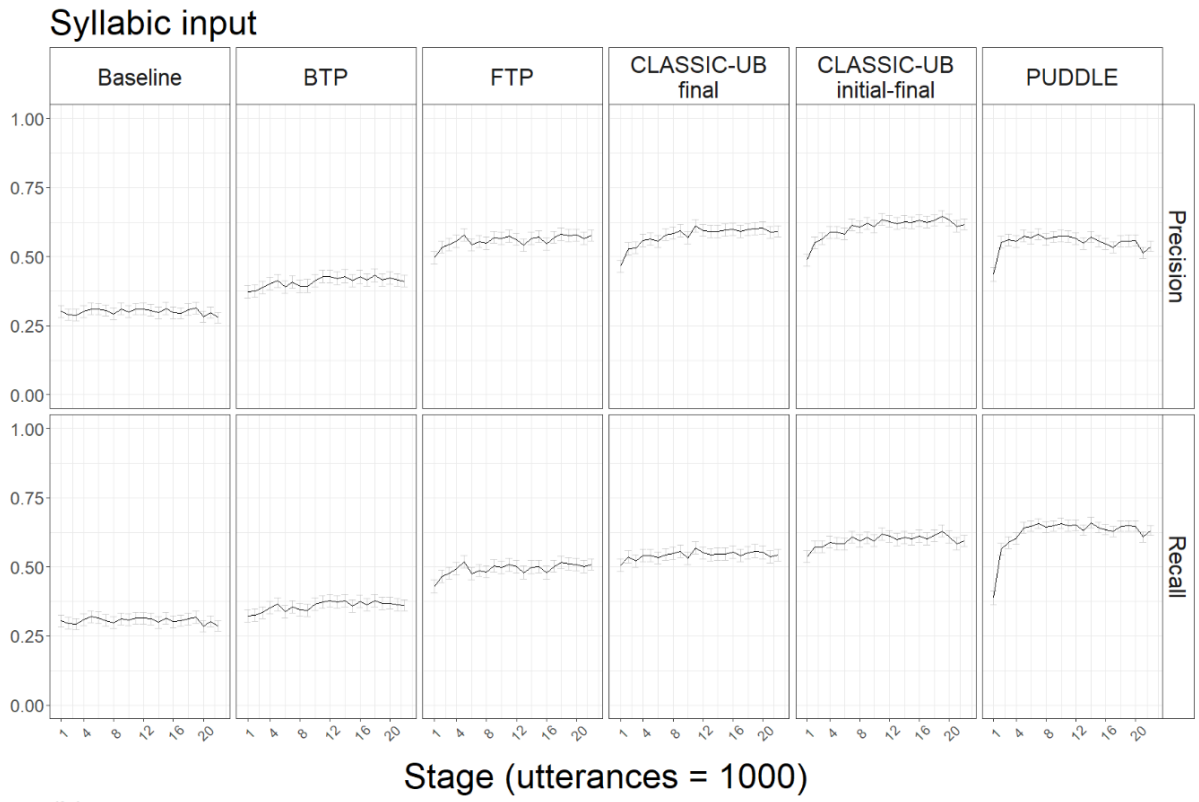
In the following section, we present results for accuracy and developmental measures of performance. We only point to statistical results included in Appendix S16-S22.

3.5.1 Precision and Recall

In Figure 7, we display phonemic and syllabic model segmentation accuracy incrementally, with average precision and recall at every 1,000 utterance steps. The performance of some models (CLASSIC-UB for phonemic and syllabic input, and transitional probability for syllabic input) showed a positive trend indicating that these models might have not reached a plateau. Given that, overall, models applied to English speech reached a plateau approximately after 40,000 input utterances (Cabiddu et al., 2023), it is possible that the upward trend is due to the limited sample size available for Italian.



(a)



(b)

Figure 7 Mean precision and recall performance with phonemic (Panel A) and syllabic (Panel B) input. The figure shows the random baseline, backward

transitional probability (BTP), forward transitional probability (FTP), CLASSIC-UB with utterance-final and initial-final markers, and PUDDLE. Performance was averaged every 1,000 utterances (Stage). Error bands for each stage indicate the 95% confidence interval around the mean.

Overall, the models segmented with lower accuracy in Italian compared to English, which was expected given previous findings showing negative correlations between models' segmentation accuracy and input average word length (e.g., Saksida et al., 2016) and morphological complexity (e.g., Loukatou et al., 2022). For example, in Cabiddu et al. (2013), the model with the best accuracy was PUDDLE, whose accuracy scores ranged between 73% and 89%. With Italian, instead, PUDDLE reached a maximum of 55% precision and 64% recall when syllabic input was used (see Figure 7b). Even baseline performance declined - with syllabic input, the performance of the random baseline reduced from 46% precision and 51% recall for English to 30% precision and 30% recall for Italian. However, despite overall lower accuracy, almost all models performed above chance (see Appendix S16), even syllabic transitional probability models that were instead found to perform worse than a random baseline in English segmentation (Cabiddu et al., 2023).

Specifically, all models performed above chance when syllabic input was used. With phonemic input, only the backward transitional probability model did not surpass the baseline in precision and recall (see top row of Figure 7a, and Appendix S16). This is line with analyses from Gervain and Guevara Erra (2012) showing that phonemic backward transitional probability performed at chance in Italian. Moreover, forward transitional probability always performed better than backward transitional probability (see Appendix S16), in line with results showing that tracking forward relations is more beneficial for segmenting words in head-initial languages like English (Cabiddu et al., 2023) and Italian (Gervain & Guevara Erra, 2012).

The models with the overall best performance across measures were the chunking models (see Figure 7 and Appendix S16), in line with what found in English (Cabiddu et al., 2023).

Next, we looked at the improvement in accuracy of each model beyond chance (baseline) when processing phonemic or syllabic input. As can be seen in Appendix S16, chunking models' relative improvement in accuracy above baseline was always higher when processing phonemic input (positive Δt values in both precision and recall). We found mixed results for transitional probability models. Backward transitional probability had a larger relative improvement when processing syllabic input (across accuracy measures), while forward transitional probability relative improvement was larger when processing syllabic input in precision, while larger when processing phonemic input in recall.

We also found mixed results regarding the role of utterance-boundary cues. In English (Cabiddu et al., 2023), adding utterance-initial cues always improved CLASSIC-UB's segmentation accuracy scores (beyond facilitation from utterance-final markers); here we found a similar advantage of CLASSIC-UB initial/final over CLASSIC-UB final, but only when syllabic input was used (see Appendix S16). This result was related to rates of oversegmentation across models, with CLASSIC-UB initial-final presenting higher oversegmentation than CLASSIC-UB final (see Figure 8a). Given the higher morphological complexity of Italian compared to English, it is possible that utterance-initial cues still benefitted segmentation, but at the morpheme level. In Figure 8b, we examined the number of morpheme tokens discovered by each model. When phonemic input was used, the addition of utterance-initial boundary cues led to discovery of a higher number of morphemes within words, which are the ones contributing to oversegmentation. The two models did not differ in the types of morphemes extracted, rather, certain morphemes were extracted more frequently in CLASSIC-UB initial/final compared to CLASSIC-UB final. The morphemes that contributed the most to the difference between models were function words ("e", "un", "a", "i"), followed by words that often occurred in utterance-initial position and that contained morphemes homophonous to those frequent function words ("cos | a", "tutt | i", "vien | i"). We return to this result in the Discussion.

The difference between CLASSIC-UB final and CLASSIC-UB initial/final was instead less evident when using syllabic input. In syllabic segmentation, the benefit

of using utterance-initial cues in segmentation might be more difficult to detect, because Italian within-word morphemes are mostly intra-syllabic (i.e., they can only be detected by phonemic models, e.g., Gervain & Guevara Erra, 2012). Indeed, a much lower number of within-word morphemes was discovered across models when syllabic input was used (see Figure 8b).

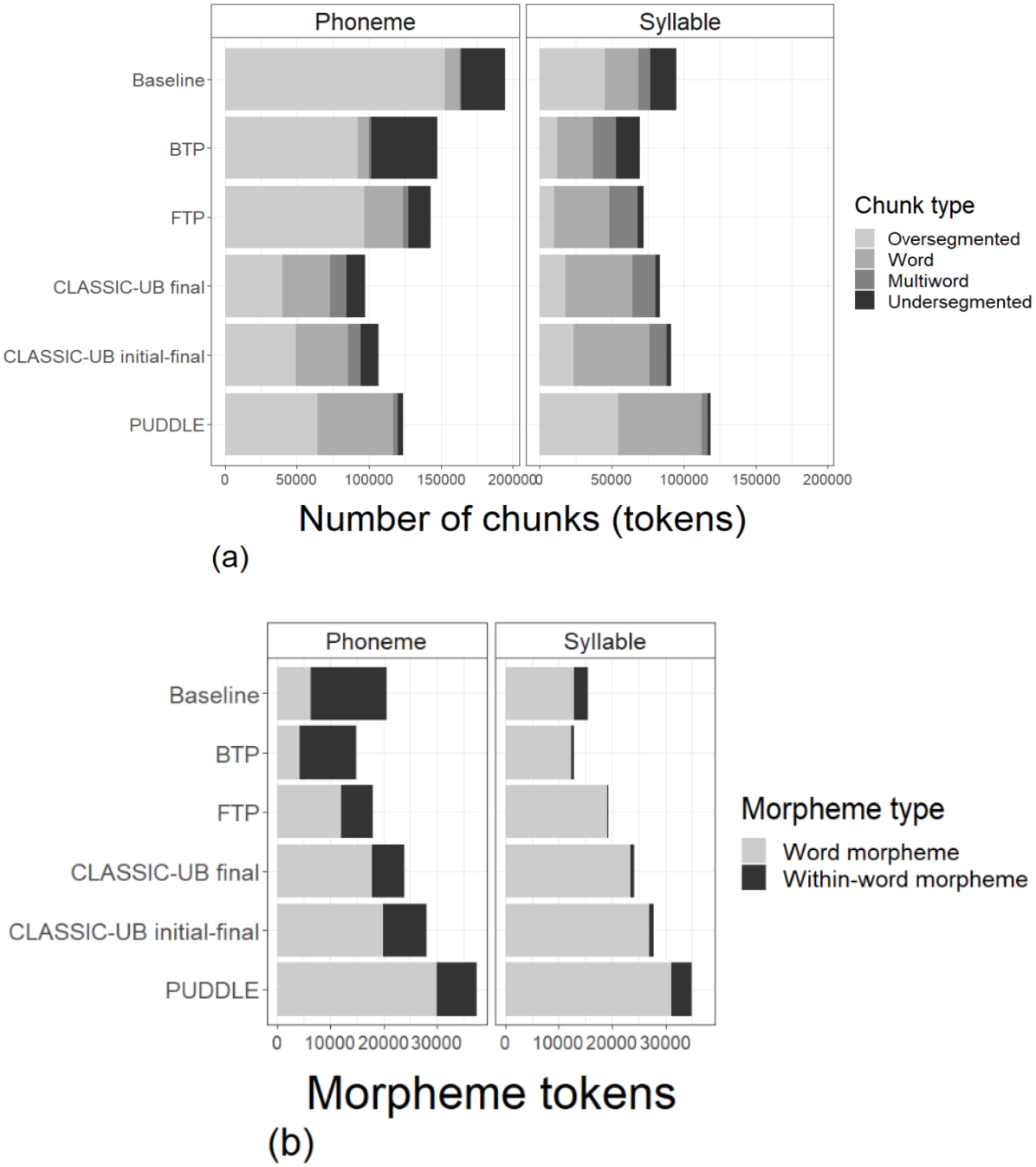


Figure 8 Panel A shows the number of tokens segmented by chunk type. Chunk types were defined as correctly identified Words, within-word Oversegmented

chunks (e.g., segmenting "og" from "dog"), correctly identified Multiword chunks (e.g., segmenting "thedog" from "thedog"), and Undersegmented chunks including at least one real word plus a part-word sequence (e.g., segmenting "edog" from "thedog"). Panel B shows the number of correctly segmented whole-word and intra-word morpheme tokens by model and input unit. FTP is forward transitional probability, BTP is backward transitional probability.

The morpheme analysis also indicates that models that oversegmented the most (transitional probability and baseline models) more likely discovered real within-word morphemes. However, better within-word morpheme segmentation was achieved at the expense of word-level segmentation accuracy in these models.

These analyses informed us about the accuracy of the models, suggesting that chunking performed better than transitional probability and baseline in word segmentation. Also, we found that overall, when controlling for chance levels, models segmented input words with higher accuracy when processing phonemic input. Now, we turn to results from developmental measures that assessed how well models' accuracy related to child data.

3.5.2 Word Age of Acquisition and Production

The first set of developmental measures focused on the timecourse of word acquisition, assessing whether segmentation models' accuracy scores could be used to predict how early a word entered children's comprehension or production vocabularies. When examining the word age of first production measure (Table 3a), no model significantly surpassed the random baseline model when run on Italian syllabic input (see pairwise comparisons in Appendix S17), similarly to that found for English (Cabiddu et al., 2023). Instead, we found that CLASSIC-UB models applied to phonemic input were the only models to significantly surpass the baseline. Also, no difference was found between CLASSIC-UB final and initial/final. The proportion of variance explained by the best model CLASSIC-UB final (AdjR^2 = .083 [.048,

.119]) for Italian (see Table 3a) was similar to that found for English (AdjR^2 = .079 [.062, .100], Cabiddu et al., 2023).

When using the age of acquisition measure based on the Italian CDI scores as the outcome, we found that none of the segmentation models explained any significant amount of variance (see Table 3b). This is in line with what found by Cabiddu et al. (2023) for English, and like for English this null result for Italian is most likely due to the limited size of the sample of word types that could be entered into this analysis (289 in Italian, 330 in English).

Focusing on age of first production, we conducted an additional analysis to assess the contribution of phonemic and syllabic input in the models' ability to capture variance in children's timecourse of word production. As shown in Appendix S17, across models, the relative improvement (beyond chance) in predictive power was higher when processing phonemic input (positive ΔAdjR^2), in line with what found in the precision and recall accuracy measures. However, the difference between phonemic and syllabic input reached significance only for the forward transitional probability model (ΔAdjR^2 = 0526 [.0176, .0894]). We suspect the lack of significance might be due to the limited sample of words used in Italian compared to English (1,653 vs. 5,480). To further examine this, we used the data from Cabiddu et al. (2023) and carried out this analysis by input type on English. We found that all segmentation models predicted the most variability in English age of first production when processing phonemic input, and the difference between phonemic and syllabic input was significant across models (see Appendix S17). This result suggests that if we were to analyse larger sample sizes in Italian, the positive contribution of phonemic input to predicting the timecourse of Italian word production might be confirmed.

Table 3 Adjusted R^2 for linear regression models predicting word age of first production (Panel A) and word age of acquisition (Panel B) as a function of weighted log10 number of times a word was correctly segmented by each model.

Panel A	Word Age of First Production			
	Phonemic input		Syllabified input	
Model	$R^2_{adjusted}$	95% CI	$R^2_{adjusted}$	95% CI
Baseline	.024	[.007, .049]	.035	[.015, .064]
Backward transitional probability	.018	[.006, .033]	.010	[.002, .026]
Forward transitional probability	.044	[.026, .068]	.003	[-.001, .011]
CLASSIC-UB final	.083	[.048, .119]	.051	[.023, .086]
CLASSIC-UB initial/final	**.**072	[.046, .105]	.046	[.023, .074]
PUDDLE	.028	[.010, .051]	.025	[.009, .046]
Panel B	Word Age of Acquisition			
	Phonemic input		Syllabified input	
Model	$R^2_{adjusted}$	95% CI	$R^2_{adjusted}$	95% CI
Baseline	.020	[-.003, .081]	-.002	[-.003, .029]
Backward transitional probability	-.002	[-.003, .032]	.001	[-.003, .030]
Forward transitional probability	.010	[-.003, .063]	-.002	[-.003, .020]
CLASSIC-UB final	-.001	[-.003, .029]	.004	[-.003, .060]
CLASSIC-UB initial/final	**.**005	[-.003, .060]	.002	[-.003, .046]
PUDDLE	.005	[-.003, .049]	.001	[-.004, .037]

Note. Heteroskedasticity-robust standard errors were computed using a HC2 estimator. The 95% confidence intervals indicate lower and upper limits of bootstrap confidence intervals around the estimate based on 1,000 iterations. Holm's correction was applied by expanding the confidence intervals.

3.5.3 Word-Level Characteristics

In Table 4, we first report the size of the vocabulary acquired by each model. Even if the models were exposed to a language where most words were multisyllabic, the models still benefitted from processing the input in syllable units, with some models (baseline, forward transitional probability, and both CLASSIC-UB models) acquiring larger vocabularies than in children's productions (N = 1,653). The models' relative advantage compared to children was similar to that found in English (Cabiddu et al., 2023). For example, compared to children's productions, CLASSIC-UB final acquired a vocabulary that was 1.21 times bigger in English and 1.27 times bigger in Italian.

Table 4 Number of word types learned by each model when run on phonemic or syllabic input.

	Word types	
	Phonemic input	Syllabified input
Baseline	532	1,747
Backward transitional probability	227	1,196
Forward transitional probability	318	2,754
CLASSIC-UB final	1,371	2,102
CLASSIC-UB initial/final	1,115	1,851
PUDDLE	1,225	1,359

Before proceeding with the comparison between word type distributions in models and children, we inspected the type of vocabularies acquired by children, compared to their child-directed speech input and the vocabularies of English children.

As can be seen in Figure 9, Italian children produced shorter (9a) and more frequent words compared to their input (9b), in line with effects of word length and frequency found in previous studies (Braginsky, Yurovsky, et al., 2019). Further, Italian children produced words with a higher number of neighbours in the language

(9c), and words with a higher internal predictability compared to their input (9d). This result is in line with the same effects of neighbourhood density and phonotactic probability found in Italian adults' word processing (Arduino & Burani, 2004), and on English children production vocabularies (Jones et al., 2021).

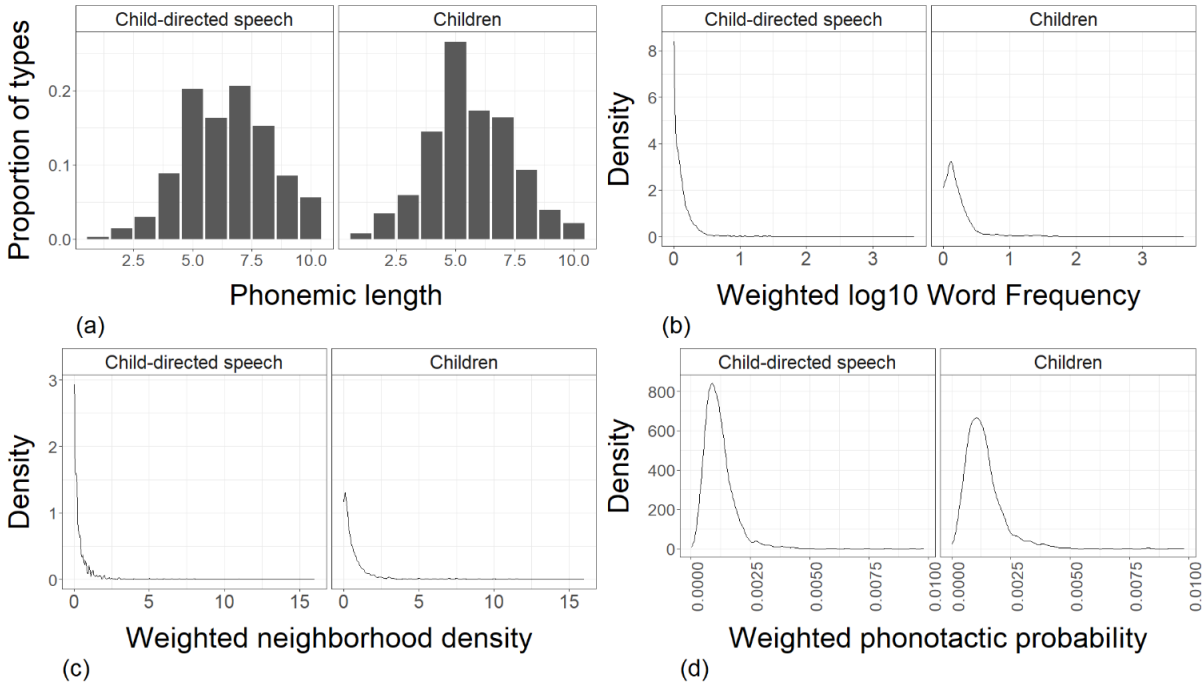


Figure 9 Panel A shows then proportion of word types in child-directed speech and produced by children by phonemic length. The other panels show the gaussian kernel density estimate of the distribution of word types by log10 word frequency (B), neighbourhood density (C), and phonotactic probability (D). The last three measures are weighted by dividing a target word value by its phonemic length. In the last three measures, the area under each curve represents 100% of data points, with curve peaks representing the mode of each distribution.

In terms of average word length, Italian child vocabularies contained longer words than in English child vocabularies, both in terms of number of phonemes (Italian: *Mean* = 5.68, *SD* = 1.84; English: *Mean* = 4.79, *SD* = 1.93) and syllables (Italian: *Mean* = 2.46, *SD* = .77; English: *Mean* = 1.95, *SD* = .92). Italian children also produced words with similar weighted log10 frequency in the input (Italian: *Mean* =

.22, *SD* = .30; English: *Mean* = .18, *SD* = .26), lower weighted neighbourhood density (Italian: *Mean* = .78, *SD* = 1.56; English: *Mean* = 1.41, *SD* = 2.34), and with higher weighted phonotactic probability (Italian: *Mean* = .0014, *SD* = .0009; English: *Mean* = .0009, *SD* = .0006)[7].

We now compare the word types learned by each model to those produced by children on the four word-level properties of length, frequency, neighbourhood density, and phonotactic probability. As for English (Cabiddu et al., 2023), no model surpassed the baseline model when the input was processed in syllable units. The baseline model reached ceiling in performance, with its distribution of word types not differing significantly from children in any of the word-level measures (see Appendix S18-S21; Phonemic length: X^2 = 7.22, *p* = .614; Word frequency: *D* = .043, *p* = .089; Neighbourhood density: *D* = .014, *p* = 1; Phonotactic probability: *D* = .027, *p* = 1). Given this ceiling effect, we did not carry out comparisons by input type, as the relative contribution of each model when processing syllabic input could not be assessed, and therefore it could not be compared to results based on phonemic input. We examine the reasons for this ceiling effect in the next exploratory section. Although, first, we discuss results on phonemic input in the remaining paragraphs of this section, and include syllabic results in Appendix S18-S21.

In line with that found for English (Cabiddu et al., 2023), chunking models performed better than transitional probability models at capturing all child word properties. Only CLASSIC-UB final performed better than the baseline at capturing the children's phonemic length distribution, and surpassed all other models apart from CLASSIC-UB initial/final (see Figure 10a, see Appendix S18). No difference was found between CLASSIC-UB models and PUDDLE at capturing child word frequency (Figure 10b, Appendix S19) and neighbourhood density distributions (Figure 10c, Appendix S20). However, chunking models performed better than baseline and traditional probability models, which mostly learned shorter words, with higher frequency, and higher neighbourhood density. Finally, only CLASSIC-UB final performed better than the baseline at capturing the children's phonotactic probability

distribution, and surpassed all other models apart from CLASSIC-UB initial/final (see Figure 10d, and Appendix S21).

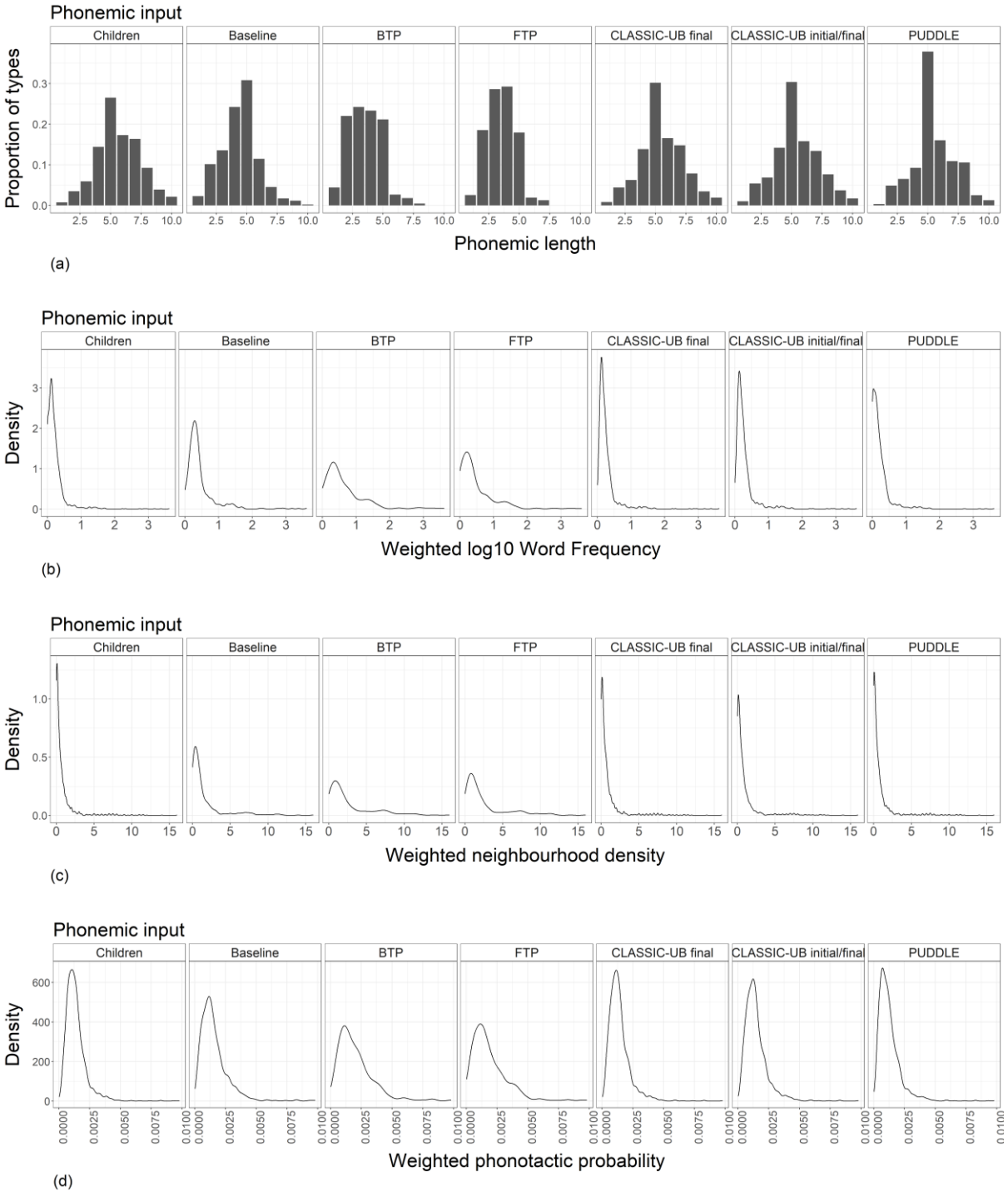


Figure 10. Panel A shows then proportion of word types produced by children and discovered by each model by phonemic length when phonemic input was used. The other panels show the gaussian kernel density estimate of the distribution of word types by log10 word frequency (B), neighbourhood density (C), and phonotactic

probability (D). The last three measures are weighted by dividing a target word value by its phonemic length. The area under each curve represents 100% of data points, with curve peaks representing the mode of each distribution.

As for English (Cabiddu et al., 2023), across word properties, information about utterance-initial cues did not improve CLASSIC-UB performance above utterance-final cues (see Appendix S18-S21). To further examine the utility of utterance-final cues for building an early vocabulary in Italian, we display the number of children and models' word types by part of speech category. As can be seen in Figure 11, CLASSIC-UB models and PUDDLE were the only models that consistently presented a noun advantage over verbs across both input types. We also statistically examined whether the size of the noun advantage over verbs differed statistically from that of children (Appendix S22). We found that the proportional noun advantage in every model was statistically smaller than children's noun advantage. We return to this result in the Discussion.

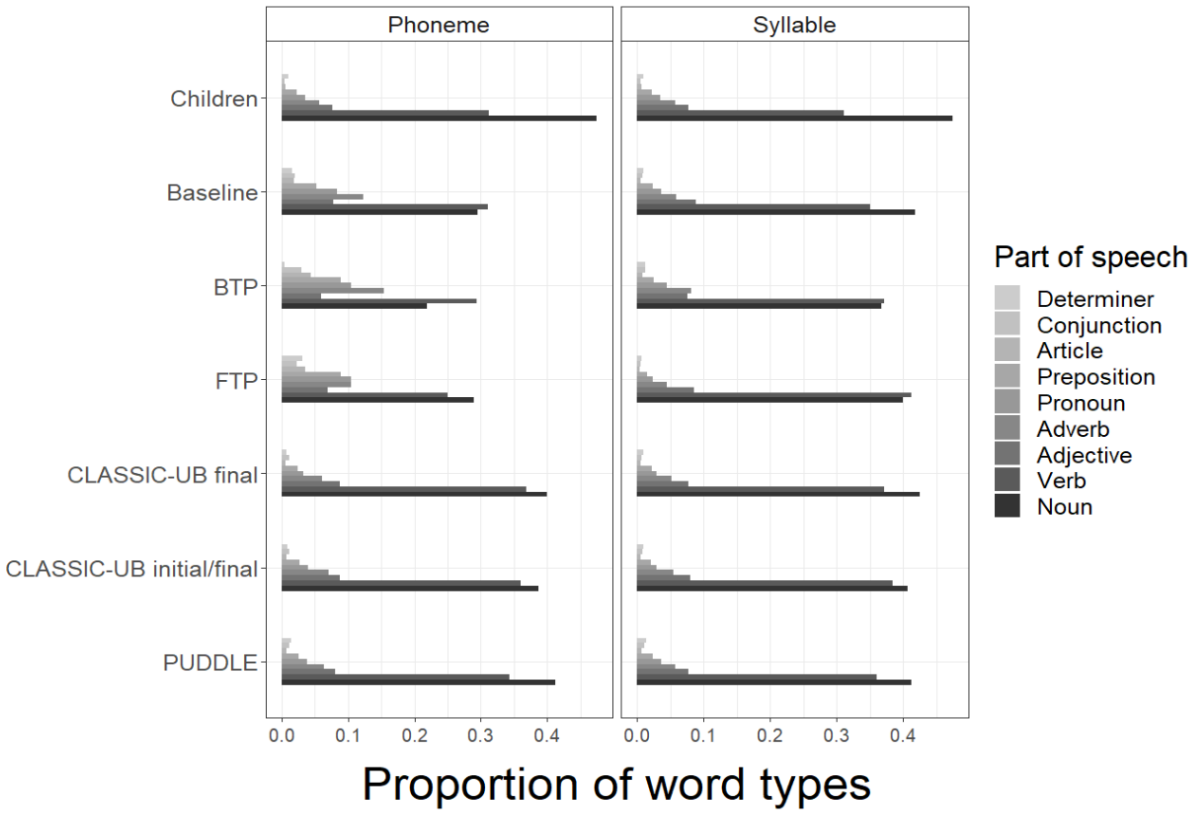


Figure 11 Proportion of word types acquired by children and each model, by input type and part of speech tag.

Finally, we explored the kind of morphological vocabulary built by each model. We focused on models run on phonemic input (Figure 12, left panel), that were the ones that surpassed the random baseline in measures that related model performance to child data. As can be seen in Figure 12, the phonemic models that overall performed better at capturing the course of vocabulary acquisition and the word properties of child vocabularies (i.e., CLASSIC-UB models) were also the ones that learned the largest number of morpheme types overall.

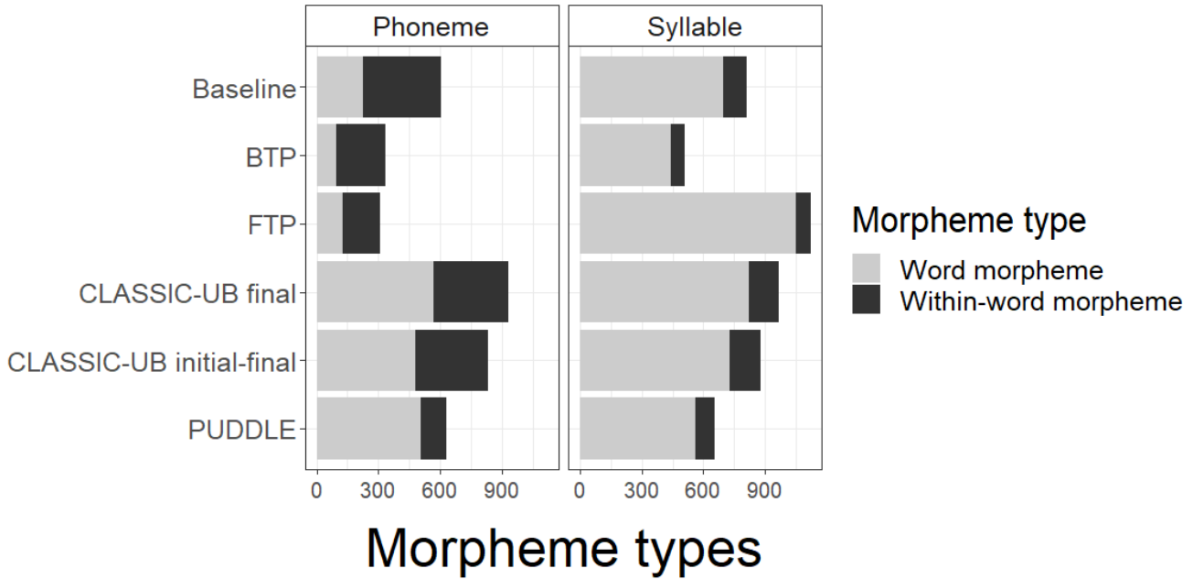


Figure 12. Number of morpheme types learned by each model, divided by morphemes corresponding to whole words (Word morpheme), and morphemes appearing as part of a word (Within-word morpheme).

3.5.4 Exploratory Analysis of Word-Level Properties at the Token Level

When assessing how well the models captured word-level properties of children's productions, we could not carry out comparisons by input type because the syllabic

baseline performed at ceiling. Using Italian speech had the expected effect on the syllabic random baseline, which learned a lower percentage of input word types (40%) compared to English (56%). However, similarly to English, the syllabic baseline still produced implausibly large vocabularies (Table 4) and fit children's word-level properties well (see Appendix S18-S21). In this section, we identified some potential explanations for this result.

First, our examination of differences between child-directed speech and children's word type distributions showed that children produced vocabularies composed of words that are easy to acquire (e.g., short, with high-frequency in the input; see Figure 9). This meant that a baseline could potentially fit children's data well even if some input word types were missed. To improve the sensitivity of the evaluation measures based on this point, one could try to increase the sample size of word types produced by children (using additional child utterances). For Zipf's law, the probability of finding low frequency word types in children's productions increases sharply when sample size increases (e.g., Cabiddu et al., 2023). This would translate into a much higher difficulty of segmentation across models, potentially increasing the sensitivity of the measures. Unfortunately, we could not increase sample size due to lack of additional corpora on CHILDES.

Second, if a baseline model correctly segments a word just once, that word will enter the lexicon and will be counted in the distribution of learned word types. This may not be a fair comparison to children's data. For instance, the word "playing" is frequently used in the input and often produced by children. If a baseline model segments "playing" correctly only one time, the word will be considered learned. However, we would expect such a word to be consistently segmented correctly, given its frequency in the input and in children's speech. The issue of not only identifying the word in the input but also consistently segmenting it is not addressed in the examination of word-level properties since these focus on the distributions of word types.

To address the above limitation, in the following exploratory analyses we reevaluated the word-level properties considering the distribution of word tokens. As can be seen across Figures 13-16 and in statistical results in Appendix S18-S21, the

syllabic baseline did not reach ceiling in any of the word-level properties with these new token-based measures (i.e., higher sensitivity). We also found that, apart from phonotactic probability, the syllabic baseline was surpassed in performance in these new token-based measures. Taking into account the consistency of segmentation (token repetitions) as well as whether a word type has been discovered by a model made the task significantly more difficult for a baseline, which segmented word tokens that were significantly shorter, more frequent, and with higher neighbourhood density than in children's token distributions.

We found that chunking models performed better than transitional probability models across measures and input types, with the best models being CLASSIC-UB initial/final and CLASSIC-UB final overall. Also, when using token-based measures, an advantage from including utterance-initial cues emerged, with CLASSIC-UB initial/final outperforming CLASSIC-UB final in the neighbourhood density measure with phonemic input, and in phonemic length, word frequency, and neighbourhood density with syllabic input. We return to this result in the Discussion.

Finally, these new token-based measures also allowed us to achieve sufficient sensitivity to assess the contribution of phonemic and syllabic input in each model's ability to capture child data. We found that, when controlling for chance levels, chunking models captured more variability in word-level properties when run on phonemic input across all word-level properties (see Appendix S18-S21). Transitional probability models showed similar results, with only two exceptions where no difference was found between phonemic and syllabic forward transitional probability to capture phonemic length (Appendix S18) and phonotactic probability (Appendix S21), and only one case where we found a larger contribution from syllabic input for backward transitional probability in word frequency (Appendix S19).

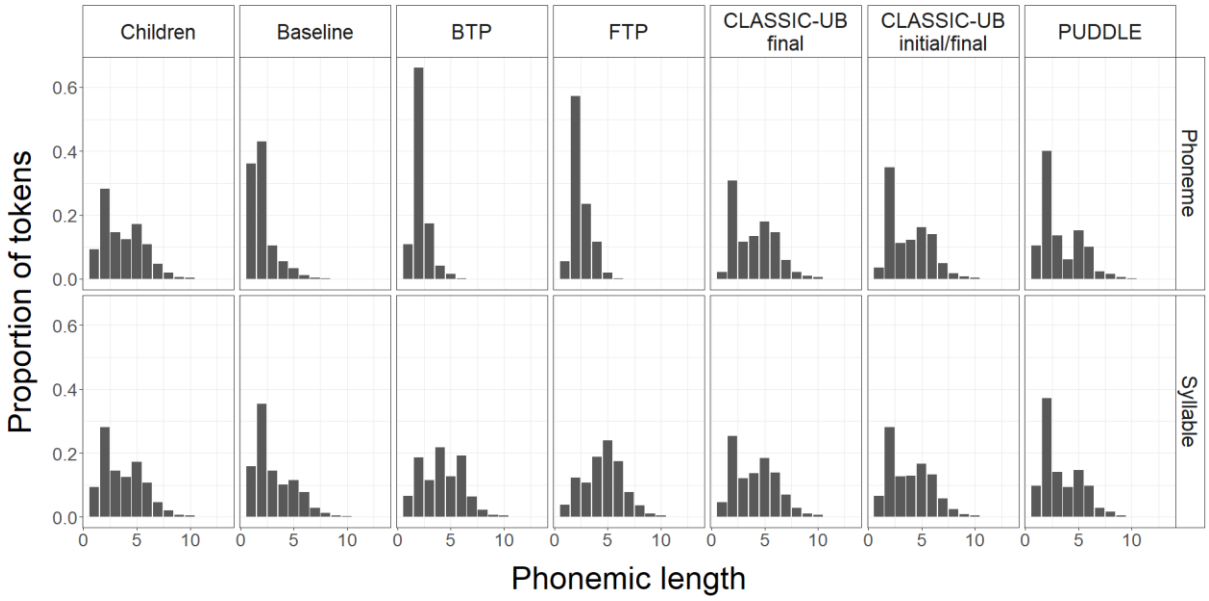


Figure 13 Proportion of word tokens at each phonemic length, produced by children and discovered by each model when phonemic and syllabic input was used.

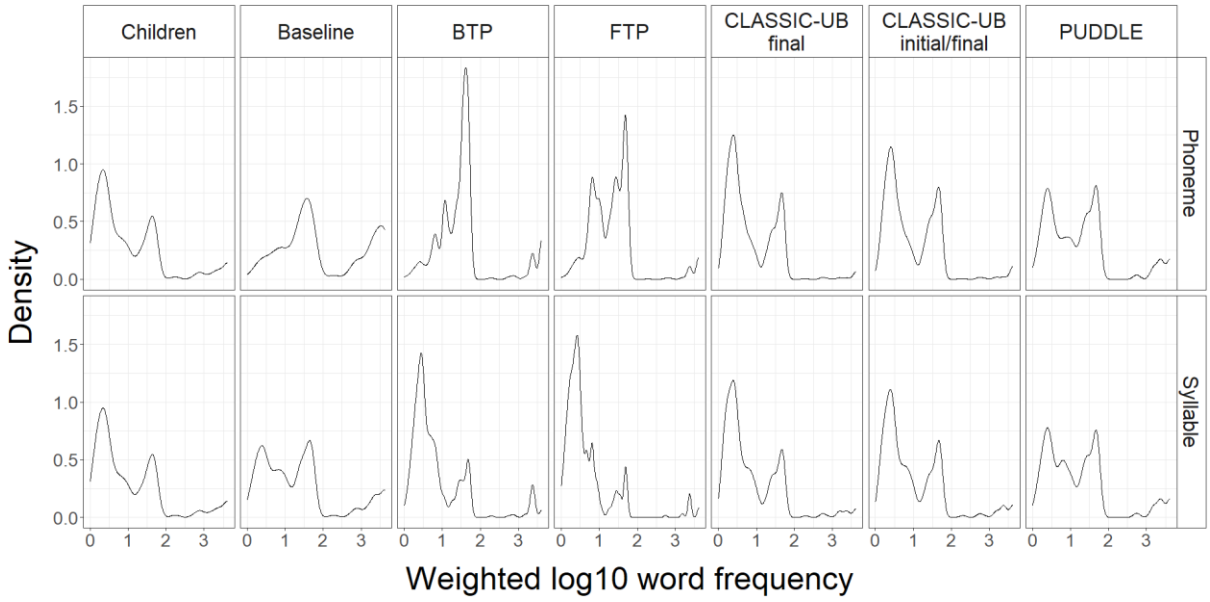


Figure 14 Gaussian kernel density estimate of the distribution of word tokens by weighted log10 word frequency, for children and models run on phonemic or syllabic input.

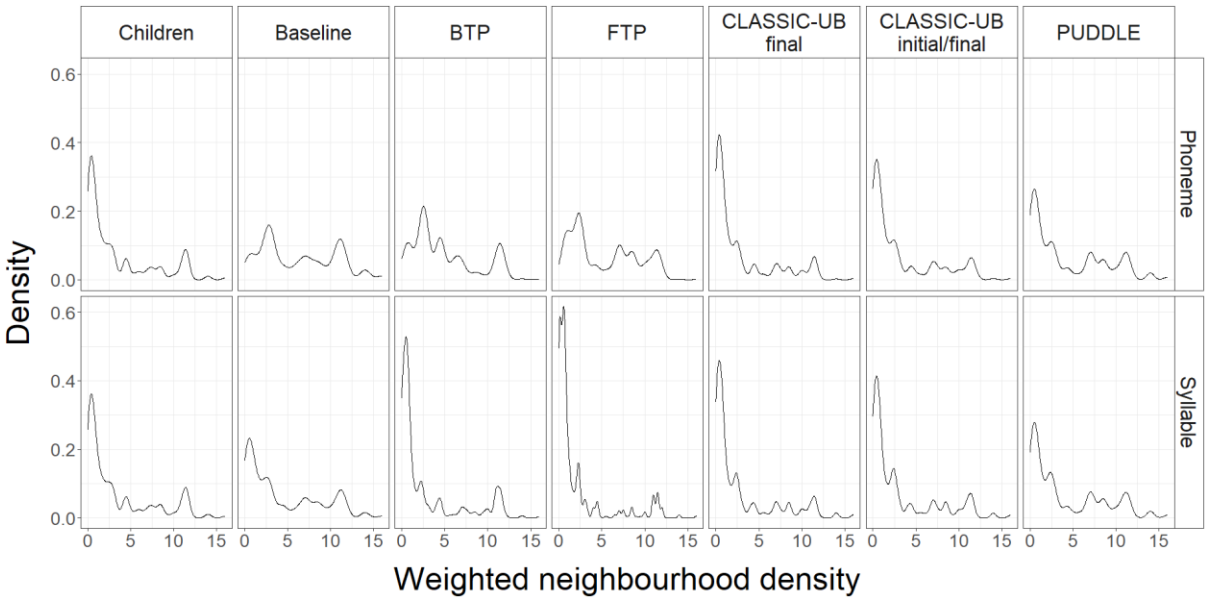


Figure 15 Gaussian kernel density estimate of the distribution of word tokens by weighted neighbourhood density, for children and models run on phonemic or syllabic input.

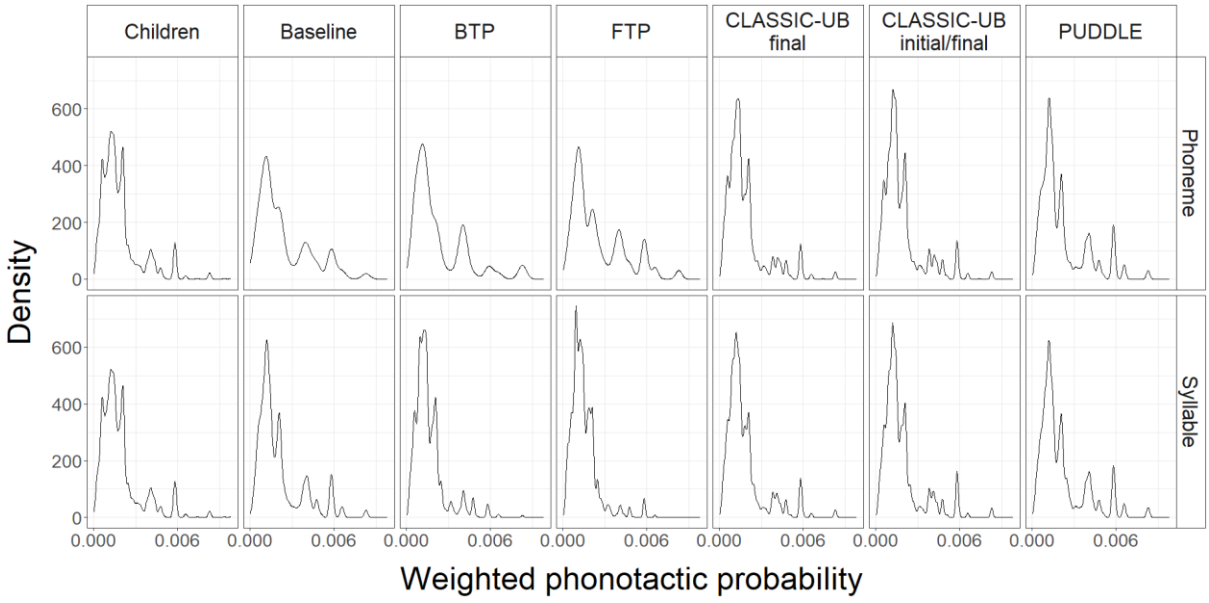


Figure 16 Gaussian kernel density estimate of the distribution of word tokens by weighted phonotactic probability, for children and models run on phonemic or syllabic input.

3.5.5 Summary of Results

Our investigation into the performance and developmental plausibility of segmentation models largely replicated previous results observed in English, using Italian child-directed input and child data. The results showed that chunking models outperformed transitional probability models in word segmentation. Additionally, chunking models explained a larger portion of variance in child developmental data. The study also provided new insights into the interaction between various input characteristics and segmentation learning mechanisms in capturing Italian vocabulary data.

Using Italian child-directed speech increased the sensitivity of the model evaluation measures, but this was observed only in segmentation performance. This increased sensitivity allowed us to examine whether segmentation models exhibited superior performance when exposed to phonemic or syllabic input. We found that phonemic input determined the most significant increase in segmentation performance in comparison to random baselines.

Similar to the findings in English, however, the random baseline model performed at ceiling when comparing word-level characteristics of models and children's vocabularies. A novel exploratory analysis suggested that this lack of sensitivity might be attributable to the fact that Cabiddu et al. (2023) inspected model and child vocabularies by focusing solely on word type distributions. These distributions only considered the introduction of a word into the lexicon, ignoring the consistency of its use. We addressed this issue by looking at word token distributions, which increased the sensitivity of the word-level measures. With these enhanced measures, we again observed that, overall, models best captured the variance in child vocabularies when exposed to phonemic input.

Furthermore, as for English, both utterance-initial and utterance-final cues played a role in segmentation performance. However, differently to what found for English, utterance-initial cues exhibited a less pronounced impact on word segmentation performance. Instead, they aided the discovery of part-word morphological units, particularly leveraging utterance-initial function words.

In line with the English findings, only utterance-final cues helped capturing the characteristics of child word type vocabularies. This supports the notion that the high diversity of new words appearing at the end of utterances in both English and Italian positively influences vocabulary development. Notably, we found that utterance-final cues induced a noun bias in models, even though the Italian input contained a higher proportion of verbs.

Finally, models proficient at capturing the characteristics of child vocabularies were also able to acquire morphological units.

In the following section, we offer a detailed discussion of each of these findings.

3.6 Discussion

We examined how well models of naturalistic word segmentation could discover words from Italian child-directed speech and capture different aspects of Italian children's vocabularies. Our main goal was to assess the plausibility of phoneme vs. syllable representations for capturing the shift from early word segmentation to vocabulary learning. We found that, for the best performing chunking models, higher segmentation accuracy was achieved when processing phonemic input. Comparisons by input type also showed a trend toward facilitation from phonemic input when predicting the timecourse of word acquisition, and an advantage for phonemic input when analysing word-level properties at the token level.

Moreover, we extended previous investigations of English cross-linguistically and found similar results for Italian, with segmentation mechanisms based on chunking performing better than transitional probability models across all measures used. We discuss these findings in the sections below.

3.6.1 Word Length

Our study aimed at increasing the sensitivity of previous measures that assessed models' segmentation accuracy and their ability to capture the composition of child

vocabularies (Cabiddu et al., 2023). To achieve higher sensitivity, we increased the difficulty of the segmentation task by using Italian child-directed speech, which presents higher ambiguity due to its average word length and morphological complexity. We found that, when segmenting Italian speech, the accuracy of the segmentation models overall decreased compared to English. At the same time, as we expected, the models' performance increased compared to random baseline models. For example, differently from English simulations, transitional probability models segmented speech above chance when syllabic input was used, in line with what found by Gervain and Guevara Erra (2012). However, despite the higher sensitivity in segmentation, we found that segmentation models exposed to syllabic input did not perform above chance in any of the developmental measures used (Appendix S17-S21), confirming the results found for English. The fact that, even in Italian, syllabic models failed to account for any variability in age of first production and child word-level properties suggests that the same results in English were not due to an artifact of word length. Additionally, the average word length in Italian was higher than in English, not only in the speech input but also in the children's productions. This finding rules out the possibility that the lack of sensitivity was due to similar word length distributions in child vocabulary across the two languages. To confirm this interpretation, future investigations could examine other languages with high average word length (e.g., German, Turkish, or Russian). Further, we carried out an additional exploratory analysis and found that we could increase the sensitivity of the word-level measures by considering the distribution of word tokens produced by children and segmented by models, rather than word types. Word token distributions are usually examined in child-directed speech and used as predictors of child outcomes (e.g., Hirsh-Pasek et al., 2015; Hoff & Naigles, 2002; Huttenlocher et al., 2010). In this study, we showed that the examination of word token distributions in child speech (i.e., consistency of child word production) might also be important to evaluate the fit of different hypotheses about the learning mechanisms of early segmentation and word learning.

When assessing models' segmentation accuracy in precision and recall, and their ability to predict child age of first production scores and token-based word-level properties, we found that overall models performed better when run on phonemic

input when controlling for chance levels (see Appendix S17-S21). These results are in line with studies showing how infants can leverage phonemic and phonotactic regularities to segment and learn words from the speech input (e.g., Fais et al., 2012; Jusczyk & Aslin, 1995; Mattys & Jusczyk, 2001).

3.6.2 Chunking vs. Transitional Probability

In line with English results (Cabiddu et al., 2023), we found that chunking models performed better than transitional probability models in all segmentation and vocabulary measures used, suggesting that chunking might also play a significant role in Italian vocabulary learning. These results suggest that the benefit from tracking lexical and sound combinations over transitional probabilities (e.g., French et al., 2011; Perruchet & Poulin-Charronnat, 2012) and the key role of chunking in early vocabulary learning (e.g., Jones et al., 2021) might apply to Italian.

Moreover, as for English, we found that CLASSIC-UB performed better than other models, likely because of its advantage in recycling phonological chunks at different grain sizes to more efficiently learn new words containing these familiar chunks (Cabiddu et al., 2023). The only difference was that, when examining the distribution of word types learned, in English CLASSIC-UB outperformed all other models in the neighbourhood density measure, but not in the phonotactic probability one. In Italian, we found the opposite. This result is likely related to the fact that the relation between word length and each of the two measures is not constant across English and Italian input, which can produce different effects across languages even when one controls for word length within each language, as we explain below.

In Italian speech, there is a prevalent presence of short biphone sequences. In contrast, English speech has a higher prevalence of longer sequences derived from phonological neighbours. CLASSIC-UB can employ either biphones or sequences from phonological neighbours to boost vocabulary learning. However, its preference for one sound combination over the other is determined by the predominant type in a given language. The predominance of a particular sound combination is influenced by the language's average word length: In languages with

longer words, like Italian, there are more biphones available to boost vocabulary learning. In contrast, languages with shorter words, such as English, contain more sequences from phonological neighbours. This distinction explains why we observed a marked advantage of CLASSIC-UB in capturing phonotactic probability in Italian, while an advantage in capturing neighbourhood density emerged in English.

To explain in detail, as word length increases, the probability of finding phonological neighbours of a target word decreases rapidly (see Appendix S23). For example, a short 4-phoneme word like *rest* has a diverse set of phonological neighbours (*best*, *chest*, *dressed*, *pressed*, *arrest*, *guessed*, *nest*, *rent*, *roast*, etc.). However, from 6-phoneme words onwards the probability of finding phonological neighbours reduces essentially to 0 (e.g., Pisoni et al., 1985; Storkel, 2004). This establishes a non-linear relation between word length and neighbourhood density. In our case, 95% of words in the Italian input had 0 to 5 phonological neighbours, while 95% of English input words in Cabiddu et al. (2023) had 0 to 17 neighbours. Put simply, Italian speech comprises longer words. These longer words have fewer phonological neighbours, leading to difficulties in identifying a significant effect of neighbourhood density. The reason is that there are not many neighbours providing sound combinations that could boost vocabulary learning.

In a similar non-linear fashion, as word length increases, phonotactic probability increases rapidly (see Appendix S23): That is, there is a much higher probability of finding frequent biphones in longer words, shifting the average biphone frequency of a target word upward. In other words, Italian speech comprises longer words than English does, and these words include more frequent biphones. CLASSIC-UB can use these biphones as sound combinations to boost vocabulary learning.

In sum, considering the definitions of neighbourhood density and phonotactic probability and their relation with word length, biphone sequences are more prevalent in Italian than sequences derived from phonological neighbours. While CLASSIC-UB taps into familiar sound combinations similarly in both English and Italian, its sensitivity to different sound combinations changes as a function of which specific sound sequences are available in the speech input (short biphones more

prevalent in Italian, longer sequences from phonological neighbours more prevalent in English). Although this interpretation would require further examination in future work, partial evidence exists of the fact that neighbourhood density effects on children's vocabularies are hard to find when focusing on longer words (Rajaram, 2022).

3.6.3 Utterance Boundaries

Our prediction that both utterance-initial and utterance-final cues would facilitate word segmentation was partially supported, with CLASSIC-UB initial/final surpassing CLASSIC-UB final only in the recall measure when phonemic input was used (Figure 7a, and Appendix S16). CLASSIC-UB initial/final did not perform better than CLASSIC-UB final in precision, due to higher rates of oversegmentation. However, we still found that such oversegmentation benefitted segmentation at the morpheme level. CLASSIC-UB initial/final learned a higher proportion of morphemes than CLASSIC-UB final. First, the model discovered more utterance-initial function words, in line with infants' sensitivity to function words with high token frequency in utterance-initial position (Gervain et al., 2008). Second, the model oversegmented the input leveraging frequent function words, which impaired its performance at the word level but benefitted morpheme segmentation. This result is in line with the prosodic bootstrapping hypothesis (Christophe et al., 1997), under which function words can be used to discover adjacent content words in speech (e.g., Johnson et al., 2014; Shi & Lepage, 2008). The same account also posits that function words might lead to oversegmentation errors (which in our case were found to be useful to discover inflections), because some of these might be recognized as homophonous sounds within other words (e.g., "i" in "tutt | i", "vien | i"), leading to restarting a lexical search. Importantly, we found that the decreased performance of CLASSIC-UB initial/final in word segmentation did not impact its performance at capturing aspects of child vocabularies.

Differently from English, Italian child-directed speech contains more verbs than nouns. In our input sample for example, 24% of input word tokens were verbs,

and 16% were nouns. However, as in English, nouns tend to appear often in utterance-final position (see Table 2). These characteristics of Italian allowed us to test whether unique sensitivity to utterance-final cues could produce a noun advantage as seen in Italian children. All chunking models used learning mechanisms that leveraged utterance-final cues, and we found that these models consistently learned more nouns than verbs, as children do (Figure 11). This result indicates that at least some variability in child word production might be explained by facilitation from utterance-final boundary cues (Longobardi et al., 2015, 2016), to which all chunking models used here were sensitive. However, the size of the advantage was not as marked for the models as in children (see Appendix S22 for a statistical comparison), indicating that other variables might be involved. For example, our models did not have access to real-world meanings, which likely facilitate children in learning nouns because these have conceptually simpler referents (e.g., objects) that can be easily identified perceptually (e.g., Gentner, 1982).

Finally, when examining the fit to word-level characteristics of child word types, we found that CLASSIC-UB did not benefit from tracking utterance-initial cues beyond utterance-final cues. This result was also found for English (Cabiddu et al., 2023) and it is in line with evidence that many different Italian words appear in utterance-final position, and these more likely enter child (word type) vocabularies compared to words in other positions (Longobardi et al., 2015).

Interestingly, however, we found that tracking utterance-initial cues facilitated CLASSIC-UB beyond tracking utterance-final cues when we analysed the models' fit to child vocabularies at the token level (see Appendix S18-S21). The reason for these contrasting results at the type and token levels lies in the fact that utterance-initial words have low type frequency and high token frequency. That is, in utterance-initial positions we find many function words like conjunctions, articles, or pronouns that are repeated frequently (high token frequency), but with the variety of distinct words in each of these categories being limited (low type frequency). As a result, utterance-initial cues became more beneficial when examining segmentation accuracy (Figure 7) or token-based word-level properties (Figure 13-16), which are

all measures that consider the output tokens discovered by the models (i.e., token frequency becomes more relevant). In contrast, when using measures that consider word types as the outcome (word age of first production, word type word-level measures), the low number of distinct function words appearing in utterance-initial position could not contribute significantly to improve CLASSIC-UB's performance. In sum, these results suggest that taking into account outcome measures based on both word types and word tokens might be important to assess the relative contribution of different cues to segmentation and vocabulary learning.

3.6.4 Morphology

Previous studies have found that word segmentation accuracy decreased when models were exposed to morphologically complex languages (Fourtassi et al., 2013; Johnson, 2008; Loukatou et al., 2018; 2019; 2022). In line with these studies, we found that segmentation models were less accurate at segmenting Italian speech compared to English, because of higher rates of oversegmentation which was related to discovery of morphological units (Figure 8).

This study was the first to assess whether segmentation models that better captured child vocabulary measures could also present evidence of morphological learning. We found that the more often models' vocabularies resembled children's (Figure 10), the larger the number of morpheme types they learned (Figure 12, phonemic input panel). Our simulations suggest that the same learning mechanisms sensitive to statistical regularities might help infants represent both word and morphological units. These results are also in line with studies showing that Italian infants start representing morphological units from 12 months of age, when they are also learning about words (Ferry et al., 2020).

3.6.5 Limitations and Future Directions

Our study suggests that processing the speech input at the sub-syllabic level might result in a more developmentally plausible output vocabulary, but there are

alternative explanations that would need to be tested and that could explain our results. A potential concern is that our results depend on the performance of a baseline model in the evaluation measures. Specifically, a random baseline exposed to phonemic input generally performs worse than one exposed to syllabic input. This is because phonemic input is harder to segment due to the increased number of word boundaries. Such a discrepancy might introduce bias to our results, leading segmentation models to explain more variance in phonemic evaluation measures simply because the random phonemic baseline is outperformed by its syllabic counterpart. However, it is crucial to note that this performance imbalance between phonemic and syllabic baselines is adjusted for by computing the relative improvements of each segmentation model in comparison to its respective baseline. For instance, when comparing the performance of CLASSIC-UB with phonemic input to a random baseline with the same input, both models encounter the same number of word boundaries to identify, thus facing an identical challenge.

Another potential concern is our use of a specific baseline to compare the performance of the models (i.e., fully random baseline). If the baseline changes, the results might differ. However, we believe that the baselines from previous studies may lead to similar conclusions or may not be suitable for our research. For instance, adopting a baseline that recognizes every phonemic or syllabic unit as a word (e.g., Bernard et al., 2020) would inherently favour the syllabic baseline. This is due to the prevalence of monosyllabic words over monophonemic ones. Thus, using a unit baseline would not alter the conclusions from our current analysis.

On the other hand, if we were to treat every utterance as a single word — essentially discovering isolated words heard in one-word utterances (e.g., Fibla et al., 2022)— we would observe identical performance for both phonemic and syllabic inputs. This approach overlooks how segmentation task difficulty varies based on the perceptual unit used. Adopting this method would be equivalent to comparing the absolute performance of the segmentation models, without controlling for the fact that syllabic input presents fewer word boundaries to identify than phonemic input. Lastly, using pseudo-random baselines introduces its set of challenges due to underlying knowledge assumptions: It is unlikely that infants possess an

understanding of the true likelihood of a word boundary occurrence in the language (oracle baseline; Bernard et al., 2020). Likewise, they may not be aware of the true average word length across languages (Loukatou et al., 2019).

Apart from evidence that supports the plausibility of phonemic segmentation in capturing child vocabularies through the use of baselines to compute relative measures of performance, we also found that models exposed to syllabic input learned larger vocabularies than those found in child productions, despite receiving limited input. This evidence can be considered additional support for the access to phonemic information in building early vocabularies. However, we acknowledge that this result may be influenced by our use of idealized segmentation models: Although all models used in this study learned in an incremental fashion and therefore are different from fully idealized learners that learn from the input in batches (i.e., simultaneous processing of the entire input, e.g., Brent & Cartwright, 1996), they still learn from the input at every opportunity (e.g., the probability of storing a chunk in the lexicon is 1 in CLASSIC-UB and PUDDLE). Aspects of idealized learning have been used to compensate for the fact that models' input is typically very small compared to what children receive. However, in our case, facilitation from syllabic processing together with learning at every opportunity might be the reason for seeing word type vocabularies that are larger than children's (Table 4), and that contain a higher proportion of low-frequency words (Appendix S19). To confirm that sensitivity to sub-syllabic units provides a better fit to child data, in future work the models' learning could be constrained in different ways. For example, one could implement aspects of memory decay or limits on attention (e.g., Frank et al., 2010; French et al., 2011; Jones & Rowland, 2017; Perruchet & Vinter, 1998) that constrain the number of chunks that can be accessed at any given time. Still, it is unclear how a model that uniquely accesses syllabic units could capture the emergence of morphological chunks in Italian, as these require access to intra-syllabic information.

Our study implemented a rigid comparison between syllabic-only and phonemic-only representations to test the implications of these two extreme scenarios on vocabulary learning. However, future work could consider the

implications of the gradual discovery of phonemic units in infants (e.g., Werker, 2018). For example, different studies have proposed the use of a proto-lexicon of n-grams as a potential solution to the phoneme learning problem (e.g., Fourtassi & Dupoux, 2014; Martin et al., 2013), and the use of a proto-lexicon is in line with studies showing how infants initially represent frequent sequences of words and nonwords (e.g., Ngon et al., 2013). For example, Martin et al. (2013) proposed a top-down solution that could be implemented within chunking models considered in our study. Knowledge of n-grams (i.e., word-like chunks of the type learned by our models), can be used to perform successful identification (at word boundaries) of segment pairs that belong to the same phoneme category. The use of n-grams is useful because it allows the model to handle the large context-dependent realization of phonemes, which creates overlaps between different phoneme categories. Of course, the facilitation from a proto-lexicon is an aspect that aligns with information that our models can currently encode, but there are also other variables that could contribute to phonetic learning such as semantic information (e.g., Fourtassi & Dupoux, 2014; Yeung & Nazzi, 2014). Finally, note that modelling the gradual discovery of phonemes is difficult because it requires large amounts of fine-grained phonetic transcriptions of child-directed speech.

3.7 Conclusion

In this computational work, we examined how assuming different learning mechanisms (chunking, transitional probability) and access to different speech perception units (phonemes, syllables) could influence the early phases of Italian infants' word segmentation and word production. We found that (1) a chunking learning mechanism might play a significant role in early Italian word acquisition as found for English, (2) a chunking learning mechanism can capture a larger variability in children's vocabulary outcomes when it has access to phonemic information compared to syllabic information, (3) the saliency of words at utterance initial and final boundaries might aid both early word segmentation and word learning, (4) a chunking learning mechanism might help Italian infants discover morphological units alongside word forms in the early phases of word acquisition. These results

emphasize the cross-linguistic significance of chunking mechanisms for early word discovery and acquisition. They also underscore the need to consider model performance across different languages, allowing researchers to investigate the effects of a broader set of characteristics from naturalistic input and their interactions with the hypothesized learning mechanisms.

Chapter 4
The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution

4.1 Abstract

Lexical ambiguity is pervasive in the English language. Recent evidence suggests that children represent and learn multiple meanings of ambiguous words from early in development (e.g., "letter" as in mail or as part of the alphabet). However, the naturalistic cues that enable young children to resolve lexical ambiguities remain unclear. Previous research indicates that verbs might serve as a critical sentence cue that children rely on for disambiguation. Yet, it remains unclear whether such facilitation originates from bottom-up cues (verb-lexical associations) or top-down cues (verb-event structures). In other words, are children able to disambiguate "letter" in "She posted a letter" because the verb "to post" co-occurs more frequently in the context of mail than in the context of the alphabet, or because they have an understanding of the kind of entities that can function as arguments of the verb? In this study, we created and used ChiSense-12, a large sense-annotated corpus of English child-directed speech, to disentangle the effects of bottom-up verb-lexical and top-down verb-event structure cues in an experimental task. Our results show that four-year-old children relied on both types of cues, providing the first evidence that children can integrate sentence cues at multiple levels for disambiguating word meanings. We conclude by discussing how our findings carry significant implications for theoretical models of word processing and disambiguation.

4.2 Introduction

Lexical ambiguity refers to the fact that a word form can carry multiple meanings depending on sentence context (e.g., "she played in a band" vs. "she twisted a band"). Lexical ambiguity is thought to improve communicative efficiency (Piantadosi et al., 2012): Speakers reuse familiar word forms, reducing cognitive demands on

the language system, while the use of rich, informative sentence contexts allows them to still convey a wide range of meanings. However, lexical ambiguity can pose a challenge to children, as they might not be able to fully leverage sentence context for disambiguation (e.g., Khanna & Boland, 2010; Rabagliati et al., 2013). In this study, we aimed to identify which aspects of sentence context children can use to resolve lexical ambiguities.

Lexical ambiguity is extremely frequent in the English language (Rodd et al., 2022), which suggests that mastering lexical ambiguity resolution may be a crucial skill for infants and children. Recent investigations have shown that English child-directed speech is also lexically ambiguous (Meylan et al., 2021). Further, toddlers and pre-schoolers master lexical ambiguity in comprehension (Floyd et al., 2020) and production (Meylan et al., 2021), employing a diversity of meanings for most words just as adults do. However, less is known about which aspects of sentence context might facilitate children's processing of ambiguous words. Previous studies have shown that children can learn different senses of ambiguous words from word-level semantic cues (e.g., children can more easily associate a word form with two object referents if the objects are similar in shape; Floyd & Goldberg, 2021; Srinivasan et al., 2019) and syntactic category cues (e.g., whether the ambiguous sense functions as a noun or verb; Dautriche et al., 2018; Lippeveld & Oshima-Takane, 2020), but few have examined which sentence context cues could facilitate children's processing of ambiguous words (Khanna & Boland, 2010; Rabagliati et al., 2013). These few studies have shown that, compared to adults, children aged 4 to 7 years have difficulty using the top-down global plausibility of sentences to disambiguate familiar word senses. In other words, children appear to struggle with lexical disambiguation based on one's real-world knowledge, which facilitates comprehension of causal relations, event sequences, and social norms conveyed by the overall discourse. In contrast, children seem to mostly rely on bottom-up word associations (i.e., tracking co-occurrences between words) to perform a shallow processing of sentence context when interpreting ambiguous words. Nevertheless, the variance in children's disambiguation performance can only be partially explained by word co-occurrences (Rabagliati et al., 2013), indicating that other top-down cues might play a role, but exactly which cues children rely on is still unknown.

In this study, we aimed to address this gap by examining whether the semantic restrictions that verbs impose on their arguments (i.e., verb-event structures) might represent a top-down sentence context cue that children could rely on from a young age. We hypothesized that these verb-event structures may play a significant role in children's lexical ambiguity resolution, given their role in children's unambiguous word processing (Andreu et al., 2013; Mani et al., 2016) and their reliance on verbs in tasks that require sentence parsing, namely syntactic ambiguity resolution (Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021).

Verbs are an important source of disambiguating information for ambiguous nouns (Hahn et al., 2015; Rabagliati et al., 2013); for example, after hearing the phrase "eat the chicken", a child is likely to interpret the noun "chicken" as referring to a type of food rather than livestock. But while we know that verbs facilitate children's interpretation of ambiguous words, it is still unclear whether such facilitation operates in a top-down or a bottom-up manner, because bottom-up and top-down cues are often entangled in naturalistic contexts (e.g., Ambridge et al., 2015). In the example above, the verb "eat" might prime the target meaning "chicken [food]" via lexical association (i.e., working as a bottom-up cue to ambiguity resolution); alternatively, the semantic restrictions imposed by the verb "eat" on its arguments (verb-event structure) might guide top-down inferences to suppress contextually irrelevant meanings (e.g., upon hearing "eat the chicken", the child may infer that "chicken" refers to a type of food because inanimate entities are more plausibly eaten than animate entities).

In this work, we created and leveraged a large sense-annotated corpus of child-directed speech, ChiSense-12, to carefully construct experimental stimuli and disentangle the effect of bottom-up and top-down verb-related cues in early lexical ambiguity resolution. Understanding the role of different types of cues is important for theories of early sentence parsing, some of which emphasize children's reliance on bottom-up cues (Snedeker & Yuan, 2008), while others propose that children consider both bottom-up and top-down cues from early on (Trueswell & Gleitman, 2007). Furthermore, it is key to understanding the learning mechanisms that might underlie sensitivity to different cues in language development. For example, usage-

based computational models of verb-event learning assume that children track word associations to learn how words are combined into sentences. Simultaneously, they use analogy mechanisms to compare sentences with similar associations and abstract verb-event constructions (e.g., Alishahi & Stevenson, 2007, 2008). This implies that children should be able to use their verb-event knowledge (alongside verb-lexical associations) in sentence parsing from early stages of development. Therefore, evidence from children's processing would bolster the proof of principle evidence from computational studies, such as Alishahi and Stevenson (2007).

In the following sections, we introduce studies that have examined the role of bottom-up and top-down cues in adults and children's lexical disambiguation. In our study, we also tested adults alongside children to serve as a comparative baseline and enable us to gain insights into the process of lexical disambiguation across different age groups. We then present a section that describes how we sense-tagged a large corpus of English child-directed speech and how we used it to construct experimental stimuli for our adults and children's study. In the remainder of the paper, we present and discuss the results of this experimental study.

4.3 The Role of Context in Lexical Disambiguation

Theories of lexical processing have faced the challenge of explaining which linguistic aspects allow the individual to access a word's meaning. Influential models of lexical processing made the simplifying assumption that a word form maps to a single correspondent meaning (e.g., Plaut et al., 1996; Seidenberg & McClelland, 1989). Yet, evidence from both adults and children indicates that they map word forms to multiple meanings. For example, adults' recognition of familiar words is slowed down when the target word maps onto different semantically unrelated senses (e.g., *dog/tree* bark), likely because of competition between representations of alternative meanings (Rodd et al., 2002). In a similar way, 4-year-old children can use sentence context to shift their interpretation of an ambiguous word from one sense to an alternative, but they still make more mistakes in choosing the correct meaning of an ambiguous word compared to an unambiguous one (Rabagliati et al., 2013). These

findings in adults and children can be accounted for by more recent models of lexical processing (Duffy et al., 2001; Rodd, 2020).

Among recent models of lexical processing, there is a shared focus on the importance of context in lexical representation. For example, the semantic settling model (Rodd, 2020) assumes that the individual represents word meanings in a high-dimensional lexical-semantic space. In other words, multiple features can define the link between a word form and its meaning, and these features relate to properties of the word itself or contextual aspects that are present when the word is used. The multiple meanings of a word are then defined by alternative paths in the lexical-semantic space, each path representing the set of features defining the mapping between a word form and each sense. Access to ambiguous word meanings is seen as a settling process, by which certain paths in the space become increasingly activated and settle toward a configuration correspondent to one of the alternative meanings. Activation is influenced by multiple cues that help the system settle on one meaning, including bottom-up information about a meaning's usage patterns (e.g., meaning expectation based on frequently co-occurring words in sentence context), or top-down information regarding the meaning of a word that can be inferred given the context (e.g., real-world knowledge used for pragmatic inferences).

The possibility of integrating a wide range of cues for word-meaning access implies that the individual possesses lexical-semantic representations that are rich and context-dependent (e.g., Elman, 2009). The same view is shared with a recent account of children's word learning in Srinivasan and Rabagliati (2021), proposing that representations of word senses are conditioned on contextual aspects (e.g., a/some chicken; thirsty/roasted chicken). Indeed, evidence exists that contextual cues can work as an aid to word sense learning: For example, when a word typically used as a verb (e.g., "eat") is presented as a noun (e.g., "an eat"), infants find it easier to associate the word with a novel animal. In contrast, when a word is strongly associated as a noun with a specific referent (e.g., "dog"), infants find it difficult to extend its use to label another novel animal (Dautriche et al., 2018). This

suggests that the use of different syntactic categories is necessary to facilitate the expansion of a word's meaning to include additional referents.

Aside from word learning, evidence supporting the idea of contextualized representations in the processing of ambiguous words comes from studies involving both adults (e.g., Colbert-Getz & Cook, 2013; Witzel & Forster, 2014) and children (e.g., Hahn et al., 2015; Khanna & Boland, 2010; Rabagliati et al., 2013). However, the question of how children's representations differ from adults' remains open. Particularly, it is unclear if children's representations can integrate both bottom-up and top-down cues to word meaning and how these cues are weighted in sentence parsing. We delve into these points in the following sections.

4.3.1 Word- and Sentence-Level Influences on Disambiguation

Adults and children's disambiguation is influenced by cues at the word and sentence level. Regarding cues at the word level, both adults (Duffy et al., 1988; Rodd et al., 2016) and children (Booth et al., 2006; Simpson & Foster, 1986) show sensitivity to meaning dominance, namely a bias toward the ambiguous word meaning that is most frequent in the language. In the adult literature, a bias has been documented that delays the interpretation of a subordinate (less frequent) meaning, even when the sentence context aligns with this subordinate meaning (the so-called subordinate-bias effect). An example of this can be seen in Duffy et al.'s (1988) study, where participants activated the dominant meaning of "ball" as an object even though the sentence context was clearly biased towards "ball" as in a dance gala (e.g., "Although attendance was not required, the ball was very important"). Similarly, when a prime word is semantically related to the dominant meaning of a target ambiguous word, children at different ages process the target faster or more accurately compared to when the prime word is related to the subordinate (9, 10, and 12 years old, Booth et al., 2006; 4 years old, Rabagliati et al., 2013; 8, 10, and 12 years old, Simpson & Foster, 1986).

Other sources of disambiguation operate at the sentence level, working in concert and in certain cases mitigating the effect of word level influences. In adults,

a strongly supportive sentence context can, in some cases, eliminate the subordinate-bias effect (Colbert-Getz & Cook, 2013): For example, when a preceding context refers to "being around water", reading time delays occur in accessing the river-related sense of "bank" because the dominant sense "bank [institution]" interferes with lexical access. However, including multiple references to the subordinate sense in the preceding context ("catching a fish", "going to the river", "being in the mud and around water") eliminates the interference and allows the individual to access the subordinate sense as quickly as the dominant in a control condition. This evidence indicates that there is a cumulative effect of sentence context and, more importantly, it shows that sentence context is an important element for disambiguation as it interacts with sense frequency to influence word-meaning access. A similar effect has been found in children by Rabagliati et al. (2013): They presented 4-year-olds with contexts that were highly constraining towards the subordinate senses of ambiguous words (e.g., "Kermit was walking in a dark cave. He was nervous about the animals, because he saw a big bat"). Children were as accurate at choosing the subordinate *animal* bat as they were at choosing an unambiguous target ("…because he saw a blackbird"), indicating no interference caused by the activation of the dominant sense *object* bat.

The above findings indicate that both adults and children leverage sentence context in lexical disambiguation. However, they do not tell us whether adults and children process sentence context in the same way. This question is important for developmental theories that assume that young children rely solely on bottom-up information (word associations) when parsing speech (bottom-up account; Snedeker & Yuan, 2008), and for those that allow additional integration of top-down information (e.g., syntactic or semantic structures) when this is perceived as sufficiently reliable (cue-validity account; Trueswell & Gleitman, 2007). To shed light on this problem, studies have used contrastive tasks in which bottom-up and top-down cues are embedded in text or spoken stories and they compete, pointing to opposite senses of ambiguous target words (Kambe et al., 2001; Khanna & Boland, 2010; Rabagliati et al., 2013; Witzel & Forster, 2014). In these contrastive tasks, adults rely on top-down cues related to global semantic plausibility. Conversely, children appear more prone to interpretation errors, suggesting they might not be

fully integrating top-down cues. We discuss adults and children's performance in contrastive tasks in the next section.

4.3.2 Use of Bottom-Up and Top-Down Cues in Adults and Children

When presented with contrastive stories, adults rely on top-down cues related to the story global semantic plausibility (e.g., Kambe et al., 2001, Rabagliati et al., 2013). For example, in Rabagliati et al. (2013), when hearing the short story "Kermit was walking in a dark cave. He was nervous about the animals, so he carried a big bat", adults assigned an object interpretation to "bat" (e.g., *baseball* bat), even if the story contained words that frequently co-occur with the sense *animal* bat in the language (i.e., "dark cave", "animals"). In other words, adults could use their event knowledge to infer that, given his emotional state, Kermit probably carried an object to protect himself. Note that this evidence does not mean that bottom-up associations have no effect on adult processing. In fact, in another study by Witzel and Forster (2014), word associations did cause some delay in the online processing of an ambiguous word when they conflicted with the overall context of the sentence. Moreover, in another study adults showed sensitivity to word associations when all cues in the sentence context fully supported one sense (i.e., non-contrastive task; e.g., Khanna & Boland, 2010), with an additive effect of bottom-up associations and top-down plausibility.

In adults, reliance on top-down plausibility is useful, as solely relying on word associations could lead the individual to commit interpretation errors in cases where word associations are not particularly strong or when the alternative senses of an ambiguous word are semantically related and might share similar context word associations (e.g., *food/animal* chicken). But differently from adults, developmental studies on lexical ambiguity have shown that 4- to 7-year-old children do commit interpretation errors, failing or only partially integrating top-down global plausibility (Khanna & Boland, 2010; Rabagliati et al., 2013). For example, consider the homophones "guest" and "guessed" (/gɛst/) in Khanna and Boland (2010). If children generate top-down inferences based on sentence context, when they hear

"The house is clean because we expect a /gɛst/", they should activate the congruent meaning "guest", but not the incongruent "guessed". Thus, they should subsequently find it easier to repeat the word "room" (which is a frequent lexical associate of "guest"), compared to hearing a context that is only compatible with the alternative meaning (e.g., Molly didn't know the answer, so she /gɛst/). However, differently from adults, 7-year-olds showed the same facilitation from both "The house is clean because we expect a /gɛst/" and "Molly didn't know the answer, so she /gɛst/" (compared to a completely unrelated sentence), suggesting that, although children were sensitive to the lexical association between /gɛst/ and "room", they were not able to integrate top-down information from the sentence context.

This evidence might suggest that children do not integrate top-down information (bottom-up account; Snedeker & Yuan, 2008). However, evidence from other studies suggest that there might be conditions in which children can leverage top-down cues (Hahn et al., 2015; Rabagliati et al., 2013). In the experiment from Rabagliati et al. (2013), 4-year-old children successfully disambiguated words based on a preceding sentence context (e.g., "Barney was on vacation. He fed/roasted the chicken, which was nice"). To test if successful disambiguation depended on the association between the preceding verb (e.g., to feed) and the ambiguous word sense (*animal* chicken), the authors ran a computational model that performed the disambiguation task uniquely leveraging the target sense frequency (dominance) and the statistical co-occurrence between context words (including the verb) and target sense in child-directed speech (bottom-up associations). Although the authors found a significant correlation between the model and children's performance, this was moderate (r = .32), indicating that children might have relied on additional top-down cues to resolve ambiguities. To examine this hypothesis, they tested children on a second contrastive task where bottom-up associations were put in competition with top-down global plausibility. They found that children could make partial use of top-down plausibility. For example, upon hearing the sentence "Elmo watched a funny movie about a castle, and a princess, and a silly dragon. That was a funny /nȧɪt/", they selected a picture depicting "night" (rather than one of a "knight") more often than when the sentence ended in "And there was a funny /naɪt/". Even if

words like "castle", "princess", "dragon", and "knight" frequently co-occur in naturalistic speech, children were able to use their top-down event knowledge and infer that people usually watch movies at "night". Therefore, children were sensitive to top-down cues. Importantly, however, in Rabagliati et al. (2013) children still relied more on lexical associations than on global plausibility: Even if a (residual) significant difference was found between the above conditions, children still selected "knight" more than 50% of the time in every condition.

The above evidence suggests that, to some degree, 4-year-old children can understand the ongoing discourse by leveraging top-down global plausibility to resolve lexical ambiguities (Elman, 2009). However, it is still unclear whether there are conditions in which children could primarily rely on top-down cues, which would strongly support a cue-validity account (Trueswell & Gleitman, 2007).

In this study, we directly compared children's reliance on bottom-up vs. top-down verb-related cues. We chose verbs because they are likely to represent a particularly valid cue that young children can rely on when processing ambiguous words. For example, the type of syntactic arguments that verbs take guide children's interpretation of ambiguous sentences (e.g., Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021). To illustrate, 3- to 5-year-old children interpret the phrase "tickle the bear with the mirror" as "tickle the bear using the mirror" (even if two bears are shown, one of which is holding a mirror) because the verb "tickle" frequently co-occurs with instrument arguments in naturalistic speech (Yacovone et al., 2021). Further, verb-event structures guide children's unambiguous word processing (Andreu et al., 2013; Mani et al., 2016). For example, 3-year-olds know that "pushing a flowerpot "is more plausible than "pushing a road" even if they have never heard either in conversation (Andreu et al., 2013).

Although some studies have investigated the role of verbs in early lexical ambiguity resolution (Hahn et al., 2015; Rabagliati et al., 2013), they have not examined the independent contribution of (bottom-up) verb-sense associations and (top-down) verb-event structure cues. This is because stimuli used in previous studies included verbs that were both lexically associated with a target sense and licensed the target sense as a plausible argument: in "Karl met the star", the verb

"meet" is likely to co-occur more often with "star [famous person]" than "star [astronomical object]" in the language, and at the same time one more plausibly "meets" an animate entity than an inanimate one (Hahn et al., 2015). Moreover, in the experiments from Rabagliati et al. (2013), although children could successfully use verbs to disambiguate a subsequent target word in a non-contrastive task, in their second experiment the interpretation of contrastive passages did not always depend on verbs (e.g., "that/there was a funny *(k)night"*, or "the teacher played music with anyone/anything, even a *band*). Thus, the specific role played by verbs in lexical disambiguation remains to be studied.

In our study, we used a large sense-annotated corpus of child-directed speech to design experimental materials which could disentangle the contribution of verb-noun lexical associations and top-down verb-event structure cues. The next section elaborates on the construction and the characteristics of the corpus.

4.4 Annotating Child-Directed Speech for Word Senses: The ChiSense-12 Corpus

Language acquisition research has benefited from the use of annotated corpora of child-directed speech to examine key questions about how children learn and process language in real-world contexts (e.g., Monaghan & Rowland, 2017). Naturalistic corpora are useful in different ways. They can be used to analyse language use patterns, test the plausibility of different learning mechanisms by applying them to naturalistic speech through computational modelling, or aid in building experimental stimuli to test the role of variables found in naturalistic conversations. However, corpora currently available such as in the CHILDES database (MacWhinney, 2000) do not provide information about different senses that words can assume depending on the conversational context. The lack of sense-annotated child-directed input makes it difficult to examine questions about child lexical ambiguity via methods that use naturalistic corpora. To address this limitation, we constructed the first large-scale child-directed speech corpus tagged for word senses. We named this new sense-tagged corpus ChiSense-12 (freely

available at https://gitlab.com/francescocabiddu/chisense-12), and we used it to carefully balance experimental materials that could disentangle the effect of verb-lexical association and verb-event structure in children's lexical ambiguity resolution.

Numerous sense-annotated corpora based on adult language exist (for an overview, see Pasini and Camacho-Collados, 2020). In these, sense annotation is usually based on Wikipedia pages or the sense inventory WordNet (Miller, 1995). The use of these sense-annotated corpora has proven to be useful in capturing adults' word sense disambiguation. For example, Loureiro et al. (2021) have recently shown that computational models based on the Transformer neural architecture (Vaswani et al., 2017) more closely approximate adult sense inter-annotator agreement when trained on sense-annotated instances compared to uniquely exploiting glosses from sense inventories. Although this evidence highlights the potential of using adult corpora for adult word sense disambiguation, the same is not necessarily true for studying child competence.

Compared to adult language, speech that young children hear is more repetitive (e.g., Jones et al., 2023), restricted to certain topics and concrete vocabulary (e.g., food, clothing, animals), with shorter sentences and simpler syntactic structure (Saxton, 2009). These characteristics may play a key role in early word processing and learning (e.g., Weisleder & Fernald, 2013), indicating that experimental or computational investigations aimed at capturing children's language understanding should be based on the specific input they receive. Furthermore, sets of ambiguous words tagged in adult corpora may consider senses that are not understood by children, or conversely, they may omit senses that are understood by children. This makes it important to select samples of word senses that young children understand.

In the first work addressing these challenges, Meylan et al. (2021) are currently tagging two large corpora of English child-directed speech (and corresponding child productions) from the CHILDES database. The child-directed corpora comprise speech directed to 18 children of age between 9 and 51 months. A total of 112,802 word tokens is being tagged using WordNet sense inventory as a reference. The sample of word types considered are based on a common measure of

child vocabulary from parental report, the Communicative Development Index (CDI; Fenson et al., 2007), covering a total of 719 lemma+part-of-speech combinations in the corpus.

Although Meylan et al.'s dataset will significantly contribute to the naturalistic study of lexical ambiguity in early childhood, it is less useful for examining the contribution of specific aspects of sentence context such as verb-event structure. First, tagging specific syntactic patterns (e.g., verb-object) was not the focus of the project. Secondly, high-frequency words in the dataset are downsampled (i.e., a random sample of 50 tokens in each 3-month recording interval is tagged) to minimize annotation time. Although this seems a reasonable strategy when focusing on word sense distributions for each word type, it limits the researcher's ability to look at the distribution of verbs that co-occur with each specific sense (i.e., the verb distribution becomes especially downsampled for senses that appear infrequently in the corpus). For this reason, we used a large English corpus of child-directed speech where we manually tagged the full sample of tokens for both word sense and verbs that take a sense as an object. Given the large-scale nature of the project, to make the annotation task manageable we only coded a pre-selected sample of 12 ambiguous words. We describe the corpus, the word sample, and our annotation strategy below.

4.4.1 Corpus

We downloaded all American and British English corpora from the CHILDES database (version 2020.1) using the R package *childesr* (Braginsky et al., 2019), which provides utterances in orthographic form using a standardized procedure to treat special codes across corpora (e.g., prosodic, discourse markers). Out of 72 corpora downloaded, we considered 53 involving target children of up to 4 years of age (59 months), resulting in speech directed to 958 target children. We further filtered the dataset for utterances containing 12 ambiguous words (see Table 5). For each word, a frequent dominant sense and a less frequent subordinate sense were considered (e.g., Bat: *dominant* = animal, *subordinate* = object). 11 words were selected from

a previous study where 4-year-olds showed understanding of both dominant and subordinate senses (Rabagliati et al., 2013). An additional word was selected with both senses having a relatively high frequency in child-directed speech (/flaʊə/: flower/flour), with its dominant meaning being known by children from around 20 months of age (Frank et al., 2017).

4.4.2 Annotation

The dataset was tagged by the first author. We only considered utterances where a target word was used in its dominant or subordinate sense. Each utterance was tagged for the word sense used (dominant/subordinate). For utterances where the sense was used as object argument, we reported the verb stem preceding the sense (see Table 5). For utterances where the word's sense was not immediately understandable, the surrounding conversational context was considered (i.e., surrounding utterances in the transcripts; see Figure 17). If the conversational context did not allow the annotator to understand the intended meaning, the utterance was discarded.

Table 5 Example of coded utterances. ID is the CHILDES database utterance number. This identifier can be used to retrieve specific corpus variables including speakers and target children's information. The remaining columns contain the target utterance (GLOSS), target ambiguous word (TARGET), specific word sense (SENSE) and verb stem used with that sense (VERB).

ID	Gloss	Target	Sense	Verb
311504	who put the rubber band on there	band	object	put on
326153	are you in a marching band	band	music group	be in
326190	oh a clown's in the band	band	music group	be in
326293	remember Child when did we see a band	band	music group	see

The final dataset included 15,581 utterances out of an initial raw sample of 21,342 (*word tokens* = 115,272; *word types* = 4,805). The dominant sense

appeared on average 73% of the time (*SD* = 13%). Descriptive statistics for each ambiguous word are presented in Table 6.

Table 6 For each target word, the table shows the raw number of utterances in which dominant and subordinate meanings appeared (N), and percentage of utterances in which dominant sense appeared (Dominance).

Word **(Dominant/Subordinate)**	***N*** **(Dominant/Subordinate)**	***Dominance***
Band (Object/Music Group)	178/58	75%
Bat (Animal/Object)	247/130	66%
Bow (Knot/Weapon)	230/27	89%
Button (Electronic/Clothing)	568/285	67%
Chicken (Animal/Food)	1463/937	61%
Flower/Flour	3521/350	91%
Glasses (Eye/Drinking)	683/620	52%
Letter (Alphabet/Mail)	1446/946	60%
Line (Geometric/Row)	471/241	66%
Moose/Mousse	178/42	81%
Nail (Finger/Tool)	460/106	81%
Sun/Son	2029/365	85%
MEAN (*SD*)	-	73% (13%)

As 8 of the 12 words included in Chi-Sense 12 are also in Meylan et al.'s (2021) dataset, we plan to analyse inter-annotator agreement as soon as this large-scale dataset is released. To give an idea of the difficulty of the annotation task, we conducted a small inter-annotator agreement study, generating a random list of 45 sentences from the coded corpus (5 per target word, excluding target words that are not homographs, i.e., moose/mousse, flower/flour, sun/son). After a short training

(using 5 training conversations), a second annotator read 45 test conversations between a child and one or more adults (see Figure 17). For each conversation, the second annotator was asked to indicate whether a target ambiguous word highlighted in red referred to its dominant meaning, subordinate meaning or to something else.

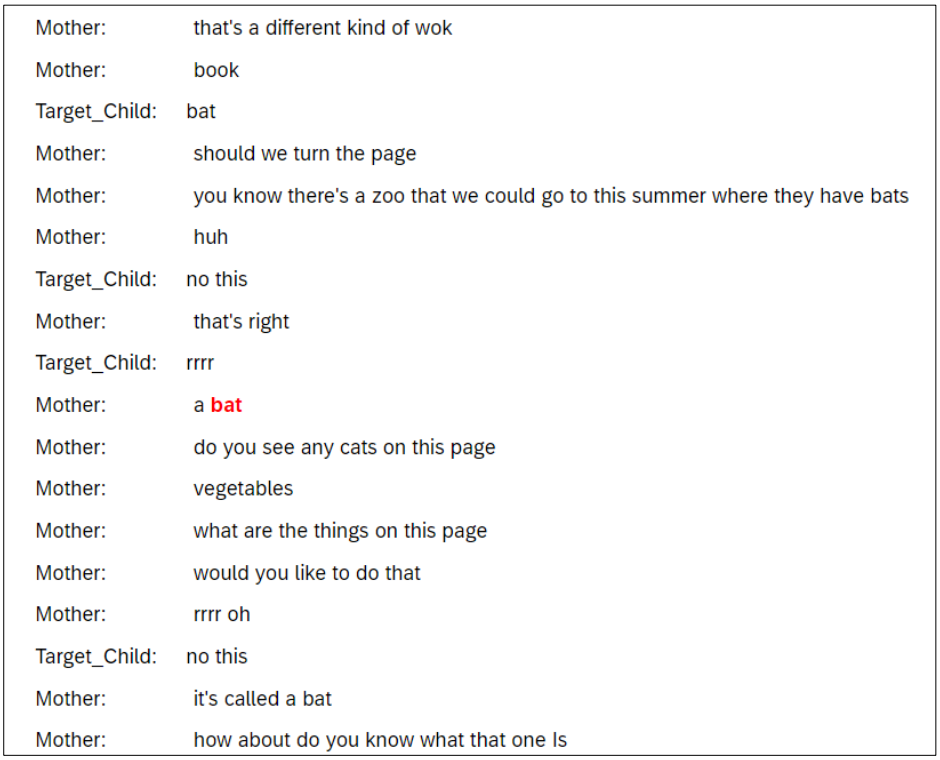


Figure 17 Example of test conversation in the small inter-annotator study. The target word in red is surrounded by its conversational context.

We found 100% agreement between first and second annotator (*Kappa* = 1). The scripts for generating the random list of sentences and the small study results can be found in the annotation project GitLab page https://gitlab.com/francescocabiddu/chisense-12.

In the following section, we describe how we designed an experimental task and used ChiSense-12 to create experimental conditions and stimuli that could

answer our research questions on the role of verb information in early lexical ambiguity resolution.

4.5 Design of Experimental Task

We designed a web-based forced-choice task similar to Rabagliati et al. (2013), to examine the role of verb-noun lexical associations and top-down verb-event structure cues in adults and children's lexical disambiguation.

Participants heard spoken stories that ended with a target ambiguous noun (see Figure 18). Two seconds before story onset, four pictures appeared on the screen and stayed on until a picture was selected. After hearing the story, participants were asked to select a picture that corresponded to the last word of the story. In each trial, two pictures depicted the two senses of a target ambiguous word (the frequent dominant meaning and the subordinate less frequent meaning) (e.g., *object* band, *music* band). The other 2 pictures depicted distractor words semantically related to these senses, which were also good completions of experimental stories. Distractors were also frequency-matched to target senses based on the sense-annotated corpus statistics. More details about distractors and target sense frequencies in the corpora and their frequency matching can be found in Appendix S24. Participants also initially saw 3 training trials, with spoken stories ending with unambiguous target words (e.g., "Emily went to the shop. Then, she bought a banana").

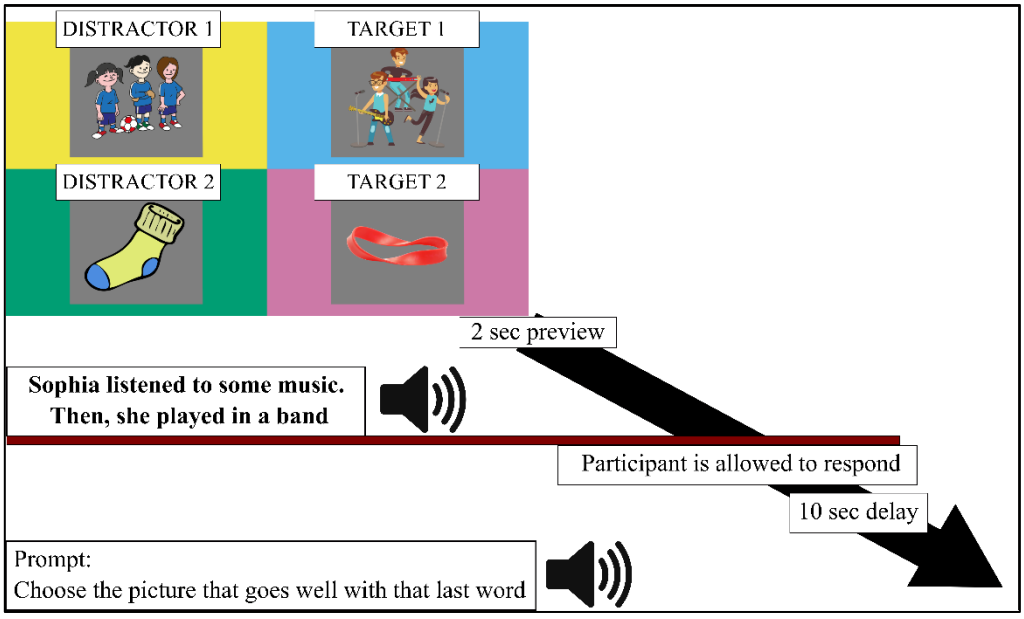


Figure 18 Example of trial. Participants saw a 2x2 grid depicting 2 target word senses (dominant: *object* band; subordinate: *music* band), and two distractor words (sock, team). Pictures appeared in random locations on every trial. After 2 seconds from picture presentation, the spoken story was played. Participants were allowed to respond only after the story ended.

Following Rabagliati et al. (2013), we constructed the experimental stories in a way that would allow us to examine whether children use top-down event structure cues when these are put in competition with bottom-up cues (i.e., to exclude the possibility that children use top-down cues only when these are the only ones available in context). Therefore, we constructed stories comprising a prior context and a target context. The prior context always contained words that frequently co-occurred with the target subordinate sense in child-directed speech. For example, in Figure 19, the prior context "Sophia listened to some music" contains the words "listen" and "music" which frequently co-occur with the subordinate meaning *music* band.

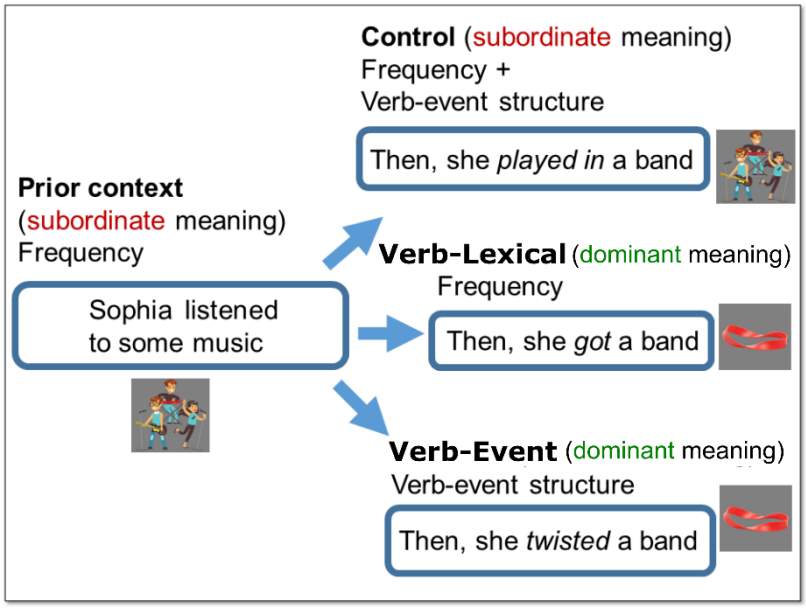


Figure 19 Example of conditions involving the target word "band". Participants could see a prior context either followed by a control, verb-lexical, or verb-event context.

The target context was manipulated in 3 within-subject conditions. In the control condition, the main verb pointed toward the target subordinate sense (i.e., the same sense that was favoured by the prior context), both in terms of lexical associations in child-directed speech and plausibility based on verb-event structure. For example, in "Then, she played in a band" (see Figure 19), the verb "play in" is lexically associated with *music* band in child-directed speech and one more plausibly plays in a *music* band than an *elastic* band. Specifically, we defined verb-sense lexical association by weighing the raw frequency of verb-sense occurrence by the number of times the sense appeared in the corpus as an object of a verb (see Appendix S26).

In the verb-lexical condition, the main verb was lexically associated to the dominant target sense ("get" frequently co-occurs with *object* band in child-directed speech; see Figure 19), therefore competing with bottom-up cues from the prior context (which pointed toward the subordinate target sense *music* band"). Importantly, verb-event structure information was compatible with both target senses in this condition (i.e., one can either "get a band [object]" or "band [music

group]"). Conversely, in the verb-event condition there was no lexical association between the main verb and either the dominant or subordinate sense. However, the verb only accepted the dominant sense as a plausible object (i.e., one can only "twist a band [object]").

Given the competition between cues from the prior and target context, in the verb-lexical and verb-event conditions one must make a higher number of inferences to link the two contexts (e.g., "Sophia listened to some music. Then, she twisted a band") than in the control condition. Therefore, with the intent of weakening the link between contexts in the control condition as much as possible, we lowered the coherence of all stories. We used a temporal connective ("Then") which is considered the lowest level of conceptual coherence save for completely unrelated sentences (see Connell & Keane, 2004; compare the control story "Sophia listened to some music. Then, she played in a band" to the alternative "Sophia wanted to create music, so she played in a band").

4.5.1 Study Hypotheses

In this section, we outline the study hypotheses. Given the prominent role of lexical association in lexical ambiguity resolution (Khanna & Boland, 2010; Rabagliati et al., 2013) and of verb bias in syntactic ambiguity resolution (e.g., Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021), we expected verbs to facilitate children's performance when the unique cue available is verb lexical association, but it is an open question whether children would be sensitive to this bottom-up cue when verb plausibility does not help.

Further, given the role of verb-event structure in early unambiguous word processing (Andreu et al., 2013; Mani et al., 2016), we would expect a strong effect of this top-down cue, but empirical evidence is needed to establish whether this would be the same for ambiguous word processing.

4.6 Method

4.6.1 Participants

All participants resided in the United Kingdom. We recruited 83 adults from the platform Prolific (age: *M* = 23 years, *SD* = 4.5 years; 55 females). Data from one adult were discarded for failing more than 1 out of 3 training trials. We used the adult dataset to carry out a series of power simulations, estimating the ideal sample size for child data collection. In doing so, we considered a range of true effect sizes, ensuring we achieved sufficient power even at smaller effect sizes than those expected from previous studies but still of practical significance. We also considered how power was affected by participant and item variation in interaction with sample size and effect size. Before collecting child data, we defined and pre-registered a data collection stopping rule in OSF at https://osf.io/a293m/?view_only==73b7fdb649ef42e0ab943d198b788c5c.This repository also contains a comprehensive report of the power analysis. The R scripts necessary to replicate the simulations are also available at https://osf.io/k2xmv/?view_only=5dbf0cb8e26f4b6e854eee28d93869e1.

To summarize the criteria established by the stopping rule: Once the child sample reached *N* = 42 (the minimum adult sample for which power was simulated), different statistical models could be fitted to the child data. The models had differing complexities of random effects structure, with the final model selected being the most complex of those that satisfied all the criteria of the stopping rule. To stop data collection, a model (starting from the one with the most complex random structure) needed to demonstrate statistical convergence. Moreover, the stopping rule required that the standard deviation estimates of the random effect parameters of the child model (participant per condition slopes for more complex models; participant and item intercepts for the intercept-only model) should not exceed a specific threshold (defined by taking adults' standard deviations as the reference) to ensure adequate and stable power. For instance, the random effect standard deviations in the most complex model could not exceed twice the corresponding standard deviations in the adult model at *N* = 42. Crucially, if convergence and/or sufficient power were not achieved for any model, additional data were collected by adding one participant for

every counterbalancing experimental block (to ensure balanced control, verb-lexical, and verb-event data points for each target word). Effectively, the sample size was increased by *N* = 3 at every step.

The criteria of the stopping rule were met upon reaching the final sample of *N* = 45. In addition to these 45 children, data from another 10 children had to be discarded during the course of data collection for various reasons (3 due to experimental errors, 1 due to fussiness, 5 due to failed training, and 1 due to language impairment). The final child sample included English-speaking children aged between 48 to 59 months (age: *M* = 52 months, *SD* = 3 months; *gender*: 8 female, 9 male, 21 non-binary / third gender, 7 prefer not to say). The complete distributions of the child socio-demographic variables are reported in Appendix S25.

This research project was approved by the ethics committee of the School of Psychology of Cardiff University (EC.18.05.08.5295GR).

4.6.2 Materials

We created experimental stories for the 12 ambiguous target words included in the ChiSense-12 corpus. The corpus was also tagged for verb stems that take ambiguous senses as object arguments. This allowed us to construct the experimental stories by computing relative frequencies of co-occurrence between verbs and target senses. Experimental stories with corresponding verb-target frequencies are shown in Appendix S26.

To ensure that all senses in the study were known by children, we asked caregivers to fill in a questionnaire where they could indicate whether a target sense or context verb was "not understood", "understood", or "understood and used" by children. Parents responses on sense and verb knowledge across children are shown in Appendix S27. We excluded 24% of trials (129/540) for which a caregiver indicated the child did not know the context verb or at least one of the two target senses (although note that we obtained the same results when including the full sample of trials, see Appendix S28).

We also asked adults to name each picture used in the experiment. Given that we matched target and distractor pictures by frequency, we ensured that adults spontaneously named the distractors (not spoken in the stories) using the labels we used for the frequency matching (e.g., when matching "chicken" with the distractor "crow", we ideally want the latter image to be named as "crow" by participants and not as "bird"). For every distractor, the expected label was always the most frequently reported, and was used by 89% adults on average (*SD* =15%).

4.6.3 Procedure

Adults completed the task independently online. Children's online task was identical to the one completed by adults, but an experimenter supervised the sessions because children were asked to give verbal responses (i.e., to say the colour of the target picture background, see Figure 18). The presence of the experimenter was also to ensure that caregivers would not interfere in child responses and that children would stay engaged in every trial.

Each participant in the experiment saw 4 control stories, 4 verb-lexical stories and 4 verb-event stories (all in randomized order), and assignment of stories to conditions was counterbalanced across participants (see counterbalancing blocks in Appendix S26).

4.6.4 Statistical Analyses

R scripts to reproduce all the figures, tables, and statistical results of this paper are available at https://osf.io/k2xmv/?view_only=5dbf0cb8e26f4b6e854eee28d93869e1.

We first conducted our analyses on adults. Before collecting any child data, we pre-registered key aspects of the study design and our two main hypotheses at https://doi.org/10.17605/OSF.IO/FK378. The analyses that examined the two main hypotheses of the study contain no deviations from what is indicated in the pre-registration documents. Additionally, we present a set of exploratory analyses not included in the pre-registration. These additional analyses were triggered by

comments received by anonymous reviewers, and we considered them valuable to shed light on the variables that might have determined adult and child performance at the experimental task.

For the pre-registered analyses, we fitted mixed-effect logistic regression models separately to adults and children's data. We used sense choice (dominant, subordinate) as the dependent variable, and condition as the independent variable (control, verb-lexical, verb-event). We specified two contrasts: control vs verb-lexical, control vs verb-event. For the adult model, the random effect structure of the models included random intercepts for participant and item, and random slopes of condition per participant and item (excluding estimated correlations between item intercepts and slopes). For the child model, we included random intercepts of participant and item, and random slopes of participant per condition. These random effect structures were the ones that allowed the models to converge and for which our simulations indicated sufficient and stable power to detect effect sizes of interest (see simulation scripts on OSF).

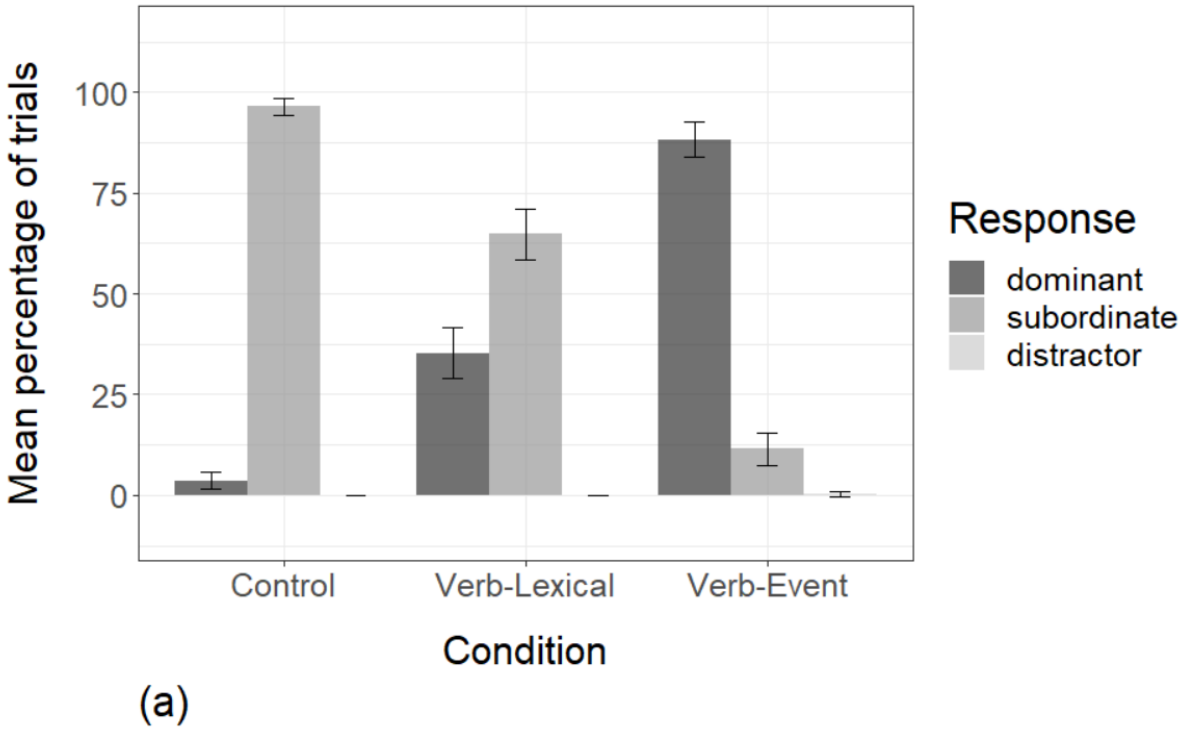
4.7 Results

In this section, we report the results related to the two main hypotheses of the study. We report key results in text and include the full output of every statistical model in Appendix S28-S30.

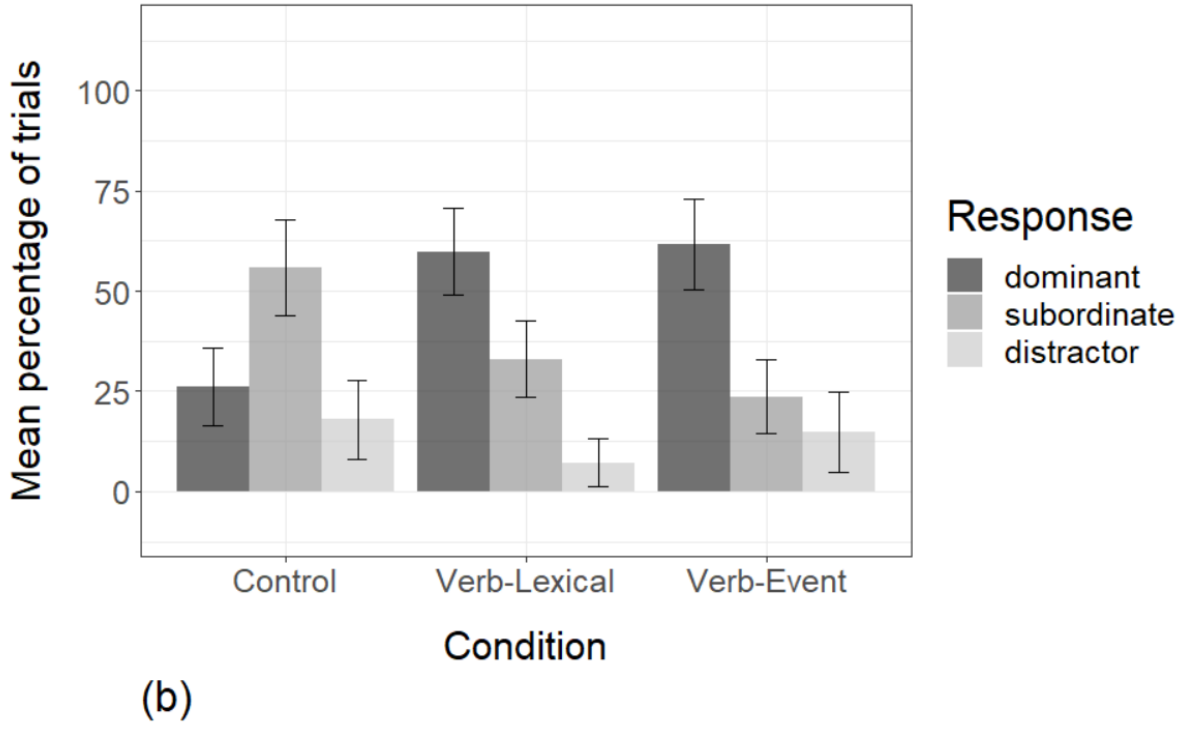
As can be seen in Figure 20, adults and children selected the *subordinate* meaning 96% and 56% of the time respectively in the control condition. This indicates that participants were able to integrate the sentence context to disambiguate the subordinate meaning of the target word.

An opposite pattern of responses, compared to the control condition, can be seen for both adults and children in the verb-event condition. Here, participants selected the *dominant* sense 88% and 62% of the time respectively. This suggests that they were able to rely on verb-event structures to select the dominant sense of the target words, over lexical associations from prior context pointing toward the subordinate. The difference in performance between control and verb-event

condition was significant for both adults (*Odds Ratio* = 759.56 [231.61, 2491.00], $p < .001$) and children (*Odds Ratio* = 7.39 [3.62, 15.11], $p < .001$).



(a)



(b)

Figure 20 Average percentage of trials in which an adult (panel a, *N* = 83) or child (panel b, *N* = 45) selected either dominant, subordinate, or distractor picture, as a function of condition (Control, Verb-Lexical, and Verb-Event). Error bars show 95% confidence intervals corrected for within-subject variance.

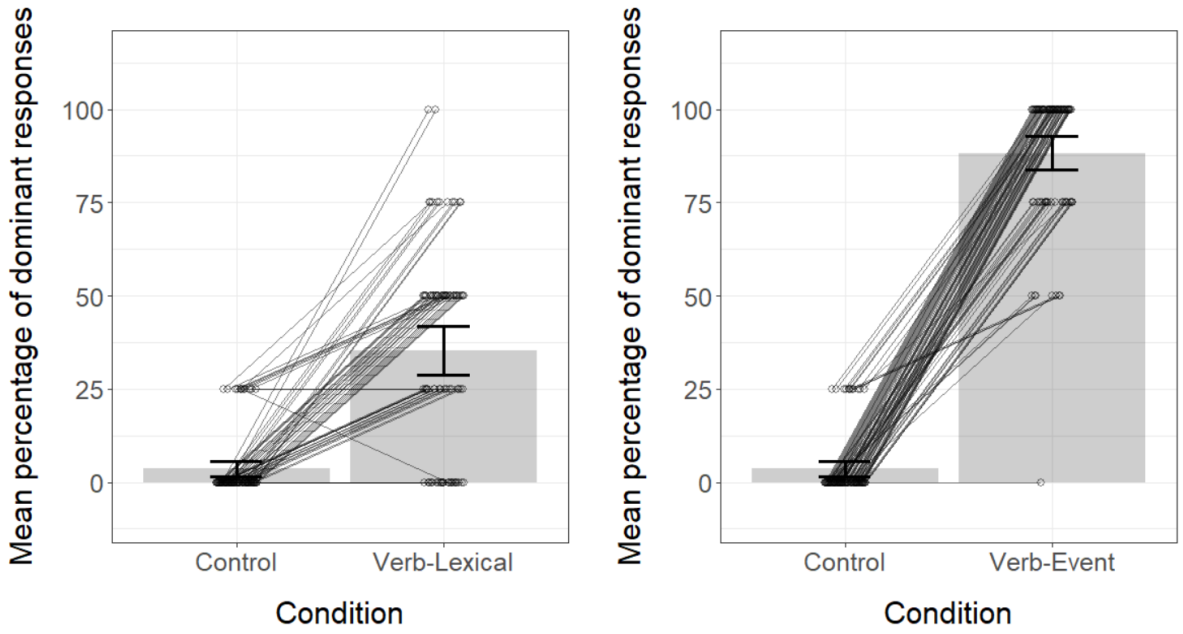
Adults and children responded differently in the verb-lexical condition, however. Adults mostly relied on prior context (65% *subordinate* meaning selection) while children relied on verb-sense lexical association instead (60% *dominant* meaning selection). In other words, when verb-event structure in the target context was neutral, adults likely preferred to rely on the global plausibility of the story triggered by the lexical associates included in the prior context (i.e., even if prior and target contexts were not strongly related in terms of coherence, still in "Sophia listened to some music. Then, she got a band" adults selected "band [music group]" because the speaker talked about "music").

Children, instead, relied on the lexical association between verb and dominant sense in the language (speakers often talk about "getting a band [object]" in real-world contexts). This result is in line with studies which showed children's reliance on verb lexical associations in sentence parsing (e.g., Kidd & Bavin, 2005; Snedeker & Trueswell, 2004; Yacovone et al., 2021).

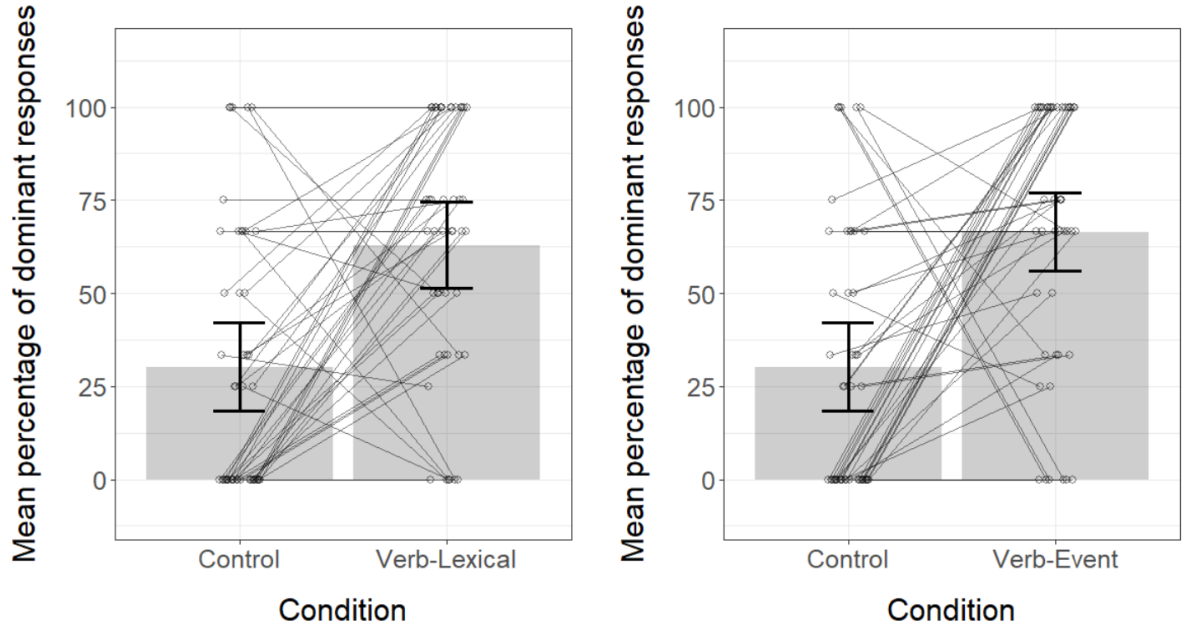
Interestingly, the difference in performance between control and verb-lexical conditions was significant not only for children (*Odds Ratio* = 4.68 [2.31, 9.49], $p < .001$), but also for adults (*Odds Ratio* = 25.29 [9.00, 71.05], $p < .001$). In other words, although adults likely relied on global plausibility guided by prior context associations, they still selected more dominant senses in the verb-lexical than in the control condition.

We further explore this result visually in Figure 21. Panel a shows that, in adults, there was more variability in dominant sense choice in the verb-lexical condition compared to the verb-event condition. This may indicate that adults also were residually sensitive to verb-lexical associations, given that the only difference between verb-lexical and verb-event conditions involved the verb preceding the

target. This result in adults is in line with their sensitivity to verb-patient lexical associations in studies where they are presented with (unambiguous) thematically appropriate patients differing in their strength of association to the verb (Andreu et al., 2013; Mani et al., 2016), and more generally with adults' sensitivity to lexical associations in sentence parsing (Khanna & Boland, 2010; Witzel & Forster, 2014).



(a)



(b)

Figure 21 Mean percentage of dominant meaning choice in each condition (with 95% confidence intervals corrected for within-subject variance) for adults (a) and children (b). Individual participants are also shown by data points, with lines that connect performance in Control and Experimental (Verb-Lexical or Verb-Event) conditions for each participant.

The pre-registered analyses conducted so far supported our two hypotheses regarding the key facilitative role of both verb-event structures and verb-lexical associations in early lexical ambiguity resolution. However, these analyses did not take into consideration how and whether the variables that characterize our experimental items (prior context associations, verb-sense associations, target sense dominance in child-directed speech) or participants' predictors (child verb knowledge from parent report) can capture participants' response variability, and whether the effect of different predictors was similar in adults and children. We have examined these questions in the following exploratory analyses, which can shed light on the cues adults and children relied on when resolving lexical ambiguities.

4.8 Exploratory Analyses

Conducting further analyses on predictors of performance can help us examine the differences between how adults and children performed. In our previous analyses, we did not directly compare the performance of the two age groups. Making a direct comparison would be useful to see if adults and children relied on similar variables but only differed in the magnitude of the observed effects. Another possibility is that the two groups differed in how sensitive they were to certain variables, leading to qualitative differences in performance.

One noticeable difference in performance between children and adults was observed in the verb-lexical condition. Adults predominantly selected the subordinate sense, whereas children mostly chose the dominant sense (Figure 20). One potential explanation is that adults may require only few lexical associations in the prior context to activate the subordinate sense, whereas children might need stronger

evidence, such as longer sentences or more robust associations, to activate the subordinate sense through bottom-up associations. In other words, the connection between "music" and "band" in "Sophia listened to some music. Then, she got a band" may be stronger in adults due to their greater language experience.

Alternatively, children might be more sensitive to sentence local information and may struggle to integrate variables from the broader context (e.g., Gertner & Fisher, 2012), resulting in a higher reliance on verb-sense associations (e.g., "getting an *elastic* band") compared to adults.

Additionally, children might be more attuned to the word-level characteristic of sense dominance, as they might not fully integrate the sentence context and instead rely more on word-level information.

To explore these possibilities, we combined the data from both adults and children. We then applied a mixed-effects model, considering the sense choice in the verb-lexical condition (dominant/subordinate) as the outcome, and the variables age group (adult/child), relative frequency of the dominant sense (dominance), verb-sense association, and prior context associations as predictors. This included two-way interactions between predictors, and three-way interactions between the age group and each pair of continuous predictors. The full output of this model is provided in Appendix S29.

Prior context associations were deduced by considering all the words in the preceding context and averaging their relative frequency of occurrence in child-directed sentences which contained the target subordinate sense. Note that we also obtained consistent results when computing prior associations from only content words, pronouns, and prepositions (Rabagliati et al., 2013), or only content words.

Our analysis revealed a significant main effect of age group (*Odds Ratio* = 4.20 [2.41, 7.32], $p < .001$). This indicates that adults selected the dominant sense significantly less frequently than children in the verb-lexical condition, thus supporting a quantitative difference in performance between the two age groups.

We also found a main effect of verb-sense association (*Odds Ratio* = 1.78 [1.25, 2.55], $p = .001$) and no significant interaction between verb-sense association

and age group (*Odds Ratio* = 0.78 [0.40, 1.37], *p* = .336). This result indicates that both adults and children were sensitive to this cue in similar way, as shown in Figure 22.

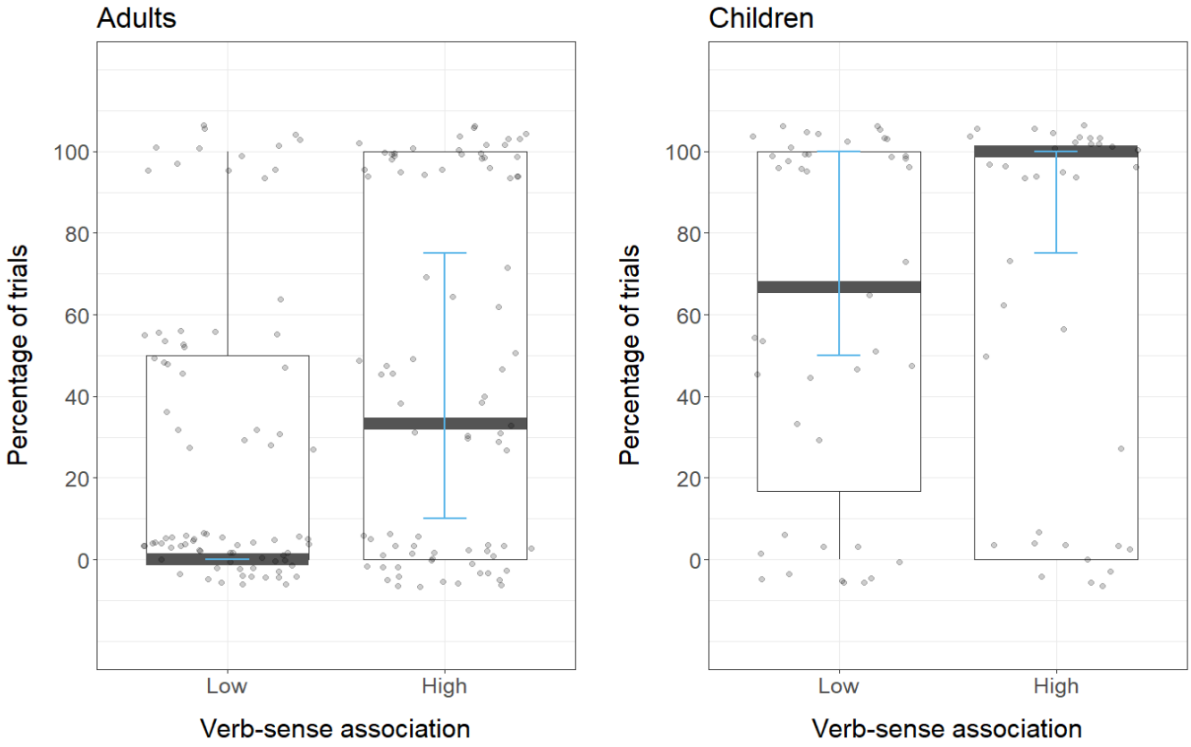
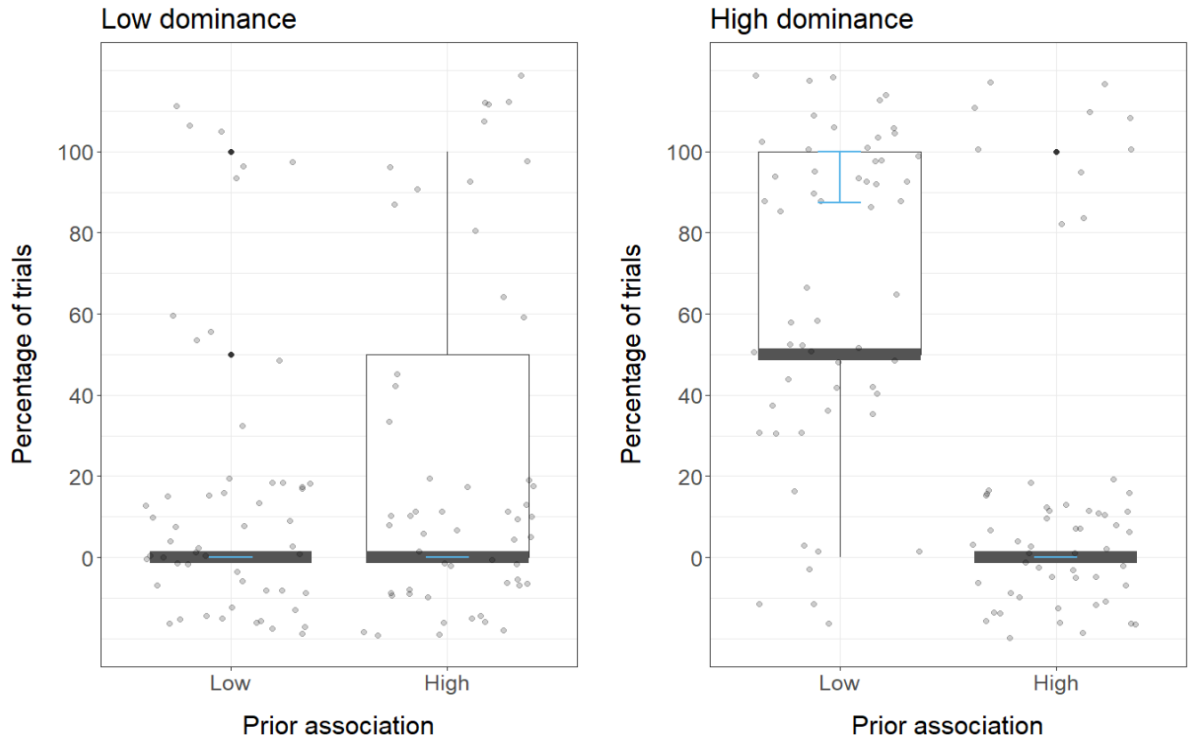


Figure 22 Boxplots showing the distribution of the percentage of dominant sense choices in the verb-lexical condition among adults and children, by verb-sense association. Black horizontal segments display the median values for each group. Verb-sense association was split at the median for graphical representation but was kept continuous in the statistical model. The data points represent individual participants and have been jittered to prevent visual overlap. The blue segments denote the 95% bootstrap confidence intervals of the median for each group, obtained from 1000 iterations.

Further, we found a significant interaction between prior association and dominance (*Odds Ratio* = 0.55 [0.38, 0.80], *p* = .002), as well as an interaction between age group and dominance (*Odds Ratio* = 0.38 [0.21, 0.67], *p* = .001). We visually examine these two interactions in Figure 23, where we plot percentages of

dominant sense choice as a function of prior association and dominance, for both adults and children. Consistent with the first interaction, for both adults and children prior associations had a significant effect only at high levels of dominance. This is attributable to a positive correlation between prior association and dominance in the experimental stories (r_s = .18), meaning the contrast between low and high prior association becomes more pronounced at high levels of dominance. As a result, at low levels of dominance, prior association showed no significant effect, regardless of age.



(a)

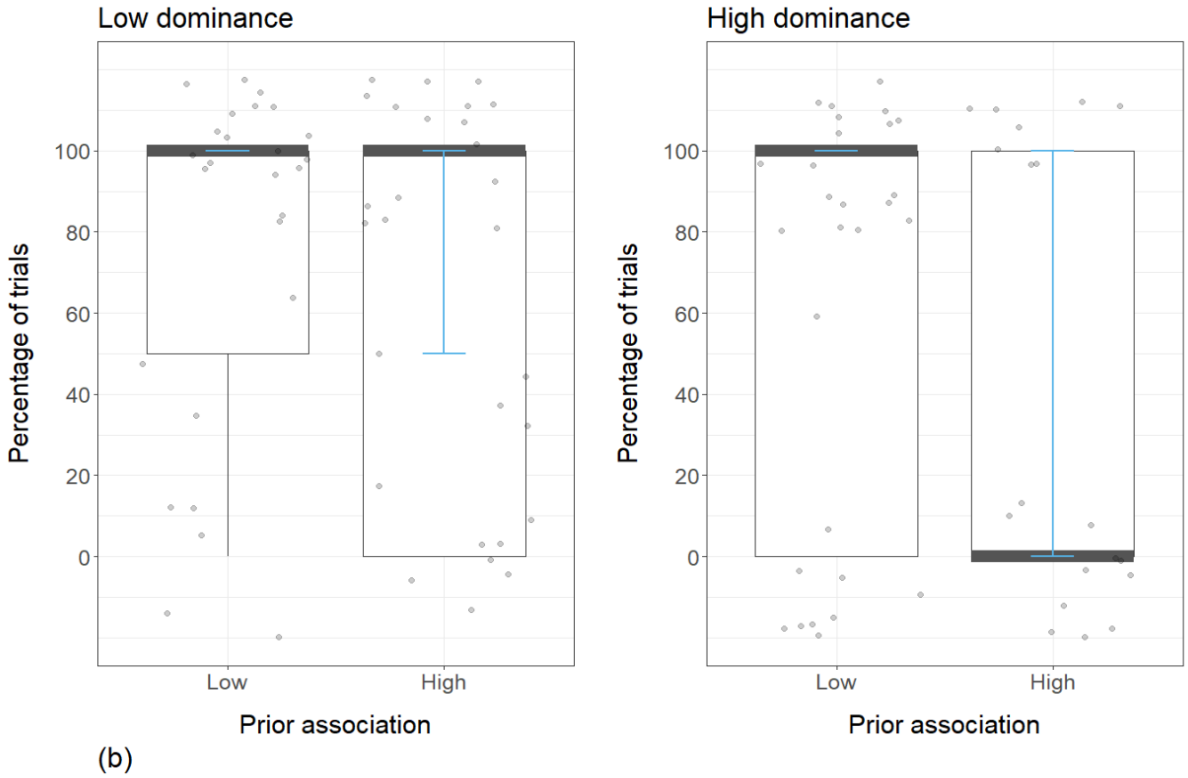


Figure 23 Boxplots showing the distribution of the percentages of dominant sense choice in the verb-lexical condition for adults (panel a) and children (panel b), as a function of prior association and sense dominance. Prior association and sense dominance were split at their median for graphical representation but were kept continuous in the statistical model. The data points represent individual participants and have been jittered to prevent visual overlap. Black horizontal segments display the median values for each group. The blue segments denote the 95% bootstrap confidence intervals of the median for each group, obtained from 1000 iterations.

Also, at low levels of dominance, children's performance was at ceiling (i.e., they almost always selected the dominant meaning) while adult performance was at floor (i.e., they almost never selected the dominant meaning, consistent with the second interaction found). This could suggest that even when dominance was not strongly pronounced, children were still clearly sensitive to it, being more receptive to the word-level frequency of the dominant sense. This might have determined a qualitative difference in performance. However, the plausibility of this explanation is unclear as we would also expect high dominance to mitigate the effect of high prior

association. Yet, what we see in children's performance is a floor effect (evidenced by the boxplot on the far right of Figure 23b). Perhaps it is more likely that adults simply exhibited less overall sensitivity to sense dominance because our dominance variable may not accurately reflect adults' representations. In fact, the dominance counts were based on child-directed speech and may not accurately reflect sense dominance in adult-directed speech. Frequency counts for adult-directed speech were unavailable, as tagged adult corpora usually consist of written text rather than spoken conversations.

Importantly, these results indicate that children, like adults, demonstrated sensitivity to prior context associations and were similarly responsive to verb-sense associations. This shows they were capable of using both broader and local contexts for disambiguation.

Finally, could it be that children's high sensitivity to dominance might have assisted them in selecting the dominant meaning in the verb-event condition, regardless of their knowledge of verb-event structures? To further examine the role of sense dominance, we fitted another exploratory mixed-effects model (Appendix S29) where sense choice in the verb-event condition was the outcome variable, with age group (adult/child), prior association, and sense dominance as predictors (including two-way and three-way interactions). We only found a main effect of age group (*Odds Ratio* = 0.32 [0.15, 0.70], $p = 0.004$), with adults selecting significantly more frequently the dominant sense in this condition, therefore indicating a quantitative difference in performance between the two age groups. Moreover, we found no effect of sense dominance (*Odds Ratio* = 0.87 [0.37, 2.05], $p = 0.758$) nor prior associations (*Odds Ratio* = 0.87 [0.36, 2.13], $p = 0.763$), suggesting that adults and children likely relied on verb-event structures to disambiguate the target words.

To further investigate this result, we fitted an additional mixed-effects model on child data only (Appendix S29), using sense choice in the verb-event condition as the outcome, with prior association, sense dominance, and verb production (*Not Produced* = Not used or Understand only; *Produced* = Understand and Use) as predictors (including two-way interactions). Verb production was computed from our

parent-report questionnaire. Interestingly, Verb production was the only significant predictor in this model (*Odds Ratio* = 3.36 [1.08, 10.49], p = 0.037), with children being more likely to select the dominant meaning if parents reported production of the preceding verb (see Figure 24).

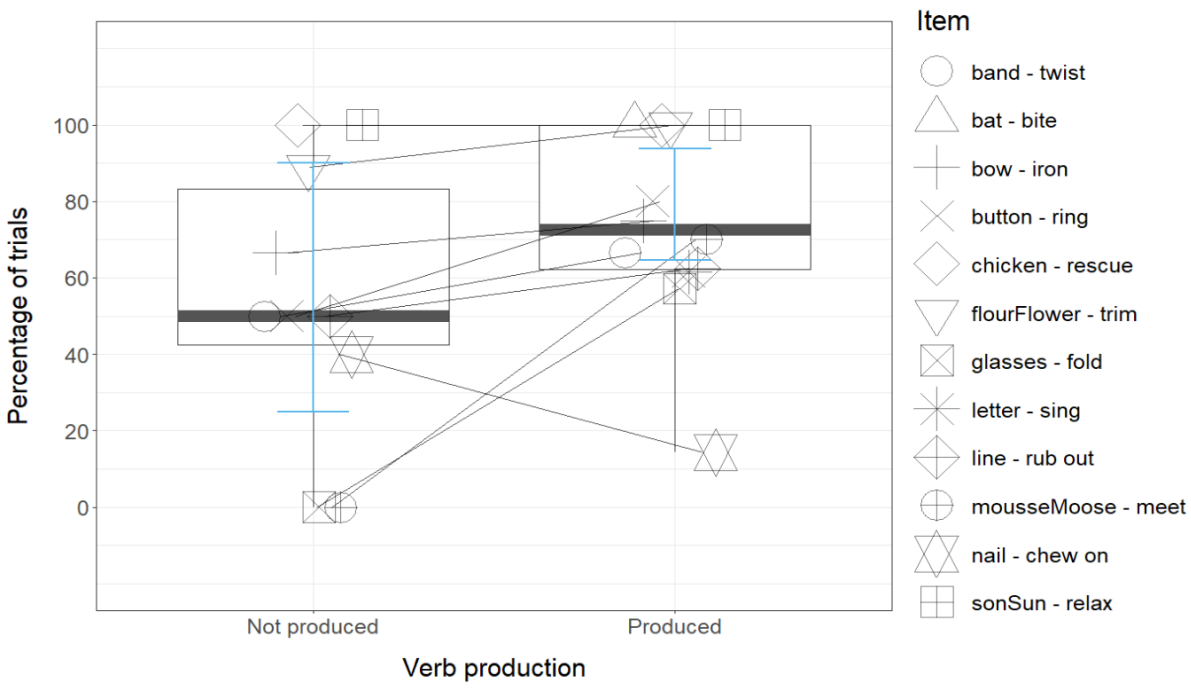


Figure 24 Boxplots showing the distribution of percentage of dominant sense choice in the verb-event condition, when caregivers reported the experimental verb was produced by the children (Produced) or not (Not produced). Item variability is shown through connecting lines. The blue segments denote the 95% bootstrap confidence intervals of the median for each group, obtained from 1000 iterations.

This provides preliminary evidence that children relied on their understanding of verb-event structures in the verb-event condition (assuming that being able to produce a verb is indicative of more consolidated knowledge of verb-event structure). It is worth noting that, if these assumptions hold true, one would also anticipate the association with verb production to be less pronounced in the verb-lexical condition (i.e., the verb-event structure is neutral in this condition, thus knowledge of the verbs should not predict children's performance). However, we

could not conduct the same analysis for the verb-lexical condition as the vast majority of verbs in this condition were reported by parents as being produced by nearly all children in our sample.

In conclusion, our exploratory analyses have overall indicated that children were sensitive to sentence context, and that our manipulations of verb-lexical associations and verb-event structures may have functioned as cues for lexical disambiguation. While children displayed a similar degree of sensitivity to sentence-context cues as adults, their overall performance was significantly lower than that of the adults. This suggests that their abilities to use the different cues presented in the experimental task may not have yet reached the adult state. Finally, the only qualitative difference in performance we found involved children's higher reliance on sense dominance in the verb-lexical condition compared to adults. However, this result requires further exploration in future work to rule out the possibility that adults' representations of sense frequency distributions might be different from that of children due to differing uses of senses in adult-directed speech.

4.9 Discussion

Although theories of word learning (e.g., Markman, 1991; Trueswell et al., 2013; Yu & Smith, 2007) and word processing (e.g., Plaut et al., 1996; Seidenberg & McClelland, 1989) typically do not account for lexical ambiguity and predict that young children do not map word forms to multiple meanings, recent evidence indicates that child-directed speech is rich in word sense ambiguity, and the same is true for children's early vocabularies (Meylan et al., 2021). In this study, we found evidence suggesting that children as young as four years of age map and can process at least two alternative meanings of familiar noun forms, supporting similar findings on early lexical disambiguation (e.g., Hahn et al., 2015; Rabagliati et al., 2013). Importantly, we examined the interplay between bottom-up and top-down cues at the sentence level. Our findings show that, when these cues were in direct competition with prior lexical associations pointing toward a subordinate meaning, children preferentially relied on verb-sense associations (verb-lexical condition) or

verb-event structures (verb-event condition) to select the dominant sense of an ambiguous target. In subsequent analyses of predictors of child performance, we identified verb-sense associations as a significant predictor in the verb-lexical condition and verb-event knowledge (measured by a proxy of parent-reported verb knowledge) as a significant predictor in the verb-event condition. These findings provide the first evidence that children can resolve a contrastive task by integrating and mostly relying on either bottom-up verb-lexical or top-down verb-event structure cues during sentence parsing for disambiguation. This supports the idea that children's word representations are contextual, rich in both bottom-up and top-down aspects from early in their development (Srinivasan & Rabagliati, 2021). Moreover, this study highlighted the importance of leveraging naturalistic conversations to disentangle the effect of bottom-up and top-down cues at the sentence level, which are often confounded due to the ubiquitous role of lexical statistics in language development (e.g., Ambridge et al., 2015). In the following paragraphs, we discuss each finding that emerged in the study.

In the pre-registered analyses, we found that children could switch from the subordinate meaning implied by the prior context to the dominant meaning implied by the target context. Additionally, in the exploratory analyses, we showed that children weighed different cues at the word and sentence level to make categorical decisions on the target word meanings. Importantly, children were sensitive to the strength of each of the cues present in the experimental task, from the relative frequency of each target meaning in the language, to more distant effects from bottom-up prior context lexical associations, and more local sentence cues like bottom-up verb-sense associations or top-down verb-event information.

Children were not only capable of handling multiple bottom-up and top-down cues in sentence parsing for disambiguation, but these cues also related to child performance in a continuous manner. Specifically, the more frequently a target meaning occurred in the language, the higher the likelihood that children selected this dominant meaning during the task. This evidence aligns with findings about children’s faster naming latencies when a prime activates a target dominant sense (Booth et al., 2006; Simpson & Foster, 1986).

At the same time, we found that sentence context can serve to mitigate the subordinate-bias effect by providing a supportive environment for less frequently used meanings of a word. In fact, the stronger the word associations in the prior context with the subordinate target meaning, the higher the likelihood of children selecting the subordinate meaning. This influence of bottom-up word associations supports findings such as those from Rabagliati et al. (2013), who found that prior associations from global context (preceding sentence) could explain some of the variance in child lexical disambiguation. It is also worth noting that, in that study, the influence of lexical associations was found in coherent tasks, where all sentence cues pointed toward one meaning. The present study replicated the result but in a contrastive task, showing that prior associations play a role also when put in competition with other cues to word meaning.

Moreover, the more frequently a verb was used with a specific target dominant sense in child-directed speech—for example, "getting an *elastic* band"—the higher the likelihood that children selected the dominant sense in the verb-lexical condition. This result aligns with other findings showing the importance of lexical statistics that verbs carry with them for child language processing (e.g., Kidd & Bavin, 2005; Mani et al., 2016; Snedeker & Trueswell, 2004; Yacovone et al., 2021), and more generally, is in line with evidence showing that children leverage sentence local information in sentence parsing (e.g., Gertner & Fisher, 2012).

Lastly, the stronger a child's knowledge of a preceding verb's semantics (as indexed by their use of the verb in production), the higher the likelihood of the child selecting the (dominant) sense in the verb-event structure condition, where that sense represented a more plausible argument for the verb. This result is line with studies on unambiguous word processing that have shown how children can use verb-event structures in sentence parsing to judge novel arguments of familiar verbs (e.g., Andreu et al., 2013).

Overall, the relations between predictors and child performance provide evidence in support of a cue-validity account (Trueswell & Gleitman, 2007): Children are able to weigh different bottom-up and top-down cues depending on their reliability to determine word meaning. This account also suggests that as children

grow, they gradually refine their estimation of the general reliability of each cue in determining word meaning. Such fine-tuning could account for differences in how children of varying ages apply these cues. Thus, to further support this account, it would be worthwhile to study the evolution of cue weighting in our task throughout childhood, both quantitatively and qualitatively. For instance, as children grow and gain a deeper understanding of the role that top-down cues play in sentence structure, we anticipate that they will increasingly rely on these cues. For example, a more comprehensive understanding of naturalistic event sequences could enable children to recognize the importance of global plausibility in contrastive lexical ambiguity tasks.

Finally, the significant differences between control and experimental conditions as well as the significance of different predictors of performance align with a semantic-settling model of lexical processing (Rodd, 2020). This model emphasizes the interactive role of word-level sense-dominance and sentence-level bottom-up and top-down cues to ambiguous word representation. Evidence supporting this account primarily comes from adult processing, so this study has contributed to extending the model's predictions to child processing.

4.9.1 Future Directions

4.9.1.1 *Adult and Child Performance*

We found both quantitative and qualitative differences in performance between adults and children. Below, we discuss how different explanations for these differences could be investigated in future work.

Both adults and children predominantly selected dominant senses when these were primed by verb-event structures (more than 50% dominance sense choices). This suggests that both groups could rely on their verb-event knowledge and override the influence from bottom-up prior context lexical associations. However, adults more consistently chose dominant senses than children, indicating a quantitative difference in performance. One reason for this difference might be that children's limited language experience with verbs could have negatively affected

their use of verb-event structures compared to adults. As observed, verb production was a significant variable in predicting child performance in the verb-event condition. Thus, differences in performance could be due to stronger and richer verb representations in adults.

Another possible explanation, which is not mutually exclusive, could relate to children's difficulties in revising their interpretations towards the dominant sense after initially committing to the subordinate upon hearing prior context associate words. This explanation would align with studies where children show difficulties in revising their interpretations (e.g., Kidd & Bavin, 2005; Qi et al., 2020; Yacovone et al., 2021). These processing difficulties could potentially be explained by immature executive function abilities (e.g., Khanna & Boland, 2010). If these difficulties are present, we expect that children would switch to the dominant sense more slowly than adults. Therefore, in future research, it would be important to implement a direct measure of switching behaviour based on real-time sentence processing. For instance, we hypothesise that when examining those trials in which participants begin by fixating on the subordinate sense, children would be slower than adults to redirect their eye-gaze to the dominant meaning upon hearing the disambiguating verb. Furthermore, we would expect measures of executive function abilities to account for age group differences beyond the influence of verb knowledge.

While we did not employ a direct measure of switching behaviour in this study, our comparison between the control and verb-event condition could still be seen as an indicator of group-level switching behaviour. In simpler terms, we evaluated how challenging it is for children to switch to the dominant meaning when considering their baseline preference for the subordinate sense in the control condition. For instance, we found that children chose the dominant sense in 62% and 25% of trials in the verb-event and control conditions, respectively. The difference of 37% could be considered a proxy of switching performance. The difference between conditions was considerably higher in adults (88% - 4% = 84%), suggesting that their switching performance may have been better. We further examined these differences between adults and children in an additional exploratory statistical model (see Appendix S30). This additional model indicated that the

difference between adults and children was significant (*Odds Ratio* Age Group * [Control vs. Verb-event] = .02 [.01, .05], $p < .001$). However, without a measure of executive function skills, we cannot determine whether this age group difference was exclusively due to children's lack of verb-event knowledge, or whether processing difficulties also played a part. Future research should explore this issue in depth.

Aside from processing limitations, it is worth noting that the involvement of limited language experience might also explain the qualitative difference in the verb-lexical condition. When exposed to stories with neutral verbs (e.g., "Sophia listened to some music. Then she got a band"), we would expect bottom-up lexical associations from the prior context (listen to, music) to trigger adults' top-down event knowledge (i.e., the only interpretation that forces the story to have global coherence is "music band"). As shown in the second experiment of Rabagliati et al. (2013), 4-year-olds might not yet have mastered the ability to integrate sentence cues into a global interpretation of discourse. This skill requires children to use their pragmatic abilities, reasoning about real-world events and how these are expressed through language. Therefore, the lack of rich language representations might have determined children's higher reliance on the word level (i.e., sense dominance) instead of the more global discourse context.

In summary, while children have demonstrated the ability to use different sentence context cues for disambiguation, they have not yet reached the adult stage. To better understand the quantitative and qualitative differences between adults and children, future work should consider examining performance at different child ages, collecting measures of cognitive abilities such as inhibition skills (e.g., Khanna & Boland, 2010), as well as measures of pragmatic knowledge of events, which have not yet been considered when assessing child lexical disambiguation (e.g., tests that can help assess children's understanding of causal relationships, event sequences, and social norms, e.g., Khan et al., 2016).

4.9.1.2 *Learning Mechanisms*

What types of learning mechanisms might be involved in children's use of verb knowledge for lexical disambiguation? A semantic-settling account posits that the significance of various cues to disambiguation changes through language experience (Rodd, 2020). This account assumes that the individual uses mechanisms that track and store highly specific knowledge about the distributional statistics of word meanings' usage. This context-dependent knowledge is hypothesized to exert an effect both in the long term and short term (Rodd et al., 2016; Wiley et al., 2018). For instance, baseball players have trouble inhibiting baseball-related senses of target words, even in fully disambiguating contexts. An example would be still activating the sense "baseball bat" in the story "Monica had a great fear of things flying around her head; she looked for the bats that lived in the shed" (Wiley et al., 2018). Further, a recent encounter with a sense of an ambiguous word boosts its availability and reactivation from minutes to a day (Gaskell et al., 2019; Rodd et al., 2016).

Rodd's (2020) proposal of distributional learning mechanisms that gradually form sense representations aligns with usage-based accounts of language development (e.g., Abbot-Smith & Tomasello, 2006; Ambridge 2019; 2020; Bybee, 2010b). These accounts suggest that the child undergoes continuous readaptation of their representations through language experience. Usage-based accounts posit that at least two mechanisms are involved in language learning: a simple associative learning mechanism, which is sensitive to different sources of statistical regularities in the linguistic input, and an analogical mechanism that analyses common features between similar linguistic exemplars encountered to abstract linguistic structures. For instance, related to the specific context of our study, Alishahi and Stevenson (2007) proposed a Bayesian computational model that learns verb-event structures by gradually generalizing over the semantic characteristics of the verb's syntactic arguments encountered during training. The learning process is enabled by a probabilistic mechanism that first records the associative relation between a verb and its semantic features, the position of the syntactic arguments, and the semantic characteristics of the items that progressively fill those arguments. The model draws

analogies across similar sentences and forms verb construction representations, which can then generalize arguments to novel items. Through incremental probabilistic learning, the model successfully simulated adult plausibility judgments of verb-selectional preferences (e.g., "The mechanic warned the driver" versus "The mechanic warned the engine").

In principle, the simulations by Alishahi and Stevenson (2007) demonstrate that usage-based mechanisms can develop verb construction representations and assist in processing novel arguments. These findings imply that a usage-based approach might be promising for understanding how verb constructions assist children in disambiguating words with multiple meanings. However, it is important to note that the simulations assumed a wealth of knowledge on the part of the learner, including a rich understanding of verb and argument semantics (derived from dictionaries), ideal identification of syntactic arguments, their semantics, and their positions relative to the main verb. Therefore, it remains unclear whether a usage-based learner could capture verb-event influences on child lexical disambiguation when applied to raw, naturalistic language, without the need to supplement the input with external resources that enhance sensitivity to syntactic and semantic information.

4.10 Limitations

An important aspect of word disambiguation, which we have not explored, involves the semantic relation between different senses of a target form. Specifically, a key distinction is often drawn between polysemous and homophonous words. Polysemous word forms correspond to senses that are semantically related (e.g., chicken as an animal or meat), while homophonous senses do not share this relation (e.g., bat as an animal or an object). Research on this subject has demonstrated that polysemy is particularly advantageous for children's mapping of word forms to multiple novel senses (e.g., Floyd & Goldberg, 2021). This is likely because they have already mapped certain semantic dimensions in common between the senses and only need to focus on the new elements that determine sense difference.

In word processing, polysemy also presents an advantage, with the reason being that homophony increases the competition between senses, resulting in slower reading disambiguation in adults (e.g., Frisson, 2009). For children, instead, no processing differences were found between polysemous and homophonous words (Rabagliati et al., 2013). However, this conclusion was drawn from an offline measure from a forced-choice task, thus it might be possible to still find differences with more fine-grained measures of performance (e.g., tracking eye-movements as the spoken stories unfold in a looking-while-listening task).

In this study, we did not investigate differences between ambiguity types due to the limited number of items used ($N = 12$). This limitation was determined by the constraints in the number of available items that were sense-tagged in the ChiSense-12 corpus. The forthcoming release of the sense-tagged corpus by Meylan et al. (2021) will increase the number of items available for testing, enabling researchers to evaluate different ambiguity types in our or similar tasks. It is important to note, however, that the distinction between polysemy and homophony is typically based on dictionary definitions. Research has demonstrated that there is considerable variability in perceived relatedness of senses in polysemy (Klein & Murphy, 2001). This variation can present complications when attempting to test ambiguity types experimentally, as these types cannot be easily defined.

Finally, while our study has benefited from the use of a sense-tagged corpus of child-directed speech, the application was constrained in its relevance to adults' performance. For instance, although children displayed a higher reliance on sense dominance compared to adults, it was not possible to conclude that sense dominance plays a more substantial role in children's than adults' disambiguation because our measure of sense dominance was derived from a corpus of child-directed speech. Sense dominance might exhibit different distribution patterns when considering adult-adult conversations. In future research, it would be advantageous to either employ sense-tagged adult-directed conversations (as soon as they become available) or to conduct a pilot study wherein sense dominance is inferred through online measures, such as reaction time or total looking time toward alternative senses upon hearing a target word in isolation. These strategies would more

effectively differentiate the effects of sense dominance in the two age groups and more accurately assess whether the extent to which the subordinate bias effect is flexible and responsive to context changes during development.

4.11 Conclusion

This study examined children's ability for lexical disambiguation, focusing on the interplay between bottom-up and top-down cues in word interpretation. The results showed that children both use and show sensitivity to the strength of multiple disambiguation cues. This supports recent models of lexical processing that highlight the contextual nature of word representations. The study also underscored both the utility and limitations of combining corpus analyses of naturalistic conversations with experimental work. On the one hand, this approach allows for meticulous control over variables that often intertwine in naturalistic speech. On the other hand, it also emphasizes the challenges associated with the limited availability of data sourced from language corpora. Lastly, the study found quantitative and qualitative differences between adults and children. These differences prompt new questions about the learning mechanisms, processing constraints, and aspects of language experience that may drive the transition to adult disambiguation skills.

Chapter 5
General Discussion

5.1 General Aims of the Thesis

This thesis focused on understanding how experiences with naturalistic speech can influence the way infants and children learn and use words. A core assumption of the usage-based theory of language development is that children form word representations by consistently attending to linguistic input through domain-general cognitive processes (Bybee, 2010b). This assumption carries two implications: First, computational models of domain-general processes - when applied to naturalistic language input - should be able to capture children's performance. Second, effects of children's naturalistic language experiences should be evident when studying how they learn and process words. In this thesis, I identified two research areas where assessing these implications of the usage-based theory of language development has been challenging, mostly due to methodological issues associated with connecting what children hear in their naturalistic environments to their actual word learning and processing. Across three empirical chapters, I introduced novel approaches to evaluate the impact of naturalistic language experiences and the role of domain-general processes in early word learning and processing.

In the following sections, I provide a summary of the findings from the empirical chapters, alongside a discussion on the implications of these studies, their limitations, and suggestions for future research.

5.2 The Role of Chunking in Early Naturalistic Word Segmentation and Word Learning

5.2.1 Chapter 2. CLASSIC Utterance Boundary: A Chunking-Based Model of Early Naturalistic Word Segmentation

Chapter 2 examined the role of a domain-general cognitive process of chunking (i.e., the ability to learn from associations) in how children segment words from

naturalistic speech. While previous studies have shown that chunking-based computational models can segment naturalistic speech with high accuracy (e.g., French et al., 2011; Monaghan & Christiansen, 2010), there has been a lack of evaluation metrics that link model segmentation accuracy to infants' real-world performance. To address this, I introduced new evaluation measures that connect model segmentation accuracy to the naturalistic word productions found in child speech. Using these measures, a developmentally plausible model is characterized as one that accurately segments words which children produce early on, and also acquires a vocabulary that mirrors that of children in terms of various word-level characteristics: Word length, word frequency, neighbourhood density, and phonotactic probability. Moreover, I introduced a new computational model of early word segmentation, CLASSIC-UB. This model is an extension of the vocabulary learning model CLASSIC (Jones & Rowland, 2017; Jones et al., 2021; Jones, 2016; Jones, Justice, et al., 2020), which I have adapted to segment naturalistic continuous speech. CLASSIC-UB performs word segmentation by combining the learning of frequently associated phonological sequences with information at the start and end of utterances. I evaluated the performance of CLASSIC-UB against previous chunking and nonchunking models of segmentation, including the chunking model PUDDLE (Monaghan & Christiansen, 2010) and two models that identify word boundaries in speech based on sound transitional probabilities (e.g., Saksida et al., 2016).

The primary finding of this study was that a model's ability to segment naturalistic input with high accuracy (i.e., identifying the largest proportion of words from the input) does not necessarily make it developmentally plausible. Interestingly, although CLASSIC-UB segmented naturalistic input with lower accuracy than the competing chunking model PUDDLE, it performed better than PUDDLE at capturing the distribution of words children produce, in terms of word length and neighbourhood density. An analysis of the performance differences among models showed that CLASSIC-UB's advantage was related to its use of overlapping phonological sequences at varied grain sizes. This characteristic helped the model in learning words typically difficult to learn (those that are long and have few phonological neighbours in the language) but are still acquired by children.

The study's second key finding was that sensitivity to both utterance-initial and utterance-final cues improved CLASSIC-UB's segmentation accuracy. However, only utterance-final cues played a role in capturing child vocabularies. This distinction is related to specific characteristics of the input at utterance boundaries. In child-directed speech, a diverse set of novel words often appears in the utterance-final position, making this cue particularly beneficial for vocabulary acquisition. In contrast, fewer novel words are found in utterance-initial position, making them less relevant for word acquisition. Still, words in the utterance-initial position are frequently repeated (e.g., function words). This aspect of token frequency allows the model to segment these words with high accuracy, thereby improving its performance in segmentation accuracy measures.

Finally, the study highlighted the better performance of chunking models (CLASSIC-UB, PUDDLE) over transitional probability models in terms of naturalistic segmentation measures and in capturing child production measures. The chunking models consistently outperformed the transitional probability models across all accuracy metrics. They also more effectively captured child vocabulary metrics at the lexical level, such as word length, word frequency, and neighbourhood density. This observation is consistent with earlier studies showing that chunking models segment naturalistic input with greater accuracy than transitional probability models (e.g., Larsen et al., 2017). Additionally, chunking models have been found to more accurately capture lexical effects on word segmentation in lab settings compared to transitional probability models (e.g., French et al., 2011; Kurumada et al., 2013). Overall, these results indicate that the cognitive process of chunking might play a significant role in early naturalistic word segmentation and learning.

5.2.2 Chapter 3. Simulating Early Word Segmentation and Word Learning from Italian Child-Directed Speech

Chapter 3 extended the first study of the thesis by examining the cross-linguistic validity of its findings, using Italian as a case study. Overall, the research confirmed the better performance of chunking models over transitional probability models in

segmenting naturalistic input and capturing child vocabularies. The advantage of CLASSIC-UB over other models was also further confirmed in the Italian context, supporting the first study's conclusions about its developmental plausibility.

This second study investigated how the variability previously observed in model segmentation accuracy across languages (e.g., Saksida et al., 2016) might relate to the models' developmental plausibility (as assessed using the new evaluation metrics introduced in Chapter 2). In this study, I conducted a detailed analysis to determine whether the results from the English simulations could be extended to Italian. Additionally, I aimed to understand how outcomes varied based on specific characteristics of the language under investigation.

The primary focus of the second study was to assess how the average word length of a language influenced simulation results. The study on English examined potential basic perception units (phonemes and syllables) that infants might initially use to tackle the segmentation problem and how these choices impacted model learning. However, conclusions regarding this aspect could not be drawn, as the new evaluation metrics applied to syllabic input demonstrated low sensitivity. I hypothesized that this issue likely arose because English contains a significant proportion of monosyllabic words. For this reason, a random baseline model exhibited high performance, leaving minimal room for other models to account for variability in child vocabularies. Therefore, for the second study, I used Italian child-directed speech, which has a smaller proportion of monosyllabic words. While using Italian did improve the sensitivity of the models, it only allowed for meaningful comparisons across syllable-based models in terms of word segmentation accuracy, not in word production. Subsequent exploratory analyses indicated that this limited sensitivity was determined by neglecting the role of word frequency in word production measures. This factor was overlooked in word-level measures that considered the pool of unique word types a model acquired from the input.

By replicating the simulations with Italian and employing more sensitive measures, I was able to link segmentation accuracy to child production vocabularies when considering both phonemes and syllables as basic units of infant speech perception. This analysis led me to conclude that segmentation models generally

performed better when exposed to phonemic rather than syllabic input. This finding aligns with existing research emphasizing the importance of subsyllabic units in early word processing and learning (Fais et al., 2012; Jusczyk & Aslin, 1995; Mattys & Jusczyk, 2001; Mani & Plunkett, 2010).

A second significant finding related to the impact of differences in average word length across languages on CLASSIC-UB's ability to capture both sublexical (phonotactic probability) and lexical (neighbourhood density) characteristics of child vocabularies. I found that CLASSIC-UB held an advantage over other competing models in capturing English children's vocabulary productions by neighbourhood density. However, it showed no such advantage when capturing word productions by phonotactic probability. Conversely, in Italian, the situation was the opposite: CLASSIC-UB outperformed other models in terms of phonotactic probability but not with neighbourhood density. In relation to these findings, I observed that the Italian language offers a greater proportion of short biphone sequences than longer phonological sequences. These short biphone sequences were likely used by the model to improve its vocabulary learning, making it particularly proficient at capturing phonotactic probability in Italian child vocabularies. In contrast, English provides longer sequences that that are captured by the definition of a phonological neighbour. CLASSIC-UB was also sensitive to these longer sequences, as its learning spans multiple chunk lengths, allowing it to capture the neighbourhood density of English child vocabularies. In essence, when we examined the impact of word length, we found that overlapping phonological sequences affected CLASSIC-UB's learning in both Italian and English. The extent of this effect was determined by the specific language input's prominent features.

Another level of investigation concerned the differences between English and Italian in terms of morphological complexity. Previous studies have shown morphological complexity to impact models' segmentation accuracy (e.g., Phillips & Pearl, 2014). The results of the second study supported past findings that higher rates of oversegmentation occurred in languages with richer morphologies, such as Italian. This oversegmentation was also related to the discovery of morphological units in the speech input. Most importantly, for the first time, I examined how rates

of oversegmentation affect model developmental plausibility. I found that oversegmentation did not negatively impact CLASSIC-UB's ability to capture the properties of child vocabularies. Furthermore, the model was able to acquire morphological units as well as word forms, in line with Italian children's early comprehension of morphology (e.g., Ferry et al., 2020).

Finally, the cross-linguistic extension of previous simulations allowed me to investigate the advantage of utterance-final boundaries in capturing child vocabularies. Italian child-directed speech is distinctive because it contains a higher proportion of verbs than nouns, yet children still learn more nouns (Longobardi et al., 2015). Italian nouns often appear in the utterance-final position, which could explain the children's noun advantage. The study revealed that chunking models sensitive to utterance-final cues exhibited a noun advantage as in child vocabularies, despite being exposed to a higher proportion of verbs. This supported the significant role of utterance-final cues in vocabulary learning. Importantly, this aspect could not be adequately investigated using English child-directed speech since, while nouns also appear in utterance-final position, they constitute the largest word category (i.e., utterance-final word frequency and word category frequency are entangled). This constitutes a confound, making the use of Italian speech important for assessing the role of utterance-final cues.

5.2.3 Implications for the Study of Early Word Segmentation and Word Learning

The findings from the studies presented in Chapters 2 and 3 have several implications for research on early naturalistic word segmentation. When modelling either English or Italian segmentation, I discovered that the results from segmentation accuracy measures were not always in agreement with those based on model performance on developmental data. This potential misalignment between these two types of measures has been highlighted in previous work (Larsen et al., 2017). The studies in this thesis confirm the need for greater attention to the type of evaluation measures used to assess developmental plausibility.

Attention has previously been given to building models based on cognitively plausible assumptions (e.g., Phillips & Pearl, 2015). Discussions have centred around which underlying learning mechanisms, cognitive constraints, and speech perception units might best approximate the actual segmentation task infants face. Additionally, the findings in this thesis emphasize the need to delve deeper into constructing evaluation measures that more closely align model performance with developmental data. This alignment is particularly challenging in the context of naturalistic segmentation. Carefully designed experimental tasks can produce scores that reflect the types of representations infants acquire in laboratory settings. These scores can be directly compared to model outputs to infer developmental plausibility (e.g., French et al., 2011; Perruchet & Vinter, 1998). However, it is more challenging to extract scores that represent infant comprehension in real-world settings. One solution I proposed in this thesis is to capitalize on the close relation between word segmentation and word learning. By using production vocabularies from extensive corpora of naturalistic conversations, one can create rich sets of developmental measures of model performance.

The strength of the approach based on child production lies not only in providing a child-based benchmark for model evaluation and comparison but also in facilitating a detailed examination of the architectural differences that cause variation in model performance. Specifically, child productions offer large sample sizes that enable an in-depth analysis of different properties of child vocabularies. This approach overcame the lack of sensitivity in the word age of acquisition measures based on CDI estimates for both English (Chapter 2) and Italian (Chapter 3). This insensitivity (also discussed in point 5 of the Notes section) was due to the samples of words available from the CDI being smaller than those available in the production vocabularies extracted from the CHILDES database. Moreover, the large samples of production vocabularies allowed me to examine interactions between variables. For instance, the advantage of CLASSIC-UB in capturing word length and neighbourhood density distributions was assessed by examining how its alignment with child distributions improved with increasing word frequency. In this context, the large sample size made it feasible to explore interactions between word frequency and other word-level measures. Another benefit of using large samples of child

productions is the opportunity to analyse the roles of type and token frequencies. This was evident when exploring the segmentation and acquisition of words at utterance boundaries in both English and Italian, also considering how these variables might influence the acquisition of words from different part-of-speech categories. Overall, the results presented in this thesis suggest that the proposed model evaluation method holds promise for detailed investigations into the plausibility of various learning mechanisms associated with child word segmentation and learning.

The results from the studies in the thesis carry implications for theoretical models of early word segmentation and word learning. Previous research in word segmentation has particularly focused on the role of transitional probabilities and chunk frequency (e.g., French et al., 2011). The advantage of CLASSIC-UB over other models is that it captures the potential role of overlapping phonological sequences in word segmentation. Although this was not the central focus of earlier studies, various word segmentation models contemplate the role of overlapping sequences at different levels. For instance, in the chunking-based segmentation model, PARSER (Perruchet & Vinter, 1998), the inclusion of a memory interference parameter could allow researchers to examine how overlapping phonological sequences might either facilitate or disrupt the activation of phonological chunks. Similarly, the neural network segmentation model, TRACX (French et al., 2011), employs distributed representations of chunks spread across its hidden layers. These distributed representations potentially offer a highly flexible similarity gradient across chunks. This flexibility does not restrict the model to evaluating overlapping sequences for adjacent sequences, as seen in CLASSIC-UB or PARSER. Instead, it can go beyond by evaluating the similarity based on sequences with intervening elements. For instance, the model's responses to nonwords like "gaboti" and "kapodi" would be more closely aligned than when exposed to "gaboti" and "pudosa", even though both pairs do not share overlapping sequences of adjacent sounds. Evidence suggests that sequences with intervening elements play a role in infant word segmentation at 17 months of age and subsequent word learning (Frost et al., 2020; Monaghan et al., 2023). This indicates that it may also be important to

extend CLASSIC-UB's architecture to incorporate chunks that encode non-adjacent dependencies.

The role of overlapping sequences in vocabulary knowledge has also been highlighted in studies that consider their effect on nonword processing (e.g., Gathercole, 1995; Jones, 2016) as well as their cascading impact on child vocabulary growth (Jones et al., 2021). Overall, evidence from previous work and in this thesis indicates that comparing models that leverage overlapping phonological sequences could provide a deeper understanding of the learning mechanisms behind early word segmentation and word learning. Such comparisons would not only clarify theoretical foundations but also improve our understanding of how various input variables—like overlapping sequences, transitional probabilities, and frequent chunks—jointly influence the segmentation performance of infants and children.

5.2.4 Limitations and Future Research

There are several limitations to the work presented on the role of chunking in early naturalistic word segmentation, which could open different lines for future research. A primary limitation is that the studies did not explore the range of potential learning mechanisms proposed in earlier work. Models chosen for comparison with CLASSIC-UB were selected based on their capacity to shed light on the efficacy of different evaluation measures related to segmentation accuracy and child word learning. Larsen et al. (2017), who first highlighted the need for assessing developmental plausibility in segmentation models, demonstrated that transitional probability models and PUDDLE had contrasting strengths based on different performance metrics. While PUDDLE was the most accurate model for segmenting speech, transitional probability was best at capturing child age of word acquisition. Beyond this initial examination of the influence of evaluation metrics on conclusions about developmental plausibility, future research should examine a broader array of segmentation mechanisms. This encompasses alternative methods that view the segmentation task as Bayesian inference, applied to transitional probabilities with the model DiBS (Diphone-Based Segmentation; Daland & Pierrehumbert, 2011), or

on chunks (Adaptor Grammar; Goldwater et al., 2009). It also includes methods that employ a mix of strategies, such as monitoring transitional probability and chunk frequency concurrently (Swingley, 2005). Moreover, as previously mentioned, it would be valuable to consider other influential chunking models that incorporate the role of overlapping phonological sequences, like PARSER (Perruchet & Vinter, 1998) and TRACX (French et al., 2011). In sum, comparing a wide range of models is crucial to support the findings of the current studies on chunking's role in word segmentation and learning. This comparison would be important to also examine the hypothesis that overlapping phonological sequences play a role in early word segmentation beyond transitional probabilities and chunk frequencies. Recent efforts, such as the WordSeg package tool (Bernard et al., 2020), aim to facilitate the comparison of computational models of word segmentation. This open-source package could be expanded to incorporate evaluation metrics of developmental plausibility introduced in this thesis, establishing a cohesive developmental benchmark for researchers interested in early word segmentation.

The simulations presented in this thesis evaluated a parsimonious model of segmentation, CLASSIC-UB, which used an unconstrained mechanism of chunking combined with sensitivity to utterance boundaries to achieve segmentation of naturalistic speech. This parsimonious approach was employed to isolate the effects of chunking sequences in long-term memory and sensitivity to cues at different utterance boundaries in early segmentation. However, this choice involved making idealized assumptions about how learners process the input. Future work should explore how additional processing constraints influence the model's performance, potentially increasing the psychological plausibility of the model. For instance, once CLASSIC-UB learns a word as a complete chunk, subsequent encounters with that word do not influence the model's processing. However, we know that children become gradually faster at accessing familiar word representations (Fernald et al., 1998). Such a speed of processing constraint could be implemented in various ways: Either as a mechanism that affects the level of activation of a chunk in long-term memory (as in PUDDLE) or as a processing advantage that speeds up the time required to access phonological sequences from long-term memory during encoding. This latter concept aligns with how parent versions of CLASSIC-UB have

incorporated timing parameters in their architectures (e.g., Lloyd-Kelly et al., 2016) and is consistent with shorter looking times to familiar sequences in infant segmentation studies (e.g., Black & Bergmann, 2017). Moreover, different chunking architectures have examined the impact of short-term memory limitations on performance (e.g., Gobet & Lane, 2010; Lloyd-Kelly et al., 2016; Perruchet & Vinter, 1998). Investigating whether modelling cognitive limitations on short-term memory can explain performance variability beyond constraints on long-term memory retrieval would be worthwhile.

Another limitation of the current studies is their focus on model performance at the group level. The child-directed speech used aggregated input from different target children. Similarly, the evaluation measures were computed across target children (e.g., the word frequency distribution of the productions across all children). This approach was chosen to maximize sample size, compensating for the phenomenon where a reduced sample size dramatically decreases the number of low-frequency words, a consequence of the Zipfian properties of speech. This methodological choice, however, restricts the scope of investigations. For instance, it prevents researchers from modelling how vocabularies grow over time and from examining the associated individual differences. Such a fine-grained analysis would offer a richer understanding of the plausibility of the models and provide insights into variables that may predict language delays (e.g., Fernald & Marchman, 2012).

In Chapter 2, I discussed how a limited word sample size might influence the study's conclusions, as certain mechanisms, like transitional probability, are favoured by training on less skewed frequency distributions (e.g., Kurumada et al., 2013). Notably, in Chapter 3, I replicated the first study using a considerably smaller sample size. This smaller sample size did not dramatically influence the performance of transitional probability models. The only significant change was observed when modelling the child's age of first production (Table 3), where forward transitional probability outperformed PUDDLE. This deviation warrants further investigation, as it might result from transitional probability benefitting from a reduced presence of low-frequency words in the smaller Italian sample. Given the advantages of using cross-linguistic data in evaluating models of early word segmentation, future research

would significantly benefit from a focus on constructing large-scale corpora derived from children's naturalistic conversations. For instance, several previous simulations of word segmentation have employed a target sample size of a minimum of 10,000 child-directed speech utterances (e.g., Caines et al., 2019; Christiansen et al., 1998; French et al., 2011; Goldwater et al., 2009; Monaghan & Christiansen, 2010). However, setting such a sample size threshold decreases the number of languages that can be investigated in the CHILDES database from 44 to 15 (e.g., Jessop et al., 2023), presenting a substantial obstacle to conducting cross-linguistic studies.

Aside from the outcome related to age of first production, the conclusions derived from the smaller sample's analysis remained consistent. Results from developmental measures were replicated even with a limited sample. While the robustness of these measures needs further verification through a study centred on sample size variations, the findings are encouraging for moving beyond group-level analyses. For example, a recent study by Jessop et al. (2023) began investigating how a model of early naturalistic word segmentation might reflect vocabulary trajectories. They proposed a new model called CIPAL (Chunk-based Incremental Processing and Learning). This model is based on the same architecture used to build CLASSIC-UB but also incorporates various cognitive limitations as mentioned above (e.g., timing parameters that modulate access to representations in long-term memory and limited short-term memory capacity). They used a range of inputs from different languages, with a minimum of 10,000 utterances per sample. The model was exposed to input directed at individual children, and vocabulary acquisition was assessed every 50 utterances to obtain repeated measures of vocabulary growth. These measures were then compared to vocabulary growth curves derived from CDI estimates from 15 languages. They discovered that the model's growth curves were similar to those of the CDI estimates in terms of quadratic growth. Namely, the vocabulary growth of the model decelerated over time, much like the estimates for children. Moreover, visual comparisons between the model and CDI curves showed similar individual variabilities. The results of this recent study are promising for future investigations using CLASSIC-UB for several reasons. First, since CIPAL is closely related to CLASSIC-UB, I expect that the latter might also be effective in predicting vocabulary growth and individual differences. Second, because vocabulary

growth curves offer a more fine-grained analysis of model performance, they might address the sensitivity issues I found when modelling the age of acquisition from CDI scores in Chapters 2 and 3. This would be useful since comprehension and expressive skills, while related, are influenced by different input variables (Swingley & Humphrey, 2018). Moreover, comparing model performance in capturing both comprehension and production vocabularies would allow an estimation of how much of CLASSIC-UB's performance is due to its ability to simulate increased fluency in production (e.g., recall and articulation) on top of segmentation abilities.

Third, combining the approach I have taken in this thesis with the method used to evaluate CIPAL could significantly advance current knowledge. CIPAL has not been compared to competing models yet, so it remains unclear how it stands against other theories regarding the learning mechanisms involved in early segmentation. Additionally, the influence of different processing limitations within CIPAL has not been evaluated. An initial comparison between CLASSIC-UB and CIPAL could provide insights into how processing limitations (inherent in CIPAL) might influence vocabulary growth modelling. Moreover, modifications to CLASSIC-UB that gradually incorporate these assumptions could further elucidate which specific assumptions improve prediction.

Lastly, it is important to note that the conclusions derived from the modelling simulations in Chapters 2 and 3 are essentially proofs of concept and need further validation in future research. For instance, the modelling results have provided intriguing predictions that warrant exploration. One such prediction involves the interaction between type and token word frequencies and cues at utterance boundaries in improving segmentation accuracy and vocabulary learning. The prediction is that the role of utterance boundaries would differ in languages with different input characteristics. As an example, Dutch and Japanese caregivers tend to position new, unfamiliar words as one-word utterances rather than at the end of multiword utterances (Han et al., 2021). I expect that only word frequency in isolation would then predict the vocabulary size of Dutch and Japanese children. Instead, I expect no significant impact from the frequency of utterance-final words in multiword utterances.

Another prediction arises from the role of overlapping phonological sequences defined within the context of either neighbourhood density or phonotactic probability (as discussed in Chapter 3). The modelling suggests that, in English, neighbourhood density might be a more influential factor in predicting children's vocabularies than phonotactic probability. Conversely, the opposite might be true for Italian. Future research could explore this hypothesis by comparing the relative significance of these variables in predicting children's productive vocabularies in both languages. Examining these differences across languages is crucial to bolster the notion that infants and children form representations of varying grain size (e.g., Jessop et al., 2023; Jones et al., 2021). In other words, it would support the idea that a single underlying mechanism can produce both sublexical phonotactic probability and lexical neighbourhood density effects, even though these are typically viewed as resulting from distinct cognitive processes (e.g., Storkel, 2009).

5.3 The Role of Chunking and Analogy in Early Word Sense Disambiguation

5.3.1 Chapter 4. The Role of Verb-Event Structure in Children's Lexical Ambiguity Resolution

Chapter 4 examined the roles of domain-general cognitive processes of chunking (the ability to learn from associations) and analogy (the ability to generalize a known linguistic structure to an item not previously heard in that structure) in early word sense disambiguation. While young children often find word sense disambiguation challenging in experimental setups (e.g., Khanna & Boland, 2010), verb cues have been shown to facilitate their performance (Hahn et al., 2015; Rabagliati et al., 2013). However, it remains unclear which aspects of verb knowledge children draw upon in word sense disambiguation. One hypothesis suggests that children's processing primarily relies on rote-learned associations between verbs and specific objects (Snedeker & Yuan, 2008). Alternatively, it is possible that young children understand verb-event structures, which they leverage to comprehend ambiguous verb objects. This second hypothesis would suggest children's reliance on abstract

verb knowledge, consistent with theories that propose early knowledge of linguistic structures for sentence processing (Trueswell & Gleitman, 2007). A challenge in contrasting these two hypotheses is that word associations and verb-event structural information are often confounded in naturalistic speech.

Nevertheless, research has attempted to examine the unique contributions of verb associations and verb-event structures to early unambiguous word processing (Andreu et al., 2013; Mani et al., 2016). These studies have shown that young children can use both cues to predict upcoming objects in a sentence. Yet, it remains unclear whether these variables similarly influence the processing of ambiguous words. These studies also have some methodological limitations. Specifically, verb-object associations were defined through association norms or ratings sourced from adult participants. Such norms might not accurately reflect which associations are actually available to children in their linguistic environments. For this reason, the experimental stimuli might have not entirely controlled for verb-object associations (Mani et al., 2016), as items labelled as presenting weak associations between a verb and its object might still be fairly typical in child environments. Even when the researchers set out to disentangle the two variables (Andreu et al., 2013) - by pairing verbs with atypical but semantically appropriate objects (that is, with null word associations), atypicality was defined using expert academic judgment. Again, this definition might not capture what is (a)typical in child environments.

In Chapter 4, I introduced a new methodological approach to evaluate the influence of verb-object associations and verb-event structures on early word sense disambiguation. I defined these variables by examining naturalistic conversations involving children up to the age of 4, which corresponds to the age of the children tested in this study. I began by manually annotating all child-directed utterances from the English section of the CHILDES database (MacWhinney, 2000). From this annotation, I extracted verb-sense associations that enabled me to design an experimental condition that tested the unique role of these associations in disambiguation. This was possible by choosing verbs with neutral verb-event structures but strong verb-sense associations. Conversely, to examine the effect of verb-event structures, I selected verbs that children do not hear in association with

the experimental word senses, as observed in the annotated naturalistic conversations. Therefore, I expected that child participants would have to rely on their knowledge of verb-event structures when using those verbs for disambiguation.

The study revealed that 4-year-old English-speaking children could use both verb-sense associations and verb-event structures to disambiguate word meanings. This supports the idea that children employ both rote-learned associations and abstract linguistic knowledge in sentence parsing (Trueswell & Gleitman, 2007). It also aligns with usage-based theories of language acquisition, which posit that both chunking and analogy play key roles in early development (Abbot-Smith & Tomasello, 2006; Bybee, 2010a; Ibbotson et al., 2012).

Further exploratory analyses produced findings consistent with previous literature and the above conclusions. First, children's performance was influenced by the frequency of word senses in naturalistic speech. This aligns with prior research highlighting children's sensitivity to the frequency dominance of certain word senses over others (e.g., Booth et al., 2006; Rabagliati et al., 2013; Simpson & Foster, 1986). Second, verb-object associations derived from naturalistic speech were predictive of children's performance, emphasizing the significance of using child-directed naturalistic input when constructing experimental stimuli. Third, children's verb knowledge, as assessed from parent-report questionnaires, was predictive of child performance at the disambiguation task. This suggests that children drew upon their understanding of verb-event structures during disambiguation.

5.3.2 Implications for the Study of Child Lexical Ambiguity Resolution

Chapter 4 highlighted the value of leveraging large corpora of child-directed speech to carefully examine the role of naturalistic variables in child word sense processing. While previous experimental evidence points to children's difficulties in word sense disambiguation, observations from naturalistic interactions suggest that children grasp lexical ambiguities in their speech from a very young age (Meylan et al., 2021). The findings in Chapter 4 highlight the importance of considering the input children typically receive. Doing so can help identify variables that may contribute to

this early proficiency. The recently released annotated corpus, ChiSense-12, offers a promising avenue for conducting corpus analyses to pinpoint these variables within the situational contexts children encounter.

It is also worth mentioning that the forthcoming release of the annotated corpus by Meylan et al. (2021) will offer a more extensive sample size of target ambiguous words. This will enable researchers to determine if the results from the current study can be generalized to a broader set of ambiguous words that children learn early on. This expanded sample size will also allow for the investigation of variables not covered by ChiSense-12, such as the influence of different sense categories (e.g., homophony, polysemy) or the role of a target word sense's syntactic category in enhancing word sense processing (e.g., Dautriche et al., 2018). Furthermore, the corpus from Meylan et al. will include words included in the Communicative Development Inventory (Fenson et al., 2007). This means that researchers will be able to examine the significance of word sense knowledge in early word learning, an area which has received limited attention. For example, a crucial question to explore is whether vocabulary growth is underestimated when considering the gradual enrichment of sense representations. Alternatively, could examining sense distributions over developmental stages capture significant variance in language delays (e.g., Norbury, 2005)? Ultimately, leveraging naturalistic conversations holds promise for deepening our insights into early word processing and learning.

The findings in Chapter 4 align with a usage-based perspective on language development (Tomasello, 2000, 2003, 2009) which assumes that children's word representations are shaped by learning mechanisms of chunking and analogy. These mechanisms allow for the integration of abstract linguistic structures, like verb-event knowledge, while also preserving idiosyncratic characteristics drawn from the learner’s experiences with language, such as verb-object associations. Furthermore, the current findings support various accounts of sentence parsing and word processing. First, they highlight the idea that some top-down abstract information sources might be more beneficial than others (Trueswell & Gleitman, 2007). For example, 4-year-old children may be proficient in using their semantic knowledge

about verb usage. However, they might find it challenging to integrate global discourse plausibility (Khanna & Boland, 2010; Rabagliati et al., 2013) since it requires several aspects of pragmatic, syntactic, and semantic knowledge.

Second, the current findings support the recent account of child word learning proposed by Srinivasan and Rabagliati (2021) and the recent semantic settling account of word processing by Rodd (2020). These findings indicate that children integrate lexical ambiguity into their vocabularies, and that their representations of word meanings incorporate both bottom-up and top-down contextual aspects. This integration should be considered in research aiming to understand child word learning, especially as we move away from the notion that child word forms map onto single meanings (e.g., Trueswell et al., 2013).

5.3.3 Limitations and Future Research

In Chapter 4, I concluded that linguistic experience influences a child's ability to disambiguate word meanings by tracking word associations and generalizing known verb-event structures. This conclusion was drawn from the significant role that verb-object co-occurrences in naturalistic speech and parent-reported verb knowledge played in predicting children's performance. I discussed the possibility that these variables, which relate to a child's language experience, might also account for the performance differences between child and adult age groups. In other words, the better performance of adults might be attributed to their more extensive language experience. However, this conclusion requires further exploration for various reasons. Firstly, one should study language experience more effectively by examining performance across different child age groups or using longitudinal designs that assess the relationship between language learning trajectories and improvements in word sense disambiguation tasks.

Second, I discussed how word sense disambiguation tasks often tap into language knowledge at various levels. For instance, in the task used in the current study, competition was introduced between different sentences within a discourse. This likely involved the use of semantic knowledge at the local sentence level (via

the manipulated verb cues) but could also potentially extend to pragmatic knowledge which allows individuals to smoothly connect sentences into a coherent narrative (this might have contributed to the qualitative differences observed between adult and child performance in the verb-lexical condition). Therefore, to better assess language experience, future research should also consider pragmatic knowledge. For instance, additionally using tests that assess children's understanding of causal relationships, event sequences, and social norms (e.g., Khan et al., 2016) would provide a richer insight into the type of knowledge children and adults use to parse discourse and perform sense disambiguation.

Third, it is possible that children's limited processing skills could account for the age differences observed in the current study. Prior research has highlighted the role of processing skills in both syntactic (e.g., Kidd & Bavin, 2005; Yacovone et al., 2021) and word ambiguity resolution (Khanna & Boland, 2010). This interest is driven by the possibility that limited processing skills might obscure children's true abilities. Consequently, examining these processing limitations could also be interesting to reveal children's understanding of abstract linguistic structures at an earlier age. In summary, a more comprehensive, concomitant evaluation of language and processing skills across developmental stages will help researchers better understand the variables and underlying processes that determine differences between children and adults in word sense disambiguation.

Finally, it is important to address the limitations concerning conclusions on the type of abstract knowledge involved in children's use of verb-event structures. While I have concluded that the use of these structures suggests the involvement of usage-based analogical learning mechanisms, the evidence does not rule out alternative mechanisms proposed by nativist approaches. This limitation stems from the fact that Chapter 4's primary aim was not to delve deep into the mechanisms determining the use of verb-event structures, but rather to examine experimental evidence supporting the idea that both chunking and analogy might be *independently* involved in early word sense disambiguation. To reiterate, the significant role of chunking was inferred from the observation that children can resolve lexical ambiguities solely based on verb-object associations. This conclusion

was reached while controlling for any verb semantics by using verbs that can accept both target word senses as plausible arguments (e.g., "She saw the [animal/food] chicken"). The role of analogy was instead highlighted by the fact that, when controlling for verb-object associations (i.e., selecting verbs that do not co-occur with either target sense in natural conversations), children still managed to resolve lexical ambiguities based on their understanding of verb semantics.

Moving forward, it is essential to delve into what exactly chunking entails and the specifics of this mechanism, similarly to what discussed in Chapters 2 and 3. For instance, does the suggested learning mechanism store verb-object chunks by focusing on the transitional probability between word pairs, as suggested by McCauley and Christiansen (2019)? Or does it leverage overlapping lexical and multiword phonological sequences, as proposed by Jones et al. (2020)? Although this question warrants further exploration, I do not believe it presents significant challenges to a usage-based approach. This is primarily because it is hard to envision a mechanism determining sensitivity to verb-object associations without fundamentally emphasizing language usage as the primary driver. However, the type of mechanism that determines sensitivity to verb-event structures is more controversial.

One possibility—which aligns with the usage-based approach I have employed throughout this thesis—suggests that learners gradually form expectations about the types of object arguments a verb can accept. This is achieved by generalizing from the characteristics of known words that have previously occupied the object argument slot in their linguistic experiences. Central to this explanation is the concept of analogy (Bybee, 2010a). However, there is an alternative perspective: That a child's innate knowledge of verbs' argument structures might include constraints on the kinds of words considered semantically plausible in the verb object slot (e.g., Gleitman & Gillette, 1995; Pinker, 1994a). This account does not provide specific indications on exactly which semantic aspects of verb arguments might be innately specified and which not. The semantic structure of verb syntactic frames might be specified at different degrees of detail (e.g., Copestake & Briscoe, 1992; Fodor & Katz, 1964; Jackendoff, 1985; Pustejovsky, 1995). For example, it

could go beyond merely defining "bat" as a patient of "swing", specifying that "bat" is a "physical object" (e.g., Fodor & Katz, 1964) or even specifying more specific attributes like "liquid" to the object argument of "drink" (e.g., Jackendoff, 1985). Despite its under specification, this perspective posits that children would not (at least not entirely) draw analogies from their past linguistic encounters; instead, they would apply predefined rules that enable the verb-argument structure to be generative, meaning they can extend it to new instances of object arguments.

To compare these two hypotheses, one approach would be to test children at a younger age, at which point they likely have not had enough language exposure to estimate the plausibility of object arguments. This would help determine whether children possess any inherent bias towards plausibility. Another approach, which I employed for Chapters 2 and 3, involves using a computational modelling approach. This would aim to demonstrate, as proof of principle, that language input alone is sufficient for using verb-event structures in word sense disambiguation.

There is indeed computational evidence suggesting how analogical learning mechanisms might operate on verb-event structures. Alishahi and Stevenson (2007) proposed a Bayesian computational model that learns verb-event structures by gradually generalizing over the semantic characteristics of a verb's syntactic arguments encountered during training. This learning is facilitated by a probabilistic mechanism that initially records the associative relation between a verb and its semantic features, the positions of the syntactic arguments, and the semantic characteristics of the items that incrementally fill those arguments. The model draws analogies across similar sentences, forming verb construction representations, which then allow it to generalize arguments to new items. Through incremental probabilistic learning, the model effectively simulated adult plausibility judgments regarding verb-event preferences (e.g., “The mechanic warned the driver” versus “The mechanic warned the engine”). In principle, the simulations by Alishahi and Stevenson (2007) demonstrate that usage-based mechanisms can develop verb-event representations and assist in processing novel arguments. This suggests that a usage-based approach might also be valuable in understanding how verb constructions help children disambiguate words with multiple meanings. However, it

is also crucial to note that the simulations assumed extensive knowledge on the learner's part, including a comprehensive understanding of both verb and argument semantics (sourced from dictionaries), ideal identification of syntactic arguments, their semantics, and their positions relative to the main verb.

In recent research, I began exploring how a usage-based learner might develop sense-specific representations and how these representations might be shaped by the sentence context, including verb information (Cabiddu et al., 2023). I evaluated a large group of models ($N = 45$) based on the Transformer neural architecture (Vaswani et al., 2017). These models perform sense disambiguation by leveraging sentence context to produce high-dimensional, contextualized representations, an approach consistent with Rodd's model (2020) and other usage-based accounts (e.g., Ambridge, 2020). Transformers can be considered usage-based learners. They retain a vast amount of context-dependent data from language exemplars while also gradually encoding context-independent information across different linguistic levels. Notably, Transformers have been shown to be sensitive to both syntactic and semantic sentence structures (e.g., Jawahar et al., 2019; Tenney et al., 2019), meaning that the models' layers contain information from which labels for various aspects, including syntactic categories, constituents, semantic roles, and coreference, among others, can be predicted. This sensitivity enabled me to apply these models directly to raw, naturalistic language without requiring additional external resources to implement sensitivity to such structures. By using the models on raw, unlabelled input, I could study how sense disambiguation could be achieved using representations drawn from naturalistic child-directed speech. Namely, the models' word sense representations were computed based on sets of utterances directed at children.

I found that Transformers could approximate findings observed for 4-year-old children as in Rabagliati et al. (2013) as well as in the current experiment of my thesis. Transformers showed the ability to use both global and local sentence contexts to disambiguate word meanings in coherent tasks. These are tasks where all cues, including word associations and event structures, point toward the intended meaning, as in "Oscar was at the beach. He caught a fish, which was exciting".

Furthermore, they also exhibited a degree of success in capturing how children could use verb-event structures to disambiguate word meanings ("Sophia twisted a [music/elastic] band"). This supports the idea that the semantic restrictions verbs impose on their arguments could be learned using some form of distributional mechanism operating on linguistic events.

However, it is also worth noting that the simulations also revealed significant challenges for Transformers. Many models within the tested pool struggled to match the performance levels of children on contrastive tasks. These tasks, such as the one presented in Chapter 4 of this thesis, require resolving the competition between word associations from prior contexts and event structures from local contexts. For instance, in my experiment, consider the sentence "Sophia listened to some music. Then, she twisted the band". The word associations from the prior context (like "listen" and "music") point toward "music band", while the verb-event structure from the local context ("twist") indicates "elastic band". Similarly, in Rabagliati et al.'s (2013) experiment, the sentence "Kermit was in a dark cave. He was nervous about the animals, so he carried a big bat" presents a competition between word associations pointing to an "animal bat" and global plausibility directing towards a "baseball bat". My ongoing research aims to explore whether these limitations in Transformers could be attributed to their lack of real-world knowledge, as they can only derive word and sentence semantics from textual input, therefore struggling to make pragmatic inferences based on real-world knowledge such as "Given that Kermit was in a dark cave and was nervous about animals, it makes sense that he would carry a bat (e.g., heavy stick) for protection, rather than a flying mammal".

This research would allow to consider how far one can push a usage-based approach and whether one would need to integrate domain-specific constraints consistent with nativist approaches (e.g., Pinker, 1989; Thornton, 2012) or domain-general innate biases (e.g., Perfors et al., 2011) to reach model developmental plausibility. In summary, it remains unclear whether Transformers generate sense predictions in a manner similar to children or adults. However, they could serve as valuable tools to test the claim that word sense processing can be approximated by

learning mechanisms that do not need the implementation of innate verb knowledge.

5.4 Conclusion

This thesis emphasized the importance of studying how naturalistic language experiences influence word learning and processing outcomes, when one's goal is to test the idea that children might be usage-based learners. This work offered three key contributions: (a) New methods were introduced to investigate early word segmentation and early word sense processing using naturalistic corpora of conversations. These methods are suitable for both computational and behavioural studies examining a usage-based perspective on infants and children's word learning and processing. (b) This research underscored the significant role of chunking as a learning mechanism. Infants and children potentially employ this mechanism to extract words from naturalistic speech and form their initial production vocabularies. (c) Chunking and analogy, as independent learning mechanisms, might support children's early processing of lexical ambiguities. This was deduced by observing young children's sensitivity to both bottom-up and top-down sentence cues to word meaning. In conclusion, this thesis holds significant implications on both theoretical and practical fronts. It enhances our understanding of how infants and children learn language in real-life contexts and may pave the way for a more profound comprehension of the sources of individual differences and learning difficulties.

Notes

1 For ease of exposition, the example uses IPA phonetic transcription. However, in our simulations, we used a transcription based on the *CMU Pronouncing Dictionary* (Lenzo, 2007; see an example in Figure 1).

2 However, CLASSIC's encoding does not allow partial activation of chunks unlike in Baayen et al.'s (2011) study.

3 Interestingly, when Larsen et al.'s (2017) measure was used, transitional probability models performed better than chunking models despite their discovering fewer words in the input as we mentioned above. For example, a transitional probability model explained 19% of variance in age of acquisition (the highest performance in the study), while the chunking model PUDDLE explained only 7% (Larsen et al., 2017).

4 The CDI words and gestures includes 373 phonological words (not considering homophone duplicates) typically acquired by infants between 8 and 18 months of age. Our final sample contained 330 words after filtering for those CDI words present in the child-directed input that the segmentation models received (i.e., CDI words that the models had the opportunity to learn).

5 A discussion about the effect of sample size reduction when using the age of acquisition measure from the CDI can be found in the file CDI_addendum at the project's OSF page (https://doi.org/10.17605/osf.io/kbnep).

6 Adjusted R^2 estimates cannot typically be directly compared to R^2 estimates. However, because of our large sample size, adjusted R^2 and R^2 estimates and confidence intervals were identical, allowing us to compare our adjusted R^2 estimates to Larsen et al.'s (2017) R^2 estimates. In fact, as sample size increases expected R^2 estimates become less biased and approach adjusted R^2 unbiased estimates of the population explained variance (Karch, 2020).

7 Italian and English child vocabularies become comparable by word-level measures only when considering similar amounts of input utterances and child word types. Otherwise, the measures can be biased by sample size. For instance, the word "dog"

might occur less frequently in Italian merely because its frequency is counted in a limited number of Italian utterances compared to English. Similarly, using two different child sample sizes for word types (Italian = 1,653; English = 5,480) can skew the comparison. Larger word type samples are more likely to contain low-frequency or long words. To correct for these biases, we sampled the same number of input utterances from English child-directed speech used in Cabiddu et al. (2023) as we had for Italian (N = 22,190). We then used this reduced input to recompute the word-level measures for English children's word types (weighted log10 frequency, weighted neighbourhood density, weighted phonotactic probability). We also downsampled the English child set of word types from 5,480 to 1,653 word types to match Italian children's sample size. It is important to note that choosing different random samples for input utterances or child word types did not significantly alter the means and standard deviations reported in the text. Additionally, we obtained similar results when comparing Italian word type distributions to the full raw English child word type distributions computed from the full English input.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*(3). https://doi.org/10.1515/TLR.2006.011

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*, *164*, 116–143. https://doi.org/10.1016/j.cognition.2017.02.009

Alcock, K. (2020). *The UK communicative development inventory database: Words and gestures ages 8–18 months 2012-2016*. [Data collection]. UK Data Service. http://doi.org/10.5255/UKDA-SN-853687

Alishahi, A., & Stevenson, S. (2007). A Cognitive Model for the Representation and Acquisition of Verb Selectional Preferences. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, 41–48. https://aclanthology.org/W07-0606

Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834. https://doi.org/10.1080/03640210801929287

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition: *First Language*. *40*(5-6), 509-559. https://doi.org/10.1177/0142723719869731

Ambridge, B. (2020). Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators. *First Language*, *40*(5–6), 640–659. https://doi.org/10.1177/0142723720949723

Ambridge, B., & Lieven, E. (2015). A Constructivist Account of Child Language Acquisition. In *The Handbook of Language Emergence* (pp. 478–510). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118346136.ch22

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273. https://doi.org/10.1017/S030500091400049X

Andreu, L., Sanz-Torrent, M., & Trueswell, J. C. (2013). Anticipatory sentence processing in children with specific language impairment: Evidence from eye movements during listening. *Applied Psycholinguistics*, *34*(1), 5–44. https://doi.org/10.1017/S0142716411000592

Antelmi, D., & Morlacchi, A. (2005). L'interpretazione del linguaggio figurato nel ritardo mentale. *L'interpretazione Del Linguaggio Figurato Nel Ritardo Mentale*, 1000–1025. https://doi.org/10.1400/57839

Arduino, L., & Burani, C. (2004). Neighborhood Effects on Nonword Visual Processing in a Language with Shallow Orthography. *Journal of Psycholinguistic Research*, *33*, 75–95. https://doi.org/10.1023/B:JOPR.0000010515.58435.68

Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language*, *48*(5), 937–958. https://doi.org/10.1017/S0305000921000386

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in fluent maternal speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Lawrence Erlbaum. https://doi.org/10.4324/9781315806822

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*, *2019*, Article e4895891. https://doi.org/10.1155/2019/4895891

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438–481. https://doi.org/10.1037/a0023851

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. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

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., 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. https://doi.org/10.1017/S0305000900008680

Batsuren, K., Bella, G., & Giunchiglia, F. (2021). MorphyNet: A Large Multilingual Database of Derivational and Inflectional Morphology. *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 39–48. https://doi.org/10.18653/v1/2021.sigmorphon-1.5

Batsuren, K., Bella, G., Arora, A., Martinovic, V., Gorman, K., Žabokrtský, Z., Ganbold, A., Dohnalová, Š., Ševčíková, M., Pelegrinová, K., Giunchiglia, F., Cotterell, R., & Vylomova, E. (2022). The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 103–116. https://doi.org/10.18653/v1/2022.sigmorphon-1.11

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics*, *47*(2), 383–411. https://doi.org/10.1515/LING.2009.014

Behrens, H. (2021). Constructivist Approaches to First Language Acquisition. *Journal of Child Language*, *48*(5), 959–983. https://doi.org/10.1017/S0305000921000556

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, *4*, 247–260. https://doi.org/10.1016/S0163-6383(81)80027-6

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology. General*, *117*(1), 21–33. https://doi.org/10.1037//0096-3445.117.1.21

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

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

Booth, J. R., Harasaki, Y., & Burman, D. D. (2006). Development of Lexical and Sentence Level Context Effects for Dominant and Subordinate Word Meanings of Homonyms. *Journal of Psycholinguistic Research*, *35*(6), 531–554. https://doi.org/10.1007/s10936-006-9028-5

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation.

Psychological science, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Bracco, G., Calderone, B., & Celata, C. (2015). Phonotactic probabilities in Italian simplex and complex words: A fragment priming study. In V. Pirrelli, C. Marzi, & M. Ferro (Eds.), *Proceedings of the NetWordS Final Conference on Word Knowledge and Word Usage: Representations and Processes in the Mental Lexicon* (extended abstract) (pp. 24-28). Pisa, Italy.

Braginsky, M., Sanchez, A., & Yurovsky, D. (2019). *childesr: Accessing the 'CHILDES' database* (0.1.2). https://github.com/langcog/childesr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind : Discoveries in Cognitive Science*, *3*, 52–67. https://doi.org/10.1162/opmi_a_00026

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

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

Bybee, J. (2001). *Phonology and language use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Bybee, J. (2010b). *Language, Usage and Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Bybee, J. (Ed.). (2010a). Analogy and similarity. In *Language, Usage and Cognition* (pp. 57–75). Cambridge University Press. https://doi.org/10.1017/CBO9780511750526.004

Cabiddu, F., Bott, L., Jones, G., & Gambi, C. (2023). CLASSIC Utterance Boundary: A Chunking-Based Model of Early Naturalistic Word Segmentation. *Language Learning*, *73*(3), 942–975. https://doi.org/10.1111/lang.12559

Cabiddu, F., Nikolaus, M., & Fourtassi, A. (2023). *Comparing children and large language models in word sense disambiguation: Insights and challenges.* [Manuscript submitted for publication]. School of psychology, Cardiff University.

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

Caselli, M. C., Rinaldi, P., Stefanini, S., & Volterra, V. (2012). Early action and gesture 'vocabulary' and its relation with word comprehension and production. *Child Development*, *83*(2), 526–542. https://doi.org/10.1111/j.1467-8624.2011.01727.x

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*(2), 97–138. https://doi.org/10.1017/S0022226700001134

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X1500031X

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

Christie, S., & Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by Structural Alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Christophe, A., Guasti, T., & Nespor, M. (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

Chuang, Y.-Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2021). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear

discriminative learning. *Behavior Research Methods*, *53*(3), 945–976. https://doi.org/10.3758/s13428-020-01356-w

Cinque, G. (1993). A Null Theory of Phrase and Compound Stress. *Linguistic Inquiry*, *24*(2), 239–297.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235–253. https://doi.org/10.1037//0096-3445.120.3.235

Colbert-Getz, J., & Cook, A. E. (2013). Revisiting effects of contextual strength on the subordinate bias effect: Evidence from eye movements. *Memory & Cognition*, *41*(8), 1172–1184. https://doi.org/10.3758/s13421-013-0328-3

Connell, L., & Keane, M. T. (2004). What Plausibly Affects Plausibility? Concept Coherence and Distributional Word Coherence as Factors Influencing Plausibility Judgments. *Memory & Cognition*, *32*(2), 185–197. https://doi.org/10.3758/BF03196851

Copestake, A., & Briscoe, T. (1992). Lexical operations in a unification-based framework. In J. Pustejovsky & S. Bergler (Eds.), *Lexical Semantics and Knowledge Representation* (pp. 101–119). Springer. https://doi.org/10.1007/3-540-55801-2_30

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, *39*(3), 388–402. https://doi.org/10.1016/j.wocn.2011.02.004

Cristià, A., Seidl, A., & Gerken, L. (2011). Learning classes of sounds in infancy. *University of Pennsylvania Working Papers in Linguistics*, *17*(1), Article 9. https://repository.upenn.edu/pwpl/vol17/iss1/9

Cruttenden, A. (1986). *Intonation*. Cambridge University Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press. https://doi.org/10.7551/mitpress/9012.001.0001

D'Odorico, L., & Carubbi, S. (2003). Prosodic Characteristics of Early Multi-Word Utterances in Italian Children. *First Language*, *23*(1), 97–116. https://doi.org/10.1177/0142723703023001005

Dal Ben, R., Souza, D. de H., & Hay, J. F. (2021). When statistics collide: The use of transitional and phonotactic probability cues to word boundaries. *Memory & Cognition*, *49*(7), 1300–1310. https://doi.org/10.3758/s13421-021-01163-4

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. https://doi.org/10.1111/j.1551-6709.2010.01160.x

Dautriche, I., Fibla, L., Fievet, A.-C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, *104*, 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

de Carvalho, A., He, A. X., Lidz, J., & Christophe, A. (2019). Prosody and Function Words Cue the Acquisition of Word Meanings in 18-Month-Old Infants. *Psychological Science*, *30*(3), 319–332. https://doi.org/10.1177/0956797618814131

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). American Psychological Association. https://doi.org/10.1037/10459-002

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446. https://doi.org/10.1016/0749-596X(88)90066-6

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*(4), 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. https://doi.org/10.1016/j.cogpsych.2016.11.004

Estes, K. G., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Fais, L., Werker, J., F., Cass, B., Leibowich, J., Barbosa, A., V., & Vatikiotis-Bateson, E. (2012). Here's looking at you, baby: What gaze and movement reveal about minimal pair word-object association at 14 months. *Laboratory Phonology*, *3*(1), 91–124. https://doi.org/10.1515/lp-2012-0007

Feldman, N. H., Goldwater, S., Dupoux, E., & Schatz, T. (2021). Do infants really learn phonetic categories? *Open Mind*, *5*, 113–131. https://doi.org/10.1162/opmi_a_00046

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 10.1037/a0034245. https://doi.org/10.1037/a0034245

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual.* MD: Paul H. Brookes Publishing Co.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222. https://doi.org/10.1111/j.1467-8624.2011.01692.x

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid Gains in Speed of Verbal Processing by Infants in the 2nd Year.

Psychological Science, *9*(3), 228–231. https://doi.org/10.1111/1467-9280.00044

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2010). Categorization in 3- and 4-Month-Old Infants: An Advantage of Words Over Tones. *Child Development*, *81*(2), 472–479. https://doi.org/10.1111/j.1467-8624.2009.01408.x

Ferry, A., Nespor, M., & Mehler, J. (2020). Twelve to 24-month-olds can understand the meaning of morphological regularities in their language. *Developmental Psychology*, *56*, 40–52. https://doi.org/10.1037/dev0000845

Fibla, L., Sebastian-Galles, N., & Cristia, A. (2022). Is there a bilingual disadvantage for word segmentation? A computational modeling approach. *Journal of Child Language*, *49*(6), 1119–1146. https://doi.org/10.1017/S0305000921000568

Floyd, S., & Goldberg, A. E. (2021). Children make use of relationships across meanings in word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*, 29–44. https://doi.org/10.1037/xlm0000821

Floyd, S., Goldberg, A. E., & Lew-Williams, C. (2020). Toddlers recognize multiple meanings of polysemous words. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society,* 2799-2804.

Fodor, J. A., & Katz, J. J. (1964). *The structure of language: Readings in the philosophy of language*. Englewood Cliffs, N.J. : Prentice-Hall.

Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of the 18th conference on computational language learning* (pp. 191–200). https://doi.org/10.3115/v1/W14-1620

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Why is English so easy to segment? *Proceedings of the Fourth Annual Workshop on Cognitive*

Modeling and Computational Linguistics (CMCL), 1–10. https://aclanthology.org/W13-2601

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, 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

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614–636. https://doi.org/10.1037/a0025255

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, *3*, 111–127. https://doi.org/10.1111/j.1749-818X.2008.00104.x

Frost, R. L. A., Jessop, A., Durrant, S., Peter, M. S., Bidgood, A., Pine, J. M., Rowland, C. F., & Monaghan, P. (2020). Non-adjacent dependency learning in infancy, and its link to language development. *Cognitive Psychology*, *120*, 101291. https://doi.org/10.1016/j.cogpsych.2020.101291

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty* [Unpublished manuscript]. Department of Linguistics, University of Pennsylvania.

Gaskell, M. G., Cairney, S. A., & Rodd, J. M. (2019). Contextual priming of word meanings is stabilized over sleep. *Cognition*, *182*, 109–126. https://doi.org/10.1016/j.cognition.2018.09.007

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*, 83–94. https://doi.org/10.3758/BF03210559

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development*, Vol. 2: Language, thought and culture (pp. 301–334). Lawrence Erlbaum.

Gentner, D. (2003). Why we're so smart. In *Language in mind: Advances in the study of language and thought* (pp. 195–235). Boston Review. https://doi.org/10.7551/mitpress/4117.001.0001

Gentner, D., & Namy, L. L. (2006). Analogical Processes in Language Learning. *Current Directions in Psychological Science*, *15*(6), 297–301. https://doi.org/10.1111/j.1467-8721.2006.00456.x

Gertner, Y., & Fisher, C. (2012). Predicted errors in children's early sentence comprehension. *Cognition*, *124*(1), 85–94. https://doi.org/10.1016/j.cognition.2012.03.010

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

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

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing Lexical Semantic Change with Contextualised Word Representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. https://doi.org/10.18653/v1/2020.acl-main.365

Gleitman, L. R., & Gillette, J. (1995). The Role of Syntax in Verb Learning. In *The Handbook of Child Language* (pp. 413–427). John Wiley & Sons, Ltd. https://doi.org/10.1111/b.9780631203124.1996.00017.x

Gobet, F. (2017). Entrenchment, Gestalt formation, and chunking. In *Entrenchment and the psychology of language learning: How we reorganize and adapt*

linguistic knowledge (pp. 245–267). De Gruyter Mouton. https://doi.org/10.1037/15969-012

Gobet, F., & Lane, P. (2010). *The CHREST Architecture of Cognition: The Role of Perception in General Intelligence*. 88–93. https://doi.org/10.2991/agi.2010.20

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*, 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4.

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

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 673–680). https://doi.org/10.3115/1220175.1220260

Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. *Journal of Child Language*, *22*(3), 703–726. https://doi.org/10.1017/s0305000900010011

Goslin, J., Galluzzi, C., & Romani, C. (2014). PhonItalia: A phonological lexicon for Italian. *Behavior Research Methods*, *46*(3), 872–886. https://doi.org/10.3758/s13428-013-0400-8

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

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, *8*, Article 555. https://doi.org/10.3389/fpsyg.2017.00555

Hahn, N., Snedeker, J., & Rabagliati, H. (2015). Rapid Linguistic Ambiguity Resolution in Young Children with Autism Spectrum Disorder: Eye Tracking Evidence for the Limits of Weak Central Coherence. *Autism Research*, *8*(6), 717–726. https://doi.org/10.1002/aur.1487

Han, M., de Jong, N. H., & Kager, R. (2021). Language specificity of infant-directed speech: Speaking rate and word position in word-learning contexts. *Language Learning and Development, 17*(3), 221–240. https://doi.org/10.1080/15475441.2020.1855182

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. https://CRAN.R-project.org/package=DHARMa

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106. https://doi.org/10.1016/j.cogpsych.2011.06.002

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. *Cognition*, *189*, 11–22. https://doi.org/10.1016/j.cognition.2019.03.005

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K. S., & Suma, K. (2015). The Contribution of Early Communication Quality to Low-Income Children's Language Success. *Psychological Science*, *26*(7), 1071–1083. https://doi.org/10.1177/0956797615581493

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

Hoff, E., Core, C., & Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-month-olds. *Journal of Child Language*, *35*(4), 903–916. https://doi.org/10.1017/S0305000908008751

Hohne, E. A., & Jusczyk, P. W. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, *56*(6), 613–623. https://doi.org/10.3758/BF03208355

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 624–646. https://doi.org/10.18653/v1/2021.conll-1.49

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

Ibbotson, P., & Tomasello, M. (2009). Prototype constructions in early language acquisition. *Language and Cognition*, *1*(1), 59–85. https://doi.org/10.1515/LANGCOG.2009.004

Ibbotson, P., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2012). Semantics of the transitive construction: Prototype effects and developmental comparisons. *Cognitive Science*, *36*(7), 1268–1288. https://doi.org/10.1111/j.1551-6709.2012.01249.x

Jackendoff, R. (1985). Semantics and Cognition. *Linguistics and Philosophy*, *8*(4), 505–519.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jessop, A., Pine, J., & Gobet, F. (2023). *Chunk-based Incremental Processing and Learning: An integrated theory of word discovery, vocabulary growth, and speed of lexical processing*. PsyArXiv. https://doi.org/10.31234/osf.io/dukpt

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, M. (2008). Unsupervised Word Segmentation for Sesotho Using Adaptor Grammars. *Proceedings of the Tenth Meeting of ACL Special Interest Group on*

Computational Morphology and Phonology, 20–27. https://aclanthology.org/W08-0704

Johnson, M., Christophe, A., Dupoux, E., & Demuth, K. (2014). Modelling function words improves unsupervised word segmentation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 282–292. https://doi.org/10.3115/v1/P14-1027

Jones, G. (2012). Why Chunking Should be Considered as an Explanation for Developmental Change before Short-Term Memory Capacity and Processing Speed. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00167

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition*, *153*, 79–88. https://doi.org/10.1016/j.cognition.2016.04.017

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

Jones, G., Cabiddu, F., & Avila-Varela, D. S. (2020). Two-year-old children's processing of two-word sequences occurring 19 or more times per million and their influence on subsequent word learning. *Journal of Experimental Child Psychology*, *199*, Article 104922. https://doi.org/10.1016/j.jecp.2020.104922

Jones, G., Cabiddu, F., Andrews, M., Rowland, C. (2021). Chunks of phonological knowledge play a significant role in children's word learning and explain effects of neighborhood size, phonotactic probability, word frequency and word length. *Journal of Memory and Language*, *119*, Article 104232. https://doi.org/10.1016/j.jml.2021.104232

Jones, G., Cabiddu, F., Barrett, D. J. K., Castro, A., & Lee, B. (2023). How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies. *First Language*, *43*(3), 253–282. https://doi.org/10.1177/01427237221150070

Jones, G., Justice, L. V., Cabiddu, F., Lee, B. J., Iao, L.-S., Harrison, N., & Macken, B. (2020). Does short-term memory develop? *Cognition*, *198*, Article 104200. https://doi.org/10.1016/j.cognition.2020.104200

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, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*, 648–654. https://doi.org/10.1037/0012-1649.23.5.648

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

Jusczyk, P. W., 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*, 822–836. https://doi.org/10.1037/0096-1523.21.4.822

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

Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition*, *29*(2), 363–372. https://doi.org/10.3758/BF03194931

Karch, J. (2020). Improving on adjusted *R*-squared. *Collabra: Psychology*, *6*(1), Article 45. https://doi.org/10.1525/collabra.343

Khan, K. S., Gugiu, M. R., Justice, L. M., Bowles, R. P., Skibbe, L. E., & Piasta, S. B. (2016). Age-Related Progressions in Story Structure in Young Children's Narratives. *Journal of Speech, Language, and Hearing Research: JSLHR*, *59*(6), 1395–1408. https://doi.org/10.1044/2016_JSLHR-L-15-0275

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *Quarterly Journal of Experimental Psychology*, *63*(1), 160–193. https://doi.org/10.1080/17470210902866664

Kidd, E., & Bavin, E. L. (2005). Lexical and referential cues to sentence interpretation: An investigation of children's interpretations of ambiguous sentences. *Journal of Child Language*, *32*(4), 855–876. https://doi.org/10.1017/S0305000905007051

Klammler, A., & Schneider, S. (2011). The size and composition of the productive holophrastic lexicon: German–Italian bilingual acquisition vs. Italian monolingual acquisition. *International Journal of Bilingual Education and Bilingualism*, *14*(1), 69–88. https://doi.org/10.1080/13670051003692840

Klein, D. E., & Murphy, G. L. (2001). The Representation of Polysemous Words. *Journal of Memory and Language*, *45*(2), 259–282. https://doi.org/10.1006/jmla.2001.2779

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

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. *Interspeech 2017*, 2198–2202. https://doi.org/10.31219/osf.io/86tu3

Lenzo, K. (2007). *The CMU pronouncing dictionary*. Carnegie Melon University.

Lieven, E. (2016). Usage-based approaches to language development: Where do we go from here? *Language and Cognition*, *8*(3), 346–368. https://doi.org/10.1017/langcog.2016.16

Lieven, E. V. M., Pine, J. M., & Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language*, *19*(2), 287–310. https://doi.org/10.1017/S0305000900011429

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In N. Arnett & R. Benett (Eds.), *Proceedings of the 30th West Coast Conference on Formal Linguistics* (pp. 237–247). Cascadilla. http://www.lingref.com/cpp/wccfl/30/paper2821.pdf

Lippeveld, M., & Oshima-Takane, Y. (2020). Children's initial understanding of the related meanings of polysemous noun-verb pairs. *Language Learning and Development*, *16*(3), 244–269. https://doi.org/10.1080/15475441.2020.1737073

Lloyd-Kelly, M., Gobet, F., & Lane, P. C. R. (2016). Under Pressure: How Time-Limited Cognition Explains Statistical Learning by 8-Month Old Infants. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016*, 1475–1480. https://livrepository.liverpool.ac.uk/3002153

Longobardi, E., Rossi-Arnaud, C., Spataro, P., Putnick, D. L., & Bornstein, M. H. (2015). Children's acquisition of nouns and verbs in Italian: Contrasting the roles of frequency and positional salience in maternal language. *Journal of Child Language*, *42*(1), 95–121. https://doi.org/10.1017/S0305000913000597

Longobardi, E., Spataro, P., L. Putnick, D., & Bornstein, M. H. (2016). Noun and Verb Production in Maternal and Child Language: Continuity, Stability, and Prediction Across the Second Year of Life. *Language Learning and Development*, *12*(2), 183–198. https://doi.org/10.1080/15475441.2015.1048339

Loukatou, G. (2019). *From phonemes to morphemes: Relating linguistic complexity to unsupervised word over-segmentation*. https://www.academia.edu/45273246/From_phonemes_to_morphemes_relating_linguistic_complexity_to_unsupervised_word_over_segmentation

Loukatou, G., Stoll, S., Blasi, D., & Cristia, A. (2022). Does morphological complexity affect word segmentation? Evidence from computational modeling. *Cognition*, *220*, 104960. https://doi.org/10.1016/j.cognition.2021.104960

Loukatou, G.-R., Stoll, S., Blasi, D., & Cristia, A. (2018). Modeling infant segmentation of two morphologically diverse languages. *Actes de La Conférence TALN. Volume 1 - Articles Longs, Articles Courts de TALN*, 47–60. https://aclanthology.org/2018.jeptalnrecital-long.4

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). *Analysis and Evaluation of Language Models for Word Sense Disambiguation* (arXiv:2008.11608). arXiv. https://doi.org/10.48550/arXiv.2008.11608

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed* (pp. xi, 366). Lawrence Erlbaum Associates Publishers.

Mani, N., & Plunkett, K. (2010). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445–470. https://doi.org/10.1111/j.1532-7078.2009.00027.x

Mani, N., Daum, M. M., & Huettig, F. (2016). "Proactive" in many ways: Developmental evidence for a dynamic pluralistic approach to prediction. *Quarterly Journal of Experimental Psychology*, *69*(11), 2189–2201. https://doi.org/10.1080/17470218.2015.1111395

Marino, C., Bernard, C., & Gervain, J. (2020). Word frequency is a cue to lexical category for 8-month-old infants. *Current Biology*, *30*(8), 1380–1386. https://doi.org/10.1016/j.cub.2020.01.070

Markman, E. M. (1991). *Categorization and Naming in Children: Problems of Induction.* The MIT Press. https://doi.org/10.7551/mitpress/1750.001.0001

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

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning Phonemes With a Proto-Lexicon. *Cognitive Science*, *37*(1), 103–124. https://doi.org/10.1111/j.1551-6709.2012.01267.x

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*(2), 91–121. https://doi.org/10.1016/s0010-0277(00)00109-8

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. https://doi.org/10.1006/cogp.1999.0721

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51. https://doi.org/10.1037/rev0000126

McGillion, M., Herbert, J. S., Pine, J., Vihman, M., dePaolis, R., Keren-Portnoy, T., & Matthews, D. (2017). What paves the way to conventional language? The predictive value of babble, pointing, and socioeconomic status. *Child Development*, *88*(1), 156–166. https://doi.org/10.1111/cdev.12671

McMurray, B. (2022). The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Language, Cognition and Neuroscience*, Advance online publication. https://doi.org/10.1080/23273798.2022.2105367

Mersad, K., Kabdebon, C., & Dehaene-Lambertz, G. (2021). Explicit access to phonetic representations in 3-month-old infants. *Cognition, 213*, Article 104613. https://doi.org/10.1016/j.cognition.2021.104613

Meylan, S. C., Mankewitz, J., Floyd, S., Rabagliati, H., & Srinivasan, M. (2021). Quantifying Lexical Ambiguity in Speech To and From English-Learning

Children. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/1pq031fn

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564. https://doi.org/10.1017/S0305000909990511

Monaghan, P., & Rowland, C. F. (2017). Combining Language Corpora With Experimental and Computational Approaches for Language Acquisition Research. *Language Learning*, *67*(S1), 14–39. https://doi.org/10.1111/lang.12221

Monaghan, P., Donnelly, S., Alcock, K., Bidgood, A., Cain, K., Durrant, S., Frost, R. L. A., Jago, L. S., Peter, M. S., Pine, J. M., Turnbull, H., & Rowland, C. F. (2023). Learning to generalise but not segment an artificial language at 17 months predicts children's language skills 3 years later. *Cognitive Psychology*, *147*, 101607. https://doi.org/10.1016/j.cogpsych.2023.101607

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and Diversity: Simulating Early Word Learning Environments. *Cognitive Science*, *42*(S2), 375–412. https://doi.org/10.1111/cogs.12592

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, *24*(1), 45–64. https://doi.org/10.1016/0010-0277(86)90004-1

Morais, J., Content, A., Cary, L., Mehler, J., & Segui, J. (1989). Syllabic segmentation and literacy. *Language and Cognitive Processes*, *4*(1), 57–67. https://doi.org/10.1080/01690968908406357

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 528–559. https://doi.org/10.1080/027249897392017

Mowrey, R., & Pagliuca, W. (1995). The reductive character of articulatory evolution. *Rivista di linguistica*, *7*, 37–124.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). Interaction of Word Learning and Semantic Category Formation in Late Talking. *Proceedings of the Cognitive Science Society*, *34*(34). http://escholarship.org/uc/item/8rf426sw

Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology*, *42*(4), 643–655. https://doi.org/10.1037/0012-1649.42.4.643

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

Norbury, C. F. (2005). Barking up the wrong tree? Lexical ambiguity resolution in children with language impairments and autistic spectrum disorders. *Journal of Experimental Child Psychology*, *90*(2), 142–171. https://doi.org/10.1016/j.jecp.2004.11.003

Pasini, T., & Camacho-Collados, J. (2020). A Short Survey on Sense-Annotated Corpora. *arXiv:1802.04744 [Cs]*. http://arxiv.org/abs/1802.04744

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

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

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

Perruchet, P., & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language*, *66*(4), 807–818. https://doi.org/10.1016/j.jml.2012.02.010

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

Phillips, L., & Pearl, L. (2014). 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. https://doi.org/10.1111/cogs.12217

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Pinker, S. (1989). Learnability and Cognition. *MIT Press*. https://mitpress.mit.edu/9780262660730/learnability-and-cognition/

Pinker, S. (1994a). How could a child use verb syntax to learn verb semantics? *Lingua*, *92*, 377–410. https://doi.org/10.1016/0024-3841(94)90347-6

Pinker, S. (1994b). *The language instinct*. *How the mind creates language.* Penguin.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. (1985). Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication*, *4*(1–3), 75–95. https://doi.org/10.1016/0167-6393(85)90037-8

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115. https://doi.org/10.1037/0033-295X.103.1.56

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, *81*(4), 927–964. https://doi.org/10.1353/lan.2005.0195

Postal, P. (1968) *Aspects of phonological theory*. Harper & Row.

Pustejovsky, J. (1995). The generative lexicon. Cambridge, MA: MIT.

Qi, Z., Love, J., Fisher, C., & Brown-Schmidt, S. (2020). Referential context and executive functioning influence children's resolution of syntactic ambiguity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(10), 1922–1947. https://doi.org/10.1037/xlm0000886

Rabagliati, H., Pylkkanen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, *49*(6), 1076–1089. https://doi.org/10.1037/a0026918

Rajaram, M. (2022). Phonological neighborhood measures and multisyllabic word acquisition in children. *Journal of Child Language*, *49*(1), 197–212. https://doi.org/10.1017/S0305000920000811

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*, *15*(2), 411–427. https://doi.org/10.1177/1745691619885860

Rodd, J. M. (2022). Word-meaning access: The one-to-many mapping from form to meaning. In A. Papafragou, J. C. Trueswell, & L. R. Gleitman (Eds.), *The Oxford Handbook of the Mental Lexicon* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198845003.013.1

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37. https://doi.org/10.1016/j.jml.2015.10.006

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2012). Dissociating Frontotemporal Contributions to Semantic Ambiguity Resolution in Spoken Sentences. *Cerebral Cortex*, *22*(8), 1761–1773. https://doi.org/10.1093/cercor/bhr252

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making Sense of Semantic Ambiguity: Semantic Competition in Lexical Access. *Journal of Memory and Language*, *46*(2), 245–266. https://doi.org/10.1006/jmla.2001.2810

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

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. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. 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

Saksida, A., Langus, A., & Nespor, M. (2016). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science*, *20*(3), Article e12390. https://doi.org/10.1111/desc.12390

Saxton, M. (2009). The Inevitability of Child Directed Speech. In S. Foster-Cohen (Ed.), *Language Acquisition* (pp. 62–86). Palgrave Macmillan UK. https://doi.org/10.1057/9780230240780_4

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., & Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, *118*(7), Article e2001844118. https://doi.org/10.1073/pnas.2001844118

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568. https://doi.org/10.1037/0033-295X.96.4.523

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

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language*, *35*(1), 1–24. https://doi.org/10.1017/S0305000907008215

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., 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

Silvey, C., Gentner, D., Richland, L. E., & Goldin-Meadow, S. (2023). Children's Early Spontaneous Comparisons Predict Later Analogical Reasoning Skills: An Investigation of Parental Influence. *Open Mind*, *7*, 483–509. https://doi.org/10.1162/opmi_a_00093

Simpson, G. B., & Foster, M. R. (1986). Lexical ambiguity and children's word recognition. *Developmental Psychology*, *22*, 147–154. https://doi.org/10.1037/0012-1649.22.2.147

Skarabela, B., Ota, M., O'Connor, R., & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition*, *211*, Article 104612. https://doi.org/10.1016/j.cognition.2021.104612

Smolík, F., & Filip, M. (2022). Corpus-based age of word acquisition: Does it support the validity of adult age-of-acquisition ratings? *PLOS ONE*, *17*(5), e0268504. https://doi.org/10.1371/journal.pone.0268504

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*(3), 238–299. https://doi.org/10.1016/j.cogpsych.2004.03.001

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language*, *58*(2), 574–608. https://doi.org/10.1016/j.jml.2007.08.001

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

Srinivasan, M., & Rabagliati, H. (2021). The Implications of Polysemy for Theories of Word Learning. *Child Development Perspectives*, *15*(3), 148–153. https://doi.org/10.1111/cdep.12411

Srinivasan, M., Berner, C., & Rabagliati, H. (2019). Children use polysemy to structure new word meanings. *Journal of Experimental Psychology. General*, *148*(5), 926–942. https://doi.org/10.1037/xge0000454

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 670–683. https://doi.org/10.1044/1092-4388(2009/08-0254)

Stokes, S. F. (2014). The impact of phonological neighborhood density on typical and atypical emerging lexicons. *Journal of Child Language*, *41*(3), 634–657. https://doi.org/10.1017/S030500091300010X

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1454–1468. https://doi.org/10.1044/1092-4388(2004/108)

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(2), 291–321. https://doi.org/10.1017/S030500090800891X

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. (2022). Infants' Learning of Speech Sounds and Word-Forms. In A. Papafragou, J. C. Trueswell, & L. R. Gleitman (Eds.), *The Oxford Handbook of the Mental Lexicon* (pp. 266–291). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198845003.013.6

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., Shatz, M., & Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language*, *24*(3), 535–565. https://doi.org/10.1017/S030500099700319X

Tenney, I., Das, D., & Pavlick, E. (2019). *BERT Rediscovers the Classical NLP Pipeline* (arXiv:1905.05950). arXiv. https://doi.org/10.48550/arXiv.1905.05950

Thornton, R. (2012). Studies at the interface of child language and models of language acquisition. *First Language*, *32*(1–2), 281–297. https://doi.org/10.1177/0142723711403881

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge University Press. https://doi.org/10.1017/CBO9780511527678

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1–2), 61–82. https://doi.org/10.1515/cogl.2001.012

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press. https://doi.org/10.2307/j.ctv26070v8

Tomasello, M. (2009). The usage-based theory of language acquisition. In *The Cambridge handbook of child language* (pp. 69–87). Cambridge Univ. Press. https://doi.org/10.1017/CBO9780511576164.005

Tonelli, L., De Marco, A., Vollmann, R., & Dressler, W. (1998). Le prime fasi dell'acquisizione della morfologia. *Parallela*, *6*, 281–301.

Trueswell, J. C., & Gleitman, L. R. (2007). Learning to parse and its implications for language acquisition. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568971.013.0039

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Valian V. V. (2015). Innateness and learnability. In Bavin E. L., Naigles L. R. (Eds.), *The Cambridge handbook of child language* (2nd ed., pp. 15–36). Cambridge, England: Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422. https://doi.org/10.1044/1092-4388(2008/030)

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329. https://doi.org/10.1111/1467-9280.00064

Volterra, V. (1984). Waiting for the Birth of a Sibling: The Verbal Fantasies of a 2-Year-Old Boy. In I. Bretherton (Ed.), *Symbolic Play* (pp. 219–248). Academic Press. https://doi.org/10.1016/B978-0-12-132680-7.50012-0

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. https://doi.org/10.1006/cogp.1995.1016

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

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728. https://doi.org/10.1017/S0142716418000152

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717. https://doi.org/10.1121/1.402450

Wiley, J., George, T., & Rayner, K. (2018). Baseball fans don't like lumpy batters: Influence of domain knowledge on the access of subordinate meanings. *The Quarterly Journal of Experimental Psychology*, *71*, 93–102. https://doi.org/10.1080/17470218.2016.1251470

Witzel, J., & Forster, K. (2014). Lexical co-occurrence and ambiguity resolution. *Language, Cognition and Neuroscience*, *29*(2), 158–185. https://doi.org/10.1080/01690965.2012.748925

Yacovone, A., Shafto, C. L., Worek, A., & Snedeker, J. (2021). Word vs. World Knowledge: A developmental shift from bottom-up lexical cues to top-down plausibility. *Cognitive Psychology*, *131*, 101442. https://doi.org/10.1016/j.cogpsych.2021.101442

Yeung, H. H., & Nazzi, T. (2014). Object labeling influences infant phonetic learning and generalization. *Cognition*, *132*(2), 151–163. https://doi.org/10.1016/j.cognition.2014.04.001

Yeung, H. H., Chen, L. M., & Werker, J. F. (2014). Referential labeling can facilitate phonetic learning in infancy. *Child Development*, *85*(3), 1036–1049. https://doi.org/10.1111/cdev.12185

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Zipf, G. (1936). *The psychobiology of language*. Routledge.

Appendix

Appendix S1: Computational Models

Transitional probability models

The implementation of forward and backward transitional probability models (FTP and BTP, respectively) followed the procedures found in previous studies (e.g., Frank et al., 2010; Larsen et al., 2017; Saksida et al., 2016) in which the transitional probabilities of a phoneme/syllable pair were computed as:

$$Forward\ TP(U_{t-1}, U_t) = \frac{F(U_{t-1}, U_t)}{F(U_{t-1})}$$

$$Backward\ TP(U_{t-1}, U_t) = \frac{F(U_{t-1}, U_t)}{F(U_t)}$$

Where $F(U_{t-1}, U_t)$ is the frequency of a pair of units (two phonemes or syllables), while $F(U_{t-1})$ and $F(U_t)$ are the frequencies of the first and second unit respectively. We used a strictly incremental version of these models in which transitional probabilities were updated at every utterance. A word boundary was placed within a phoneme/syllable target pair if the transitional probabilities of the surrounding pairs were both greater than the target pair transitional probability (i.e., relative threshold). Utterance boundaries were used as additional units available to the models, therefore for the phoneme pair ↲*h* in [↲*hellobaby*↲], $F(U_{t-1})$ would correspond to the frequency of the utterance-initial marker ↲ and $F(U_t)$ to the frequency of the phoneme *h*.

Although using an absolute threshold has been shown to increase models' performance at precision and recall measures (Gambell & Yang, 2006; Saksida et al., 2016), we instead used a relative threshold where word boundaries were posited based on the transitional probabilities of the surrounding biphones or syllable pairs. The choice of a relative threshold was consistent with studies showing that infants segment at local minima of transitional probability (e.g., Saffran et al., 1996; 1999), while we were not aware of any experimental findings that have provided direct evidence for an absolute threshold mechanism. Further, we used a strictly incremental version of the transitional probability models (i.e., word boundaries are

set based on current transitional probabilities of surrounding pairs), to match CLASSIC-UB and PUDDLE's incremental way of learning. One could apply the same incremental principle to an absolute threshold by updating a running average; indeed, absolute transitional probabilities can fall out of predictive incremental learning models (e.g., Baayen et al., 2013; Harmon & Kapatsinski, 2021).

PUDDLE

PUDDLE (Monaghan & Christiansen, 2010) parses utterances phoneme by phoneme, searching for a matched string in its lexicon (we also adapted the original model to process the input syllable by syllable). At the start of the segmentation process, whole sentences are stored in the lexicon as this is initially empty. Items in the lexicon are ranked by absolute frequency of occurrence (which guides further string matching). The frequency of an item is updated every time it is discovered in the input, making the model strictly incremental. The lexicon in PUDDLE stores chunks that can begin or end utterances, and these can comprise phonemes, phoneme pairs, or longer sequences of phonemes up to whole utterances. When PUDDLE finds a match in the lexicon, it only recognizes the item if (a) there is an item on its left which ends with a previously encountered ending, and (b) there is an item on its right which begins with a previously encountered beginning.

Random baseline

We chose to implement a fully random baseline which relies on a random coin toss to place a boundary after each input unit (Lignos, 2012). This baseline represents a scenario in which a child would segment the input by making random guesses on word boundary locations and tends to mostly segment short and frequent words as the input gives more opportunities to correctly segment them (Grimm et al., 2017), that is, these words are more likely to be discovered by chance. Comparing to chance is informative because, ideally, one would want a more complex model, which implements a specific segmentation mechanism, to at least perform better than chance. A fully random baseline is also more informative than baselines which

consider each utterance or each unit as a word (e.g., Bernard et al., 2020). These baselines would only discover a very low proportion of word types from the phonemic input (an utterance baseline would only discover types that appear as one-word utterances, while a unit baseline would only discover mono-phonemic word types). Finally, pseudorandom baselines are problematic because of their prior knowledge assumptions. For example, it is unlikely that infants have knowledge of the true probability of a word boundary to occur in the language (oracle baseline; e.g., Bernard et al., 2020), or the true average word length in cross-linguistic terms (Loukatou et al., 2019).

References

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*(3), 329–347. https://doi.org/10.1177/0023830913484896

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

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

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty* [Unpublished manuscript]. Department of Linguistics, University of Pennsylvania.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational

investigation. *Frontiers in Psychology*, *8*, Article 555. https://doi.org/10.3389/fpsyg.2017.00555

Harmon, Z., & Kapatsinski, V. (2021). A theory of repetition and retrieval in language production. *Psychological Review*, *128*(6), 1112–1144. https://doi.org/10.1037/rev0000305

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. *Interspeech 2017*, 2198–2202. https://doi.org/10.31219/osf.io/86tu3

Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In N. Arnett & R. Benett (Eds.), *Proceedings of the 30th West Coast Conference on Formal Linguistics* (pp. 237–247). Cascadilla. http://www.lingref.com/cpp/wccfl/30/paper2821.pdf

Loukatou, G. R., Moran, S., Blasi, D. E., Stoll, S., & Cristia, A. (2019). Is word segmentation child's play in all languages? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3931-3937). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1383

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*(3), 545–564. https://doi.org/10.1017/S0305000909990511

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Saksida, A., Langus, A., & Nespor, M. (2016). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science*, *20*(3), Article e12390. https://doi.org/10.1111/desc.12390

Appendix S2: Input Preprocessing

The seven CHILDES corpora used were: Belfast (Henry, 1995), Manchester (Theakston et al., 2001), Thomas (Lieven et al., 2009), Tommerdahl (Tommerdahl & Kilpatrick, 2013), Wells (Wells, 1981), Forrester (Forrester, 2002), Lara (Rowland & Fletcher, 2006). The corpora were imported into the R environment (R Core Team, 2018) using the package childesr (Braginsky et al., 2019), which guarantees a standardized procedure for obtaining the utterance samples. The corpora were phonetically transcribed using the *CMU Pronouncing Dictionary* (Lenzo, 2007). The transcription process was carried out without considering word stress markers in the dictionary. Utterances containing one or more words not appearing in the *CMU Pronouncing Dictionary* were discarded.

The advantage of using a transcription dictionary is that it allows automatic transcription of large input corpora into phonetic form. However, it has the important limitation of assuming that words always consist of the same phonemes in running speech. This is not the case as words undergo significant phonetic reduction in conversational speech (e.g., *until* [ʌntɪl] may be also realized via phoneme deletion [_ntil] or substitution [ʌntəl]; see Johnson, 2004). Addressing this limitation would require access to either phonetically transcribed corpora which include different word realizations (e.g., Schuppler et al., 2011) or to systems that directly operate on raw speech (e.g., Arnold et al., 2017; ten Bosch et al., 2022).

The corpora differed by mean length of utterance (MLU; see Table S2.1). If utterances are not shuffled, the models' performance oscillates depending on the corpus MLU. This happens because long sentences are more difficult to segment for all segmentation models. Given the input to different children is likely to show variability in MLU across time, we controlled for this variation by randomly shuffling the order of the utterances; given that this variation influenced all models equally, this choice should not affect comparisons between models.

When required, the syllabification of the input was performed using the WordSeg package (Bernard et al., 2020), which applies the maximal onset principle (Phillips & Pearl, 2015). We have not claimed that such a procedure corresponds to

how an infant would segment the input into syllables (for work focused on this problem, see Räsänen et al., 2018), as by definition the maximal onset principle requires prior knowledge of word onsets. Rather, it is a convenient deterministic strategy for presyllabifying the corpora, which can then be used as input for the models under the assumption that infants might be already organizing speech as strings of syllabic constituents before they have started representing word forms (e.g., Bertoncini & Mehler, 1981; Bertoncini et al., 1988; Bijeljac-Babic et al., 1993). Further, it is worth noting that the maximal onset principle is not the only strategy that could be used, as other variables can influence English syllabification (e.g., word-edge frequency, stress, vowel quality, sonority, morphology; Derwing, 1992; Derwing & Eddington, 2014; Olejarczuk & Kapatsinski, 2018).

The input for the models were all utterances from the seven corpora that were directed to children of age 2 years (see Table S2.1). We believed that this choice to focus on age 2 years was justified for two reasons. First, at age 2 years a larger amount of data on children's own productions is available in the corpora. Since we evaluated our models on measures that were based on child productions (i.e., age of first production and word-level measures), focusing on age 2 years allowed us to test the models on a much larger sample of word types. At ages earlier than 2 years child productions decrease significantly in type frequency (e.g., at Year 1, child word types are about one quarter of Year-2 word types) which would significantly limit the sample of words used to compute our evaluation measures.

Table S2.1 Descriptive statistics of phonetically transcribed CHILDES English corpora filtering for utterances directed to children of age 2 years

Corpus	Utterances	MLU	Word tokens	Word types
Forrester	3,183	4.89	15,567	1,576
Tommerdahl	5,700	4.84	27,610	1,646
Wells	16,292	3.62	59,042	3,053
Belfast	17,923	5.52	99,004	3,922
Lara	59,598	3.68	219,184	4,316
Thomas	194,695	5.23	1,018,726	8,160
Manchester	307,079	3.96	1,215,740	9,587

Note. For each corpus, the table indicates the number of input utterances, mean length of utterance (MLU, i.e., mean number of words in an utterance), number of words including repetitions (Word tokens), number of different words (Word types)

Second, the corpora also contain many more utterances directed to 2-year-olds than to younger children. The input available in CHILDES at Year 0 or 1 is significantly smaller in size compared to input directed to Year 2. For example, our 2-year-olds' input comprised 604,000 utterances, while 1-year-olds' input only contained 54,274 utterances. This was problematic because a smaller sample of utterances was more likely to be biased and less likely to preserve the characteristics of naturalistic speech directed to young children. Thus, focusing on age 2 years represented a compromise: This was the youngest age group for which a large enough (and thus representative) sample of child-directed speech was available.

To illustrate this point further, we have generated Figure S2.1 below. In this figure (panels Raw of each lexical measure), one can see that the characteristics of the input change significantly from age 0 to 2 years, with the presence of more long, infrequent, low-neighborhood, and high-phonotactic words as age increased.

Crucially, we can show that these differences were mostly due to differences in sample size (see also Montag et al., 2018). This was because the likelihood of finding long/infrequent/low-neighborhood/high-phonotactic words increases as sample size increases. Indeed, when we matched input at different age bins by sample size (i.e., sampling the same number of utterances as in the smallest age sample) we saw that the differences between corpora decreased substantially (see Matched panels of Figure S2.1). Therefore, the loss from choosing input directed to age 2 years (i.e., maximizing sample size at the expense of input age) was lower compared to choosing input smaller in size at earlier ages, which would instead grossly misrepresent the characteristics of the naturalistic input.

In conclusion, although some differences between speech at different ages remained when we controlled for input sample size, and it was thus possible that the age-2-year input was not entirely representative of the input at younger ages, the input that we had available for younger children was almost certainly not representative of naturalistic input directed to children of those ages either, because so little of it is available in existing corpora. Future studies may try to replicate these analyses using speech directed at earlier ages, once large-scale corpora of language input directed at such earlier ages become available to researchers.

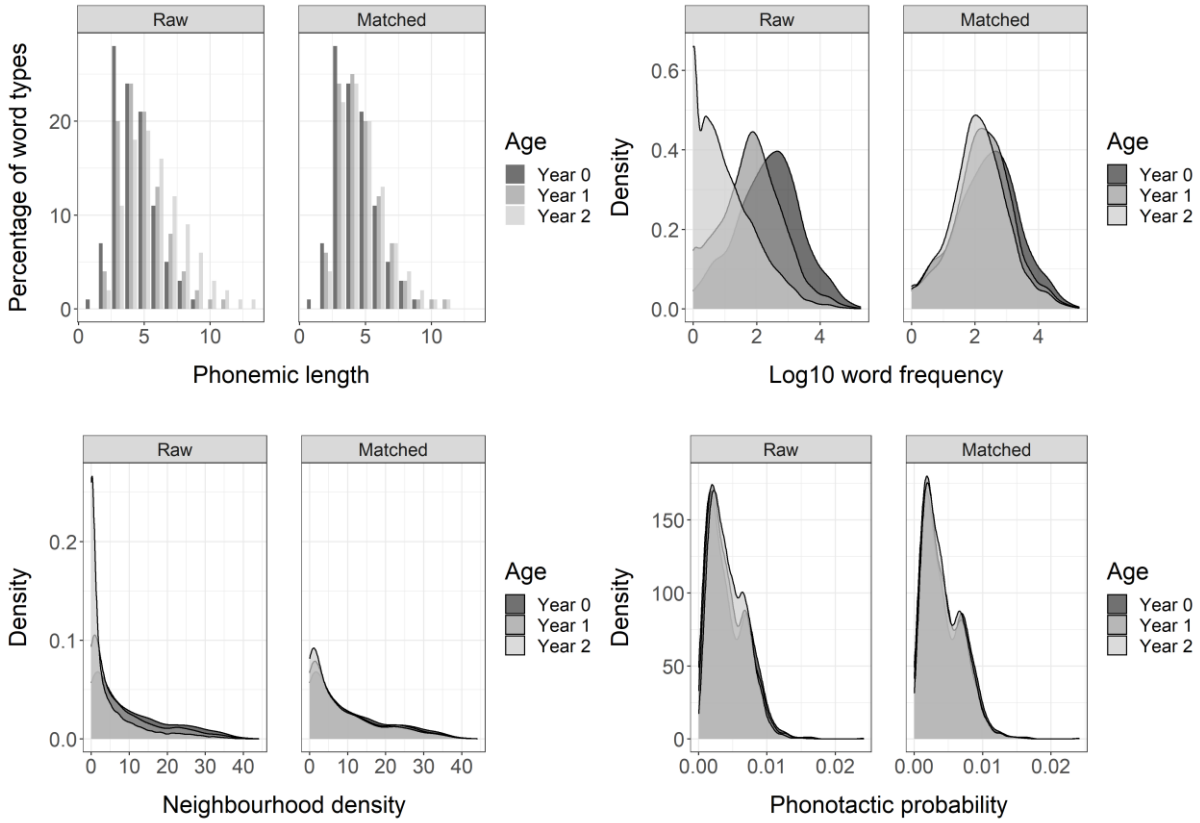


Figure S2.1 Word characteristics of word type distributions for input directed at Year 0, 1, and 2. Raw panels show word characteristics when considering all utterances available at each age (age 0 years = 11,745; age 1 year = 54,274; age 2 years = 604,000). Age-1- and 2-year utterances were taken from the same corpora used in Chapter 2, while age-0-year utterances were taken from the Korman corpus (Korman, 1992), which contains maternal speech directed to infants aged between 4 and 16 weeks. Matched panels refer to word type distributions when each input is matched by age-0-year sample size, therefore randomly sampling 11,745 utterances from age-1- and age-2-year corpora. Results did not depend on the particular random samples computed as repeating the sampling procedure produced identical distributions.

References

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., & Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between

meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, *12*(4), Article e0174623. https://doi.org/10.1371/journal.pone.0174623

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., Fibla, L., Dupoux, E., Daland, R., Cao, X. N., & Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, *52*(1), 264–278. https://doi.org/10.3758/s13428-019-01223-3

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, *4*, 247–260. https://doi.org/10.1016/S0163-6383(81)80027-6

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology. General*, *117*(1), 21–33. https://doi.org/10.1037//0096-3445.117.1.21

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

Braginsky, M., Sanchez, A., & Yurovsky, D. (2019). *childesr: Accessing the 'CHILDES' database* (R package, Version 0.1.2) [Computer software]. https://github.com/langcog/childesr

Derwing, B. L. (1992). A 'pause-break' task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results for five diverse languages. *Language and Speech*, *35*(1–2), 219–235. https://doi.org/10.1177/002383099203500217

Derwing, B. L., & Eddington, D. (2014). The experimental investigation of syllable structure. *The Mental Lexicon*, *9*(2), 170–195. https://doi.org/10.1075/ml.9.2.02der

Forrester, M. (2002). Appropriating cultural conceptions of childhood: Participation in conversation. *Childhood*, *9*, 255–278. https://doi.org/10.1177/0907568202009003043

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting.* Oxford University Press.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29–54). The National International Institute for Japanese Language.

Korman, M. (1992). *CHILDES English Korman Corpus*. https://doi.org/10.21415/T59G7B

Lenzo, K. (2007). *The CMU pronouncing dictionary*. Carnegie Melon University.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481–507. https://doi.org/10.1515/COGL.2009.022

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

Olejarczuk, P., & Kapatsinski, V. (2018). The metrical parse is guided by gradient phonotactics. *Phonology*, *35*(3), 367–405. https://doi.org/10.1017/S0952675718000106

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854. https://doi.org/10.1111/cogs.12217

R Core Team (2018). *R: A language and environment for statistical computing* (Version 4.0.3) [Computer software]. R Foundation for Statistical Computing.

Räsänen, O., Doyle, G., & Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, *171*, 130–150. https://doi.org/10.1016/j.cognition.2017.11.003

Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, *33*, 859–877. https://doi.org/10.1017/S0305000906007537

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96–109. https://doi.org/10.1016/j.wocn.2010.11.006

ten Bosch, L., Boves, L., & Ernestus, M. (2022). DIANA, a process-oriented model of human auditory word recognition. *Brain Sciences*, *12*(5), Article 681. https://doi.org/10.3390/brainsci12050681

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*, 127–152. https://doi.org/10.1017/S0305000900004608

Tommerdahl, J., & Kilpatrick, C. (2013). Analyzing reliability of grammatical production in spontaneous samples of varying length. *Journal of Child Language Teaching and Therapy, 29*(2), 171–183. https://doi.org/10.1177/0265659012459528

Wells, C. G. (1981). *Learning through interaction: The study of language development.* Cambridge University Press. https://doi.org/10.1017/CBO9780511620737

Appendix S3: Word Age of First Production Estimation

Word age of first production has been used in Grimm et al.'s (2017; 2019) studies as an index of word learning. If a word is first produced early in development, it is assumed that this is in part because it is easy to learn. To compute word age of first production estimates, we used Grimm et al.'s (2017; 2019) procedure, as its validity was assessed in two ways: corpora age of first production estimates showed a fairly strong correlation with American English communicative-development-inventories from parent-report measures of child expressive vocabulary (r_s = .50, p < .001) and a stronger correlation with the only estimates for British English that are directly derived from children (i.e., from a picture-naming task; Morrison et al., 1997; r_s = .65, p < .001).

To estimate word age of first production, we used mean length of utterance as a proxy of the developmental stage at which a word is acquired. For a given word, we first computed mean length of utterance for each transcript via bootstrapping (to compensate for differences in number of utterances). The lowest mean length of utterance across transcripts was then taken as age of first production value in order to correct for inflation (as it is likely that children knew a target word before they produced it in the recordings). This method also avoided having to find a set of common words across corpora to calculate a mean stage; finding a set of common words across corpora would have meant discarding a high amount of low frequency words that do not appear in all corpora, resulting in a skewed set of high-frequency words.

References

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, *8*, Article 555. https://doi.org/10.3389/fpsyg.2017.00555

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2019). Children probably store short rather than frequent or predictable chunks: Quantitative evidence from a

corpus study. *Frontiers in Psychology*, *10*, Article 80. https://doi.org/10.3389/fpsyg.2019.00080

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 528–559. https://doi.org/10.1080/027249897392017

Appendix S4: Comparison of Precision and Recall Measures

We have included a narrative account of the findings in Table S4.1 and S4.2 in the paper in the section Results and Discussion / Precision and Recall.

Table S4.1 Comparison of models for the precision and recall measures for phonemic input

								95% CI	
Model comparison	Measure	M1[a]	M2[b]	ΔM[c]	*t*	*p*	*df*	*LL*	*UL*
BTP vs. PUDDLE	Recall	.45	.79	−.33	−73.05	< .001	19,044.5	−.348	−.320
BTP vs. PUDDLE	Precision	.42	.73	−.32	−66.73	< .001	19,681.9	−.333	−.304
BTP vs. Baseline	Recall	.45	.17	.28	60.43	< .001	19,391.0	.266	.293
BTP vs. Baseline	Precision	.42	.14	.27	59.57	< .001	19,094.8	.258	.284
BTP vs. CLASSIC-UB initial/final	Precision	.42	.5	−.09	−16.96	< .001	19,989.6	−.101	−.071
BTP vs. CLASSIC-UB final	Precision	.42	.49	−.07	−13.26	< .001	19,936.1	−.084	−.055
BTP vs. FTP	Recall	.45	.51	−.05	10.31	< .001	19,998.0	−.066	−.038
BTP vs. FTP	Precision	.42	.47	−.05	−10.26	< .001	19,997.8	−.067	−.037
BTP vs. CLASSIC-UB initial/final	Recall	.45	.5	−.04	−8.58	< .001	19,988.7	−.058	−.028
BTP vs. CLASSIC-UB final	Recall	.45	.45	.01	1.17	.243	19,907.3	−.004	.016
FTP vs. Baseline	Recall	.51	.17	.33	71.56	< .001	19,381.2	.318	.346
FTP vs. Baseline	Precision	.47	.14	.32	70.78	< .001	19,073.4	.310	.338
FTP vs. PUDDLE	Recall	.51	.79	−.28	−61.57	< .001	19,032.8	−.294	−.266
FTP vs. PUDDLE	Precision	.47	.73	−.26	55.74	< .001	19,668.1	−.279	−.251
CLASSIC-UB final vs. Baseline	Precision	.49	.14	.34	72.09	< .001	18,639.9	.326	.355

CLASSIC-UB final vs. PUDDLE	Recall	.45	.79	−.34	−71.29	< .001	18,476.8	−.355	−.321
CLASSIC-UB final vs. Baseline	Recall	.45	.17	.27	56.75	< .001	18,895.5	.259	.289
CLASSIC-UB final vs. PUDDLE	Precision	.49	.73	−.25	50.59	< .001	19,362.6	−.265	−.231
CLASSIC-UB final vs. FTP	Recall	.45	.51	−.06	−11.13	< .001	19,911.2	−.073	−.042
CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.45	.50	−.05	−9.44	< .001	19,955.8	−.066	−.034
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.49	.50	−.02	3.36	.003	19,973.2	−.032	.004
CLASSIC-UB final vs. FTP	Precision	.49	.47	.02	3.28	.003	19,942.2	.004	.028
CLASSIC-UB initial/final vs. Baseline	Precision	.50	.14	.36	77.5	< .001	18,935.1	.345	.374
CLASSIC-UB initial/final vs. Baseline	Recall	.50	.17	.32	68.94	< .001	19,245.2	.311	.339
CLASSIC-UB initial/final vs. PUDDLE	Recall	.50	.79	−.29	−62.62	< .001	18,872.7	−.304	−.274
CLASSIC-UB initial/final vs. PUDDLE	Precision	.50	.73	−.23	−47.98	< .001	19,575.9	−.247	−.215
CLASSIC-UB initial/final vs. FTP	Precision	.50	.47	.03	6.79	< .001	19,991.8	.020	.047
CLASSIC-UB initial/final vs. FTP	Recall	.50	.51	−.01	1.63	.206	19,989.9	−.020	.002

PUDDLE vs. Baseline	Recall	.79	.17	.61	149.07	< .001	19,950.6	.598	.624
PUDDLE vs. Baseline	Precision	.73	.14	.59	138.79	< .001	19,825.3	.574	.601

Note. Pairwise comparisons via Welch's *t* test for unequal variances; *p* values and bootstrap 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

[a]M1 = first model mean. [b]M2 = second model mean. [c]ΔM = mean difference.

Table S4.2 Comparison of precision and recall measures for syllabified input

Model comparison	Measure	M1[a]	M2[b]	ΔM[c]	*t*	*p*	*df*	95% CI	
								LL	*UL*
BTP vs. PUDDLE	Recall	.38	.89	−.51	−116.28	< .001	15,269.446	−.528	−.503
BTP vs. PUDDLE	Precision	.46	.85	−.40	−87.41	< .001	17,137.092	−.411	−.384
BTP vs. CLASSIC-UB initial/final	Precision	.46	.66	−.20	−39.55	< .001	19,734.761	.217	−.185
BTP vs. CLASSIC-UB initial/final	Recall	.38	.58	−.20	−37.96	< .001	19,913.453	−.219	−.186
BTP vs. CLASSIC-UB final	Precision	.46	.57	−.11	−20.43	< .001	19,984.445	−.123	−.093
BTP vs. CLASSIC-UB final	Recall	.38	.48	−.11	−19.2	< .001	19,996.639	−.12	−.088
BTP vs. Baseline	Recall	.38	.46	−.08	−14.24	< .001	19,997.407	−.095	−.063
BTP vs. Baseline	Precision	.46	.51	−.06	−10.37	< .001	19,997.833	−.072	−.041
BTP vs. FTP	Precision	.46	.49	−.04	−6.79	< .001	19,987.974	−.052	−.022
BTP vs. FTP	Recall	.38	.41	−.03	−5.52	< .001	19,997.237	−.047	−.016
FTP vs. PUDDLE	Recall	.41	.89	−.48	-109.94	< .001	15,324.968	−.499	−.471
FTP vs. PUDDLE	Precision	.49	.85	−.36	−80.70	< .001	17,360.737	−.376	−.346
FTP vs. Baseline	Recall	.41	.46	−.05	−8.75	< .001	19,997.989	−.064	−.033
FTP vs. Baseline	Precision	.49	.51	−.02	−3.69	.001	19,990.389	−.034	−.006
CLASSIC-UB final vs. PUDDLE	Recall	.48	.89	−.41	−93.02	< .001	15,343.657	−.422	−.394
CLASSIC-UB final vs. PUDDLE	Precision	.57	.85	−.29	−64.64	< .001	17,397.080	−.302	−.272
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.57	.66	−.09	−18.51	< .001	19,838.308	−.107	−.079

CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.48	.58	−.10	−18.31	< .001	19,933.371	−.112	−.081
CLASSIC-UB final vs. FTP	Precision	.57	.49	.07	13.78	< .001	19,997.734	.057	.088
CLASSIC-UB final vs. FTP	Recall	.48	.41	.08	13.71	< .001	19,997.914	.059	.091
CLASSIC-UB final vs. Baseline	Precision	.57	.51	.05	9.95	< .001	19,987.279	.037	.068
CLASSIC-UB final vs. Baseline	Recall	.48	.46	.03	4.95	< .001	19,997.843	.013	.042
CLASSIC-UB initial/final vs. PUDDLE	Recall	.58	.89	−.31	−74.00	< .001	15,872.605	−.325	−.297
CLASSIC-UB initial/final vs. PUDDLE	Precision	.66	.85	−.19	−46.49	< .001	18,269.313	−.208	−.179
CLASSIC-UB initial/final vs. FTP	Precision	.66	.49	.16	32.86	< .001	19,825.279	.151	.180
CLASSIC-UB initial/final vs. FTP	Recall	.58	.41	.17	32.39	< .001	19,928.617	.157	.189
CLASSIC-UB initial/final vs. Baseline	Precision	.66	.51	.15	28.66	< .001	19,747.424	.129	.161
CLASSIC-UB initial/final vs. Baseline	Recall	.58	.46	.12	23.37	< .001	19,926.899	.108	.140
PUDDLE vs. Baseline	Recall	.89	.46	.44	98.97	< .001	15,318.378	.422	.448
PUDDLE vs. Baseline	Precision	.85	.51	.34	75.27	< .001	17,165.941	.326	.351

Note. Pairwise comparisons via Welch's *t* test for unequal variances; *p* values and bootstrap 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

[a]M1 = first model mean. [b]M2 = second model mean. [c]ΔM = mean difference.

Appendix S5: Frequency-Weighted Age of First Production Analyses: Pairwise Differences Between Models' Adjusted R^2

A narrative account of the phonemic analysis is available in section Results and Discussion / Word Age of First Production of Chapter 2. Instead, here we focus on findings when models were run on syllabified input. Table S5.1 shows that models run on syllabified input did not perform better than the baseline. PUDDLE performed better than CLASSIC-UB initial-final and CLASSIC-UB final at predicting children's word age of first production. However, the difference between PUDDLE and the baseline was not significant and neither was the difference between the baseline and CLASSIC-UB initial-final.

CLASSIC-UB initial-final performed better than CLASSIC-UB final at predicting children's word age of first production.

PUDDLE explained a significantly higher proportion of variance in word age of first production than forward and backward transitional probability models.

Table S5.1 Frequency-weighted age of first production analyses: Pairwise differences between adjusted R^2 values of models when phonemic or syllabified input was used

			95% CI	
Model comparison	Input type	ΔR^2	*LL*	*UL*
BTP vs. Baseline	Phoneme	.008	−.011	.026
FTP vs. Baseline	Phoneme	.010	−.011	.031
FTP vs. BTP	Phoneme	.002	−.013	.016
CLASSIC-UB final vs. Baseline	Phoneme	.043	.020	.069
CLASSIC-UB final vs. BTP	Phoneme	.035	.013	.059
CLASSIC-UB final vs. FTP	Phoneme	.033	.011	.057
CLASSIC-UB final vs. PUDDLE	Phoneme	.001	−.015	.018
CLASSIC-UB initial/final vs. Baseline	Phoneme	.048	.024	.073
CLASSIC-UB initial/final vs. BTP	Phoneme	.040	.010	.061
CLASSIC-UB initial/final vs. FTP	Phoneme	.038	.016	.059
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.006	−.016	.029
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.005	−.006	.016
PUDDLE vs. Baseline	Phoneme	.042	.021	.064
PUDDLE vs. BTP	Phoneme	.034	.011	.054
PUDDLE vs. FTP	Phoneme	.032	.014	.054
Baseline vs. BTP	Syllable	.041	.024	.058
Baseline vs. FTP	Syllable	.028	.012	.047
Baseline vs. CLASSIC-UB final	Syllable	.020	.005	.040
Baseline vs. CLASSIC-UB initial/final	Syllable	.003	−.012	.020
FTP vs. BTP	Syllable	.013	.005	.022
CLASSIC-UB final vs. BTP	Syllable	.021	.011	.032

CLASSIC-UB final vs. FTP	Syllable	.008	−.004	.020
CLASSIC-UB initial/final vs. BTP	Syllable	.038	.025	.054
CLASSIC-UB initial/final vs. FTP	Syllable	.025	.010	.044
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.017	.010	.026
PUDDLE vs. BTP	Syllable	.061	.043	.082
PUDDLE vs. FTP	Syllable	.048	.023	.068
PUDDLE vs. CLASSIC-UB final	Syllable	.040	.024	.062
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.023	.008	.041
PUDDLE vs. Baseline	Syllable	.020	−.001	.039

Note. ΔR^2 = difference between adjusted R^2 values. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S6: Frequency-Unweighted Age of First Production Analyses

Interestingly, Larsen et al. (2017) found that a forward transitional probability model run on syllabified input showed the best performance, predicting 19% of variance in word age of acquisition. In contrast, we found that a forward transitional probability model run on syllabified input predicted a low proportion of variance, $R^2_{adjusted}$ = .013, 95% CI [.007, .021] (see Table 1 in the main paper). We suggest this difference was related to differences in the predictor measure; we weighted the predictor measure by the frequency of a target word in the input, while Larsen used raw counts. Accordingly, when we used raw counts and syllabified input, we were able to replicate Larsen's finding (see Tables S6.1 and S6.1), with the forward transitional probability model showing the best performance followed by CLASSIC-UB final. Importantly, however, even in this analysis we found that no model outperformed the baseline, with the baseline performing significantly better than forward transitional probability. Larsen did not include a comparison to a random baseline. We also obtained the same result when using raw counts and phonemic input, with CLASSIC-UB final showing the best performance but not being able to outperform the baseline. These results indicated that controlling for input word frequency and including a random baseline are both important to draw conclusions about the developmental plausibility of different segmentation models. A discussion on the role of the random baseline is included in Appendix S13.

Table S6.1 Adjusted R^2 for linear regression models predicting word age of first production as a function of unweighted log10 number of times a word was correctly segmented by each model

	Phonemic input			Syllabified input		
		95% CI			95% CI	
Model	$R^2_{adjusted}$	*LL*	*UL*	$R^2_{adjusted}$	*LL*	*UL*
Baseline	.273	.252	.295	.340	.310	.364
Backward transitional probability	.153	.140	.167	.225	.202	.249
Forward transitional probability	.168	.151	.185	.311	.284	.338
CLASSIC-UB final	.227	.205	.250	.301	.274	.327
CLASSIC-UB initial/final	.196	.176	.219	.278	.252	.302
PUDDLE	.195	.175	.214	.217	.194	.238

Note. Heteroskedasticity-robust standard errors were computed using a HC2 estimator. The 95% confidence intervals indicate lower and upper limits of bootstrap confidence intervals around the estimate (based on 1,000 iterations). Holm's correction was applied to the confidence intervals.

Table S6.2 Pairwise differences between adjusted R^2 of unweighted age of first production models

Model comparison	Input type	ΔR^2	95% CI	
			LL	*UL*
Baseline vs. BTP	Phoneme	.120	.101	.140
Baseline vs. FTP	Phoneme	.105	.088	.126
Baseline vs. PUDDLE	Phoneme	.078	.060	.098
Baseline vs. CLASSIC-UB initial/final	Phoneme	.077	.056	.097
Baseline vs. CLASSIC-UB final	Phoneme	.046	.026	.069
FTP vs. BTP	Phoneme	.015	.000	.029
CLASSIC-UB final vs. BTP	Phoneme	.074	.051	.098
CLASSIC-UB final vs. FTP	Phoneme	.059	.031	.083
CLASSIC-UB final vs. PUDDLE	Phoneme	.032	.009	.056
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.031	.020	.042
CLASSIC-UB initial/final vs. BTP	Phoneme	.043	.022	.065
CLASSIC-UB initial/final vs. FTP	Phoneme	.028	.008	.048
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.001	−.014	.017
PUDDLE vs. BTP	Phoneme	.042	.022	.060
PUDDLE vs. FTP	Phoneme	.027	.009	.047
Baseline vs. PUDDLE	Syllable	.123	.101	.145
Baseline vs. BTP	Syllable	.115	.095	.137
Baseline vs. CLASSIC-UB initial/final	Syllable	.062	.042	.08
Baseline vs. CLASSIC-UB final	Syllable	.039	.020	.058
Baseline vs. FTP	Syllable	.029	.010	.047
BTP vs. PUDDLE	Syllable	.008	−.015	.030
FTP vs. PUDDLE	Syllable	.094	.067	.122

FTP vs. BTP	Syllable	.086	.064	.109
FTP vs. CLASSIC-UB initial/final	Syllable	.033	.012	.056
FTP vs. CLASSIC-UB final	Syllable	.010	−.007	.026
CLASSIC-UB final vs. PUDDLE	Syllable	.084	.061	.109
CLASSIC-UB final vs. BTP	Syllable	.076	.053	.100
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.023	.013	.033
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.061	.041	.080
CLASSIC-UB initial/final vs. BTP	Syllable	.053	.023	.079

Note. ΔR^2 = difference between adjusted R^2 values. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

References

Larsen, E., Cristia, A., & Dupoux, E. (2017). Relating unsupervised word segmentation to reported vocabulary acquisition. *Interspeech 2017*, 2198–2202. https://doi.org/10.31219/osf.io/86tu3

Appendix S7: Approximation of Child Production Vocabulary by Phonemic Length

We ran analyses on both phonemic and syllabified input. We have given a narrative account of the phonemic-input analysis in the section Results and Discussion / Word-Level Measures / Phonemic Length of Chapter 2; below we focus on syllabified input.

When we used syllabified input (see Figure S7.1, Table S7.1, and Table S7.2), the model with the best performance was CLASSIC-UB final, but even this model did not outperform the baseline at approximating children's vocabularies by phonemic length. We have included a discussion on the role of the random baseline in Appendix S13.

CLASSIC-UB initial-final and CLASSIC-UB final showed a better performance than PUDDLE at approximating children's vocabularies by phonemic length. We found no significant difference when comparing CLASSIC-UB final and CLASSIC-UB initial-final. Finally, PUDDLE performance did not differ statistically from the backward transitional probability model and was significantly worse than the forward transitional probability model.

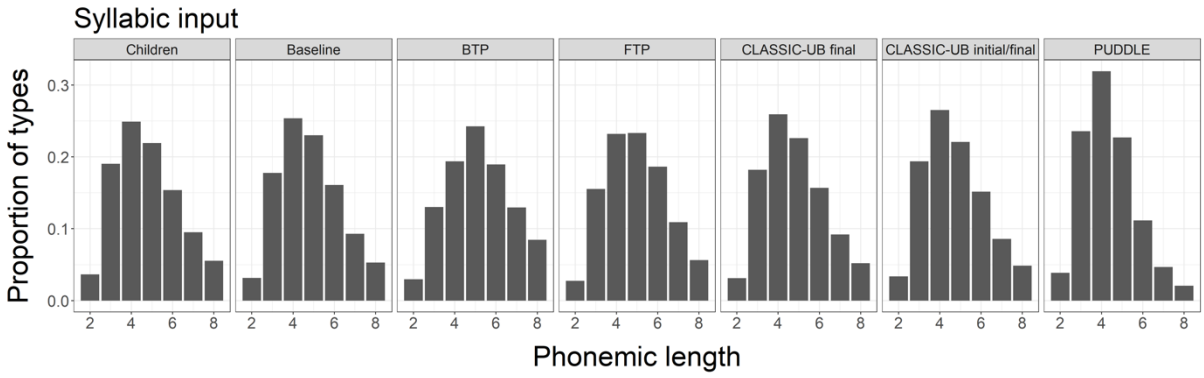


Figure S7.1 Proportion of unique words (types) produced by children and discovered by each model by phonemic length, when syllabified input was used.

Table S7.1 Child-model comparison by phonemic length.

Comparison	Input type	χ^2	*df*	*p*	95% CI	
					LL	*UL*
Children vs. Baseline	Phoneme	528.99	6	< .001	421.46	691.44
Children vs. BTP	Phoneme	1,314.99	6	< .001	1,112.25	1,552.42
Children vs. FTP	Phoneme	1274.04	6	< .001	1,107.59	1,486.48
Children vs. CLASSIC-UB final	Phoneme	244.9	6	< .001	167.47	357.26
Children vs. CLASSIC-UB initial/final	Phoneme	311.02	6	< .001	223.76	440.03
Children vs. PUDDLE	Phoneme	1178.97	6	< .001	969.29	1,406.66
Children vs. Baseline	Syllable	16.62	6	.022	6.1	64.2
Children vs. BTP	Syllable	401.07	6	< .001	268.08	598.9
Children vs. FTP	Syllable	130.19	6	< .001	67.29	244.6
Children vs. CLASSIC-UB final	Syllable	14.62	6	.023	7.11	56.59
Children vs. CLASSIC-UB initial/final	Syllable	19.34	6	.011	6.62	74.01
Children vs. PUDDLE	Syllable	439.45	6	< .001	317.47	604.34

Note. We compared the probability of observing words of different phonemic lengths in the models' vocabularies against the expected probability of words being of a given phonemic length in children's vocabularies. Comparisons were tested via a chi-square goodness of fit test. The chi-square statistic always compares the distance of a model's distribution from children's. The table shows the type of comparison, the input type used, the chi-square statistic, degrees of freedom, *p* value and cut-offs of 95% bootstrap confidence interval of the statistic. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S7.2 Pairwise differences between the chi-square statistics reported in Table S7.1, comparing how well two models' observed probabilities of phonemic lengths fit children's expected probabilities, when phonemic or syllabified input is used

			95% CI	
Model comparison	Input type	$\Delta\chi^2$	*LL*	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	284.09	146.62	416.98
Baseline vs. CLASSIC-UB initial/final	Phoneme	217.97	69.46	393.14
BTP vs. CLASSIC-UB final	Phoneme	1,070.09	874.65	1,291.03
BTP vs. CLASSIC-UB initial/final	Phoneme	1,003.97	813.07	1,255.92
BTP vs. Baseline	Phoneme	785.99	564.44	983.90
BTP vs. PUDDLE	Phoneme	136.01	−75.75	368.09
BTP vs. FTP	Phoneme	40.94	−133.07	239.68
FTP vs. CLASSIC-UB final	Phoneme	1029.14	856.78	1260.98
FTP vs. CLASSIC-UB initial/final	Phoneme	963.02	736.17	1207.71
FTP vs. Baseline	Phoneme	745.05	551.63	946.67
FTP vs. PUDDLE	Phoneme	95.07	−111.00	287.36
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	66.12	−41.65	172.99
PUDDLE vs. CLASSIC-UB final	Phoneme	934.07	717.61	1,153.22
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	867.95	679.12	1,084.11
PUDDLE vs. Baseline	Phoneme	649.98	423.18	858.81
Baseline vs. CLASSIC-UB final	Syllable	2.00	−25.63	31.97
BTP vs. CLASSIC-UB final	Syllable	386.45	203.26	578.85
BTP vs. Baseline	Syllable	384.45	223.81	556.98
BTP vs. CLASSIC-UB initial/final	Syllable	381.73	200.61	570.78
BTP vs. FTP	Syllable	270.88	122.27	419.30
FTP vs. CLASSIC-UB final	Syllable	115.57	31.55	219.54

FTP vs. Baseline	Syllable	113.57	31.45	212.32
FTP vs. CLASSIC-UB initial/final	Syllable	110.85	13.61	226.86
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	4.72	−33.69	43.77
CLASSIC-UB initial/final vs. Baseline	Syllable	2.72	−36.82	44.92
PUDDLE vs. CLASSIC-UB final	Syllable	424.83	299.34	576.60
PUDDLE vs. Baseline	Syllable	422.83	253.88	609.58
PUDDLE vs. CLASSIC-UB initial/final	Syllable	420.11	274.95	581.06
PUDDLE vs. FTP	Syllable	309.26	78.34	508.42
PUDDLE vs. BTP	Syllable	38.38	−211.69	295.25

Note. The $\Delta\chi^2$ measure examined whether two models' distributions were at the same distance from children's expected probabilities. The order of each pairwise difference was set as in the column Comparison (e.g., in Baseline vs. CLASSIC-UB final, the CLASSIC-UB final χ^2 estimate is subtracted from the Baseline χ^2 estimate). Lower and upper limits of bootstrap 95% confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S8: Approximation of Child Production Vocabulary by Weighted Log10 Word Frequency

We ran analyses on both phonemic and syllabified input. We have provided a narrative account of the phonemic-input analysis in the section Results and Discussion / Word-Level Measures / Word Frequency of Chapter 2; below we focus on syllabified input.

When syllabified input was used (see Figure S8.1, Table S8.1, and Table S8.2), PUDDLE outperformed CLASSIC-UB initial-final and CLASSIC-UB final at approximating children's vocabularies by weighted log10 word frequency. However, neither PUDDLE nor CLASSIC-UB initial-final were able to outperform the baseline. We have included a discussion on the role of the random baseline in Appendix S13.

CLASSIC-UB final did not differ statistically from CLASSIC-UB initial-final. Finally, PUDDLE outperformed forward and backward transitional probability at approximating children's vocabularies by weighted log10 word frequency.

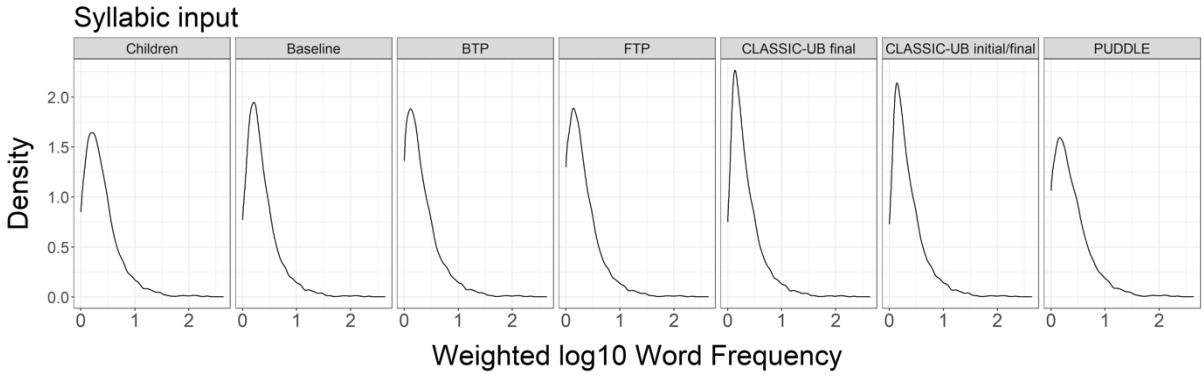


Figure S8.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted log10 word frequency. Syllabified input was used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

Table S8.1 Child-model comparison by weighted log10 word frequency

Model comparison	Input type	*D*	*p*	95% CI *LL*	95% CI *UL*
Children vs. Baseline	Phoneme	.29	< .001	.27	.32
Children vs. BTP	Phoneme	.26	< .001	.23	.30
Children vs. FTP	Phoneme	.23	< .001	.20	.27
Children vs. CLASSIC-UB final	Phoneme	.13	< .001	.11	.15
Children vs. CLASSIC-UB initial/final	Phoneme	.16	< .001	.14	.19
Children vs. PUDDLE	Phoneme	.13	< .001	.11	.16
Children vs. Baseline	Syllable	.05	< .001	.03	.07
Children vs. BTP	Syllable	.11	< .001	.09	.14
Children vs. FTP	Syllable	.10	< .001	.08	.12
Children vs. CLASSIC-UB final	Syllable	.10	< .001	.08	.12
Children vs. CLASSIC-UB initial/final	Syllable	.07	< .001	.06	.10
Children vs. PUDDLE	Syllable	.04	< .001	.03	.06

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted log10 word frequency were compared to child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S8.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S8.1

Model comparison	Input type	Δ*D*	95% CI	
			LL	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	.169	.136	.198
Baseline vs. PUDDLE	Phoneme	.163	.127	.192
Baseline vs. CLASSIC-UB initial/final	Phoneme	.132	.087	.159
Baseline vs. FTP	Phoneme	.064	.028	.101
Baseline vs. BTP	Phoneme	.030	−.006	.069
BTP vs. CLASSIC-UB final	Phoneme	.138	.098	.175
BTP vs. PUDDLE	Phoneme	.133	.096	.175
BTP vs. CLASSIC-UB initial/final	Phoneme	.101	.063	.139
BTP vs. FTP	Phoneme	.034	-.001	.072
FTP vs. CLASSIC-UB final	Phoneme	.105	.067	.145
FTP vs. PUDDLE	Phoneme	.099	.063	.132
FTP vs. CLASSIC-UB initial/final	Phoneme	.068	.028	.106
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.037	.006	.072
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.032	−.006	.069
PUDDLE vs. CLASSIC-UB final	Phoneme	.005	−.020	.036
Baseline vs. PUDDLE	Syllable	.003	−.017	.028
BTP vs. PUDDLE	Syllable	.066	.039	.091
BTP vs. Baseline	Syllable	.063	.028	.089
BTP vs. CLASSIC-UB initial/final	Syllable	.035	.009	.060
BTP vs. FTP	Syllable	.015	−.008	.034
BTP vs. CLASSIC-UB final	Syllable	.014	−.010	.038
FTP vs. PUDDLE	Syllable	.052	.028	.081

FTP vs. Baseline	Syllable	.048	.023	.073
FTP vs. CLASSIC-UB initial/final	Syllable	.021	−.002	.042
CLASSIC-UB final vs. PUDDLE	Syllable	.053	.028	.079
CLASSIC-UB final vs. Baseline	Syllable	.049	.019	.074
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.022	−.002	.045
CLASSIC-UB final vs. FTP	Syllable	.001	−.017	.016
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.031	.006	.058
CLASSIC-UB initial/final vs. Baseline	Syllable	.028	.000	.052

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted log10 word frequency when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S9: Approximation of Child Production Vocabulary by Weighted Neighbourhood Density

We ran analyses on both phonemic and syllabified input. We have included a narrative account of the phonemic-input analysis in the section Results and Discussion / Word-Level Measures / Neighbourhood Density of Chapter 2; below we focus on syllabified input.

When we used syllabified input (see Figure S9.1, Table S9.1, and Table S9.2), CLASSIC-UB final showed the best performance at approximating children's vocabularies by weighted neighbourhood density, but it was not able to outperform the baseline. We have included a discussion on the role of the random baseline in Appendix S13.

CLASSIC-UB final did not differ statistically from CLASSIC-UB initial-final. Finally, PUDDLE did not differ statistically from backward transitional probability and performed significantly worse than forward transitional probability at approximating children's vocabularies by weighted neighbourhood density.

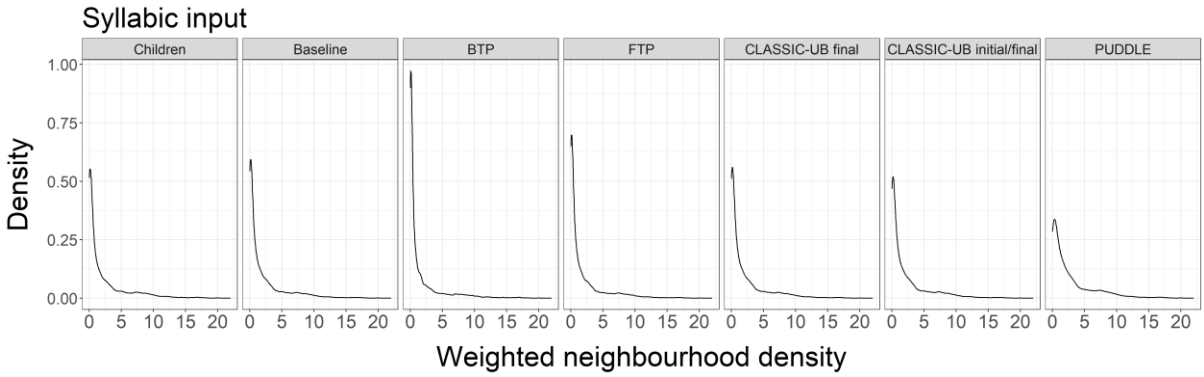


Figure S9.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model by weighted neighborhood density. Syllabified input was used.

Table S9.1 Child-model comparison by weighted neighbourhood density

Model comparison	Input type	*D*	*p*	95% CI *LL*	95% CI *UL*
Children vs. Baseline	Phoneme	.20	< .001	.18	.23
Children vs. BTP	Phoneme	.37	< .001	.34	.40
Children vs. FTP	Phoneme	.34	< .001	.32	.37
Children vs. CLASSIC-UB final	Phoneme	.14	< .001	.12	.17
Children vs. CLASSIC-UB initial/final	Phoneme	.18	< .001	.16	.21
Children vs. PUDDLE	Phoneme	.29	< .001	.26	.32
Children vs. Baseline	Syllable	.03	.029	.02	.04
Children vs. BTP	Syllable	.12	< .001	.10	.14
Children vs. FTP	Syllable	.05	< .001	.03	.07
Children vs. CLASSIC-UB final	Syllable	.03	.005	.02	.05
Children vs. CLASSIC-UB initial/final	Syllable	.05	< .001	.03	.07
Children vs. PUDDLE	Syllable	.17	< .001	.14	.19

Note. Model distributions of unique words by weighted neighbourhood density were compared to child distribution. The 95% bootstrap confidence intervals of the statistic were adjusted using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S9.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S9.1

Model comparison	Input type	Δ*D*	95% CI	
			LL	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	.063	.034	.092
Baseline vs. CLASSIC-UB initial/final	Phoneme	.021	−.007	.050
BTP vs. CLASSIC-UB final	Phoneme	.228	.191	.261
BTP vs. CLASSIC-UB initial/final	Phoneme	.185	.147	.222
BTP vs. Baseline	Phoneme	.164	.127	.197
BTP vs. PUDDLE	Phoneme	.081	.049	.110
BTP vs. FTP	Phoneme	.028	−.002	.058
FTP vs. CLASSIC-UB final	Phoneme	.199	.168	.229
FTP vs. CLASSIC-UB initial/final	Phoneme	.157	.122	.191
FTP vs. Baseline	Phoneme	.136	.101	.171
FTP vs. PUDDLE	Phoneme	.053	.023	.083
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.042	.007	.074
PUDDLE vs. CLASSIC-UB final	Phoneme	.146	.115	.178
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.104	.065	.135
PUDDLE vs. Baseline	Phoneme	.083	.049	.118
BTP vs. Baseline	Syllable	.092	.055	.123
BTP vs. CLASSIC-UB final	Syllable	.086	.044	.119
BTP vs. FTP	Syllable	.073	.045	.100
BTP vs. CLASSIC-UB initial/final	Syllable	.066	.028	.103
FTP vs. Baseline	Syllable	.019	−.010	.044
FTP vs. CLASSIC-UB final	Syllable	.013	−.019	.039
CLASSIC-UB final vs. Baseline	Syllable	.006	−.007	.02

CLASSIC-UB initial/final vs. Baseline	Syllable	.026	.001	.046
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.019	−.002	.038
CLASSIC-UB initial/final vs. FTP	Syllable	.007	−.023	.036
PUDDLE vs. Baseline	Syllable	.141	.113	.164
PUDDLE vs. CLASSIC-UB final	Syllable	.135	.110	.155
PUDDLE vs. FTP	Syllable	.122	.079	.155
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.115	.087	.142
PUDDLE vs. BTP	Syllable	.049	−.002	.090

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted neighbourhood density when phonemic or syllabified input is used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S10: Approximation of Child Production Vocabulary by Weighted Phonotactic Probability

We ran analyses on both phonemic and syllabified input. We have provided a narrative account of the phonemic-input analysis in the section Results and Discussion / Word-Level Measures / Phonotactic Probability of Chapter 2.

When we used syllabified input, we found no significant differences between models' performance at approximating children's vocabulary by weighted phonotactic probability (see Figure S10.1, Table S10.1, and Table S10.2).

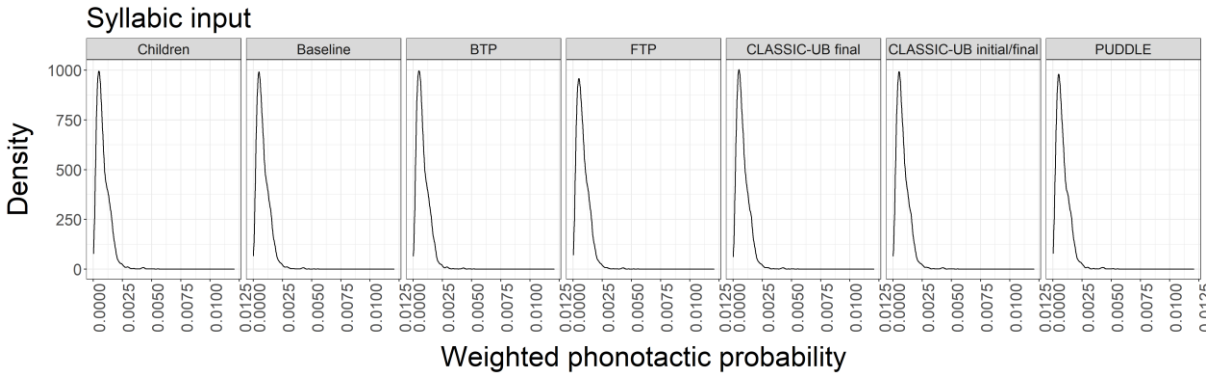


Figure S10.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted phonotactic probability. Syllabified input was used.

Table S10.1 Child-model comparison by weighted phonotactic probability

Model comparison	Input type	*D*	*p*	95% CI *LL*	95% CI *UL*
Children vs. Baseline	Phoneme	.05	.002	.03	.08
Children vs. BTP	Phoneme	.08	< .001	.05	.12
Children vs. FTP	Phoneme	.08	< .001	.05	.11
Children vs. CLASSIC-UB final	Phoneme	.07	< .001	.05	.10
Children vs. CLASSIC-UB initial/final	Phoneme	.09	< .001	.06	.12
Children vs. PUDDLE	Phoneme	.05	< .001	.03	.08
Children vs. Baseline	Syllable	.02	.770	.01	.04
Children vs. BTP	Syllable	.01	.770	.01	.03
Children vs. FTP	Syllable	.02	.368	.01	.05
Children vs. CLASSIC-UB final	Syllable	.02	.368	.01	.04
Children vs. CLASSIC-UB initial/final	Syllable	.03	.044	.02	.06
Children vs. PUDDLE	Syllable	.02	.770	.01	.04

Note. Model distributions of unique words by weighted phonotactic probability were compared to child distribution. The 95% bootstrap confidence intervals of the statistic were adjusted using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S10.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S10.1

Model comparison	Input type	Δ*D*	95% CI	
			LL	*UL*
BTP vs. Baseline	Phoneme	.031	−.001	.073
BTP vs. PUDDLE	Phoneme	.028	−.021	.072
BTP vs. CLASSIC-UB final	Phoneme	.007	−.029	.045
BTP vs. FTP	Phoneme	.002	−.027	.033
FTP vs. Baseline	Phoneme	.029	−.007	.064
FTP vs. PUDDLE	Phoneme	.026	−.018	.078
FTP vs. CLASSIC-UB final	Phoneme	.005	−.032	.034
CLASSIC-UB final vs. Baseline	Phoneme	.024	−.014	.059
CLASSIC-UB final vs. PUDDLE	Phoneme	.021	−.028	.075
CLASSIC-UB initial/final vs. Baseline	Phoneme	.042	.008	.081
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.038	−.011	.098
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.017	−.014	.050
CLASSIC-UB initial/final vs. FTP	Phoneme	.012	−.022	.045
CLASSIC-UB initial/final vs. BTP	Phoneme	.01	−.032	.050
PUDDLE vs. Baseline	Phoneme	.003	−.038	.042
Baseline vs. BTP	Syllable	.004	−.016	.024
FTP vs. BTP	Syllable	.009	−.014	.030
FTP vs. Baseline	Syllable	.005	−.017	.023
FTP vs. PUDDLE	Syllable	.003	−.018	.019
CLASSIC-UB final vs. BTP	Syllable	.01	−.014	.028
CLASSIC-UB final vs. Baseline	Syllable	.006	−.017	.022
CLASSIC-UB final vs. PUDDLE	Syllable	.003	−.017	.020

CLASSIC-UB final vs. FTP	Syllable	.001	−.015	.015
CLASSIC-UB initial/final vs. BTP	Syllable	.017	−.017	.037
CLASSIC-UB initial/final vs. Baseline	Syllable	.013	−.008	.034
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.011	−.012	.032
CLASSIC-UB initial/final vs. FTP	Syllable	.008	−.015	.029
CLASSIC-UB initial/final vs. CLASSIC-UB final	Syllable	.008	−.014	.027
PUDDLE vs. BTP	Syllable	.006	−.021	.027
PUDDLE vs. Baseline	Syllable	.002	−.017	.022

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted phonotactic probability when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S11: CLASSIC-UB Initial-Final Versus CLASSIC-UB Final

In this section, we briefly discuss whether the comparison between CLASSIC-UB initial-final and CLASSIC-UB final (on all the measures considered in our study) suggested that the addition of utterance-initial markers improved model performance.

As can Figure 2 in the main text and Appendix S4 indicate, CLASSIC-UB initial-final showed a better performance than did CLASSIC-UB final in the traditional measures, reaching .50 for precision and .50 for recall with phonemic input (vs. .49 for precision and .45 for recall), and .66 for precision and .58 for recall with syllabified input (vs. .57 for precision and .48 for recall). This suggested that the inclusion of initial (in addition to final) utterance-boundary markers was useful in segmenting the speech input as other studies have shown (Seidl & Johnson, 2006, 2008).

However, results for the developmental measures suggested that utterance-initial markers did not significantly improve model performance. CLASSIC-UB initial-final did not explain more variability in child age of first production compared to CLASSIC-UB final, suggesting that an initial utterance-boundary marker might not be necessary for predicting word age of first production (see Table S6.1). Similarly, adding utterance-initial markers did not significantly improve the model's ability to capture any of the word-level characteristics of children's vocabularies (see Tables S7.2, S8.2, S9.2, and S10.2).

When measures that are not weighted by input frequency are considered (i.e., traditional measures, unweighted age of first production, word-level measures; see Appendices S4 and S6–S10), this result was likely due to the ratio of type to token frequency of the words present in the input at utterance-initial and final position. Namely, token frequency (i.e., frequency of a word including repetitions) was lower for words appearing at the end of utterances, $M = 305.35$, $SE = 28.81$, than were words appearing at the start of utterances, $M = 652.05$, $SE = 62.81$. At the same time, the input contained higher type frequency (i.e., more different words) at the end of utterances ($N = 5{,}485$) than at the beginning ($N = 786$). This

suggested that CLASSIC-UB's segmentation accuracy increased when provided with utterance-initial markers because there were more repeated words that the model was able to segment correctly at the start of utterances, but their role became marginal for building a lexicon as the majority of novel words appeared at utterance ends (e.g., Fernald & Mazzie, 1991).

This result provided evidence in support of previous work (e.g., Pearl et al., 2010) suggesting that utterance-initial words might be segmented with higher accuracy because they have a higher token frequency (e.g., pronouns, determiners) than have more variable utterance-final ones (e.g., nouns, verbs). Additionally, using measures based on child data, we showed that the high type frequency of utterance-final words might be important in the process of building a lexicon from the segmented words. In other words, even if the perceptual salience of word boundaries at utterance-initial and final edges equally facilitates word extraction in the laboratory (Seidl & Johnson, 2006; 2008), their role in the naturalistic environment might be moderated by frequency information. The repeated presentation of few different words in utterance-initial position might increase the likelihood of segmenting those words correctly. Conversely, encountering a large number of different words at utterance ends might increase the chance of building a more diverse (i.e., larger) vocabulary. Finally, this also means that facilitatory effects of utterance boundaries in naturalistic settings might be different for languages where, for example, new words do not tend to be placed at utterance ends as in English child-directed speech (e.g., Dutch, Japanese; Han et al., 2021).

References

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology, 27*(2), 209–221. https://doi.org/10.1037/0012-1649.27.2.209

Han, M., de Jong, N. H., & Kager, R. (2021). Language specificity of infant-directed speech: Speaking rate and word position in word-learning contexts. *Language*

Learning and Development, 17(3), 221–240. https://doi.org/10.1080/15475441.2020.1855182

Phillips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*(8), 1824–1854. https://doi.org/10.1111/cogs.12217

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

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of child language*, *35*(1), 1–24. https://doi.org/10.1017/S0305000907008215

Appendix S12: Does PUDDLE Represent a Child With More Advanced Vocabulary Knowledge?

The difference between CLASSIC-UB and PUDDLE in the word-level measures (i.e., with CLASSIC-UB better approximating children's vocabularies by phonemic length and neighbourhood density) might be explained by differences in vocabulary size. At the end of learning, PUDDLE had a larger vocabulary than CLASSIC-UB, and might be taken to represent a child with more advanced vocabulary knowledge. Conversely, it is possible that an earlier stage of PUDDLE with smaller vocabulary might show similar performance to CLASSIC-UB on our developmental measures. To assess this possibility, we were able to look at the models' developmental cascades to see whether model differences still held when we considered the stage at which PUDDLE had reached a vocabulary equal in size to that of CLASSIC-UB. We carried out this analysis only for phonemic input because CLASSIC-UB developed a smaller vocabulary than PUDDLE only when using phonemic input but not when using syllabified input (see Table S12.1).

Table S12.1 Raw number of word types learned by CLASSIC-UB models and PUDDLE when run on phonemic or syllabic input, ranked from largest to smallest.

Model	Input type	Word types learned
CLASSIC-UB final	Syllables	8,047
CLASSIC-UB initial/final	Syllables	7,451
PUDDLE	Syllables	5,903
PUDDLE	Phonemes	3,967
CLASSIC-UB final	Phonemes	3,611
CLASSIC-UB initial/final	Phonemes	3,049

In Figure S12.1, black vertical lines indicate the stage at which PUDDLE had reached a vocabulary size equal to that of CLASSIC-UB final or that of CLASSIC-UB initial/final (as indicated by the text labels). If differences between models are explained by vocabulary size, PUDDLE word-level distributions at the vertical lines should become similar to CLASSIC-UB's distributions at stage 20 (i.e., at the end of its learning). Instead, for those measures that were found to show significant differences, that is, phonemic length and neighbourhood density, differences between PUDDLE and CLASSIC-UB models held across stages, with PUDDLE's learning being always biased toward short (three-phoneme and four-phoneme) words and high-neighbourhood words compared to CLASSIC-UB models.

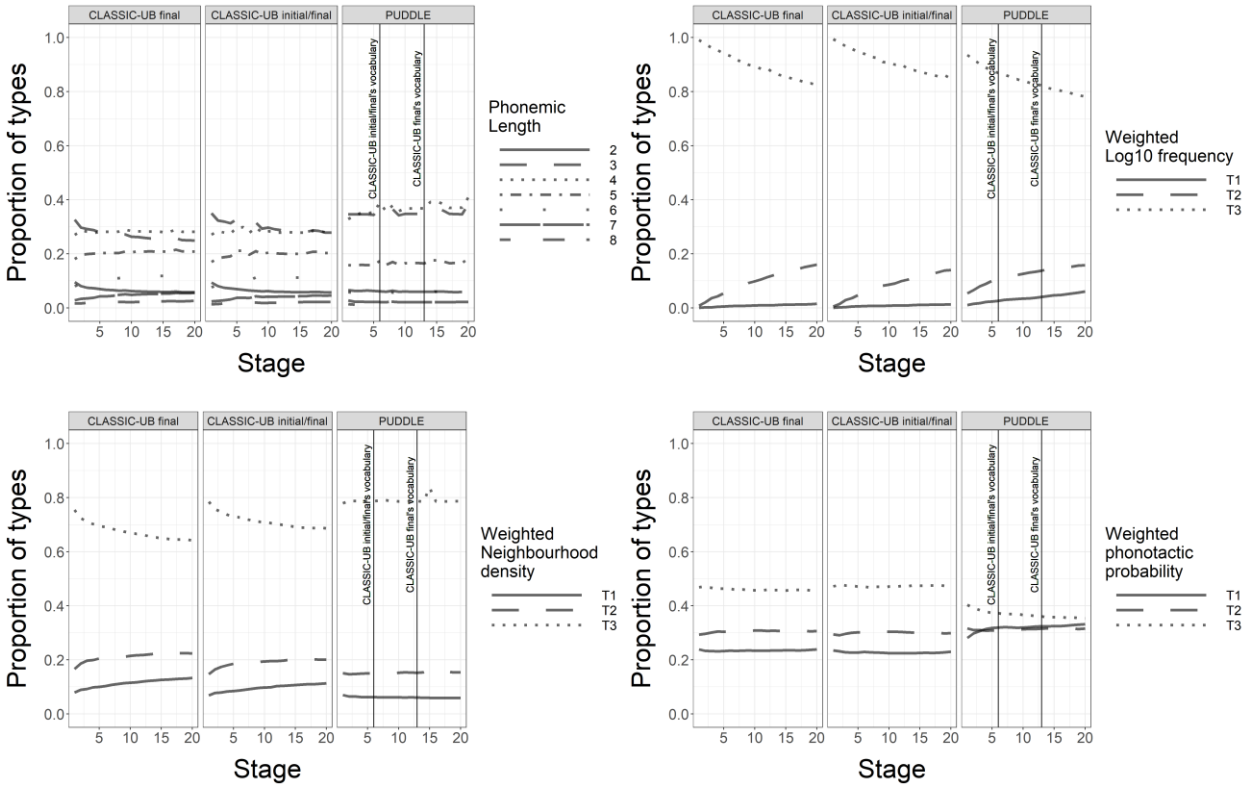


Figure S12.1 Proportion of types discovered at each input stage for each word-level measure. Proportion of types was computed by dividing the cumulative number of word types by the total number of types at a specific stage. Stage was computed by dividing the segmented utterances into 20 equal stages; the 604 stages used for the precision and recall measures were divided into wider stages, 20, because the probability of discovering new word types decreases substantially at later stages. For

continuous word-level measures (i.e., word frequency, neighbourhood density, and phonotactic probability), word types were divided into groups based on child-directed speech tertiles. For example, T1 in the word frequency measure identifies words that have a low frequency in child-directed speech (<= 33rd percentile), while T3 refers to high-frequency words in child-directed speech (> 66th percentile). Black vertical lines indicate the stages at which PUDDLE has reached a vocabulary size equal to CLASSIC-UB final or CLASSIC-UB initial/final.

As we discussed in the main paper (see Measures of Developmental Plausibility section of the Discussion), differences in performance can be explained by CLASSIC-UB's ability to learn words with overlapping phonological sequences (see Jones, 2016). Indication of this can be seen when one looks at the length and neighbourhood findings separately by word frequency. In Figure S12.2 below, one can see that CLASSIC-UB became more accurate at capturing child vocabularies as frequency increased. This happened because frequent words are more likely to share phonological sequences with previously learned words, consequently boosting CLASSIC-UB's learning compared to other models which do not show such facilitation (as their learning mechanism is uniquely based on tracking target sequences' frequency). Namely, other models' performance at capturing child phonemic length and neighbourhood density did not improve as word frequency increased.

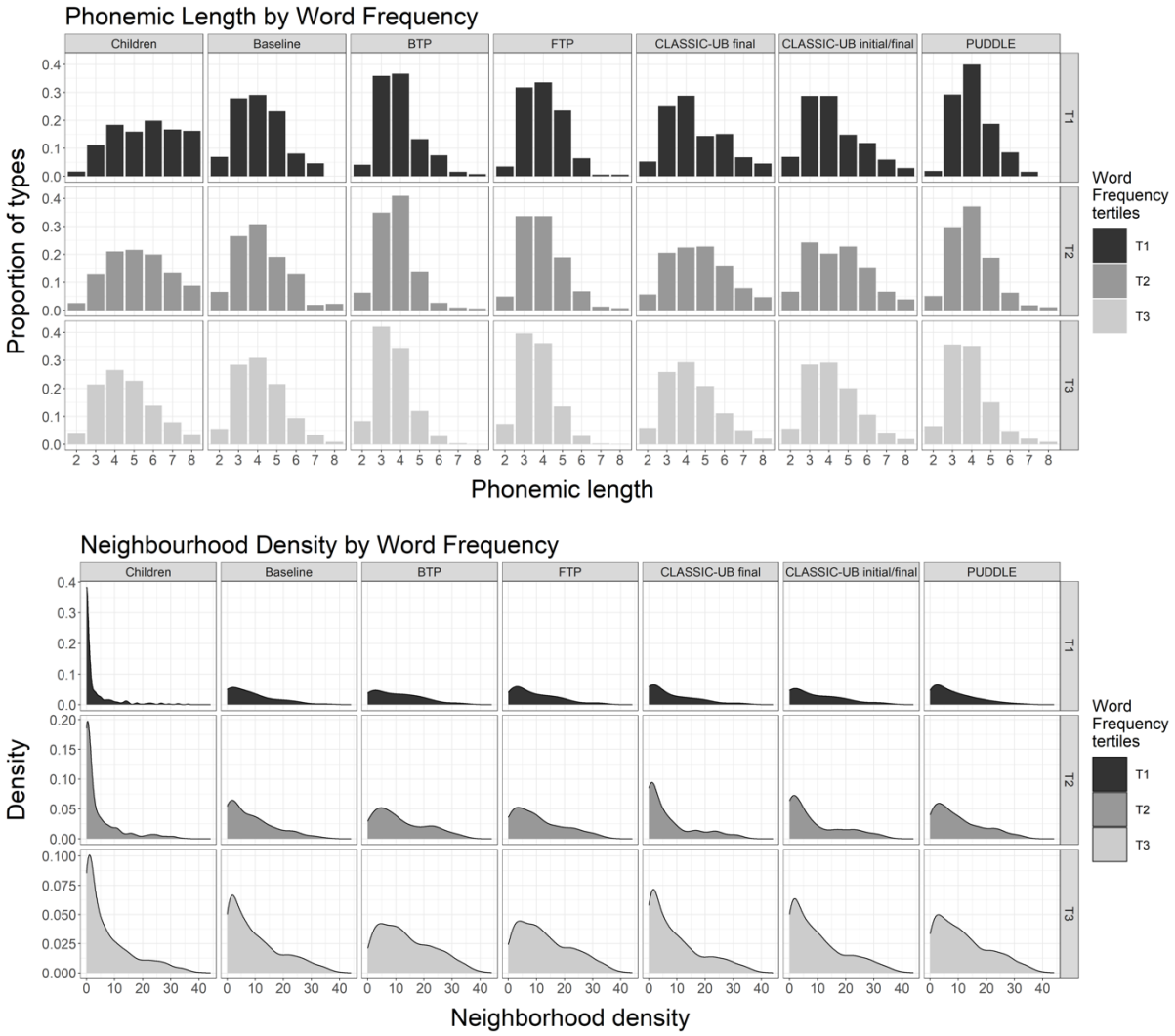


Figure S12.2 Child and models' phonemic length and neighbourhood density distributions at different child-directed word frequency tertiles.

We also conducted a final exploratory analysis to support our claim that CLASSIC-UB captures long and low-neighbourhood words from the child vocabularies better than PUDDLE. Specifically, we wanted to check whether CLASSIC-UB actually learned more long and low-neighbourhood words than did PUDDLE rather than it simply missing a portion of children's short, high-neighbourhood words (that PUDDLE instead captured), producing in turn an increase in the relative proportion of long, low-neighbourhood words in its vocabulary.

Thus, we looked at the absolute number of children's word types captured by each model, as shown in Figure S12.3. In this figure, we plotted the raw number of

types produced by children, alongside the number of children's words that CLASSIC-UB final or PUDDLE captured or missed (by phonemic length and neighbourhood density). This analysis excluded a portion of words that the models learned from the input but that were not produced by children; when including this set of words, the results that we obtained were consistent with the analysis reported below. As Figure S12.3 shows, differences in phonemic length were not due only to the fact that PUDDLE captured more three- and four-phoneme children's words than did CLASSIC-UB but also to the fact that CLASSIC-UB captured a higher absolute number of five- to eight-phoneme words than did PUDDLE. Similarly, although PUDDLE captured a higher number of high-neighbourhood words (T3), it also captured a lower absolute number of words in the low and middle neighbourhood range (T1 and T2) than did CLASSIC-UB. In sum, this analysis supported our claim that CLASSIC-UB's learning mechanism facilitates the learning of words that are generally more difficult to learn (i.e., long and with a low number of similar words in the input) but that children nevertheless acquire.

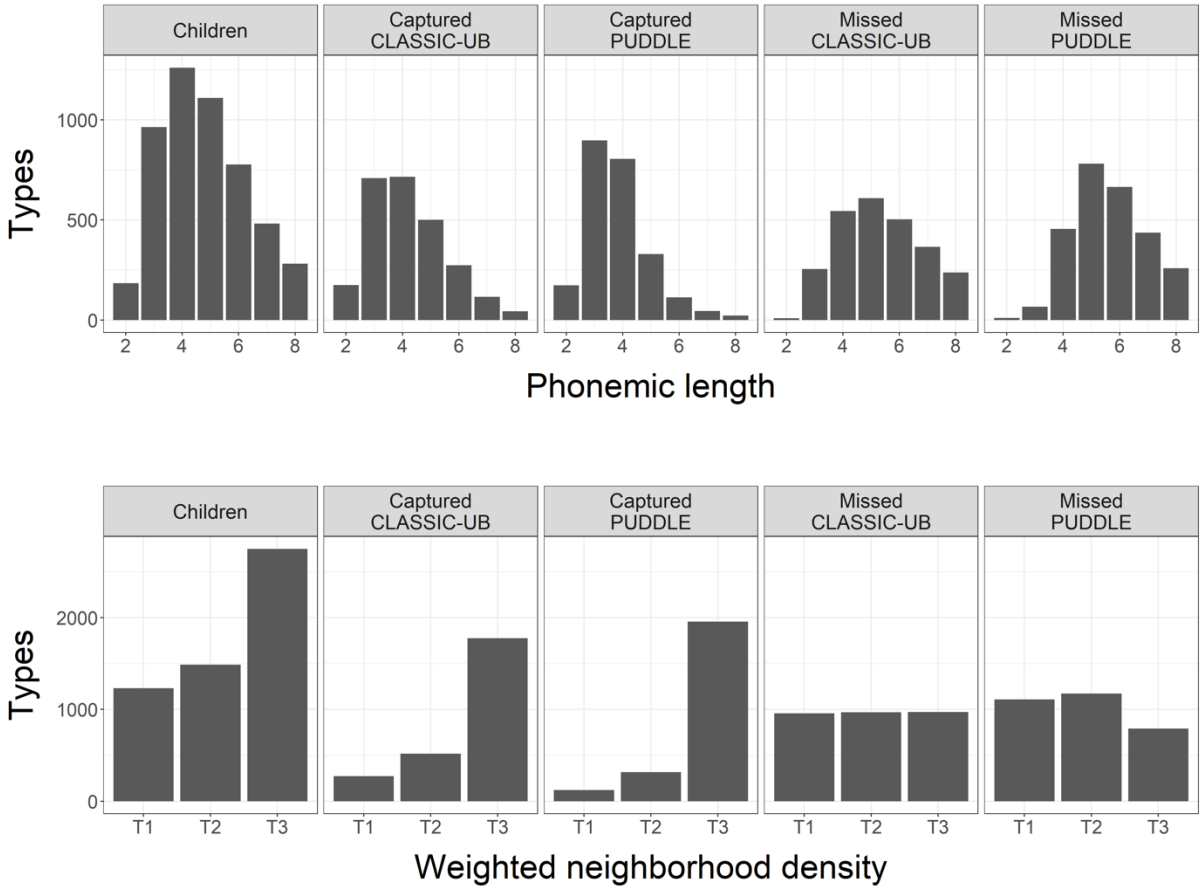


Figure S12.3 The plot shows the raw number of word types produced by children alongside the raw number of word types produced by children that CLASSIC-UB final and PUDDLE learned (captured) or not learned (missed). Phonemic length considers children's words from two to eight phonemes, while weighted neighbourhood density considers children's words in low (T1), middle (T2) and high (T3) neighbourhood child-directed speech tertiles.

References

Jones, G. (2016). The influence of children's exposure to language from two to six years: The case of nonword repetition. *Cognition, 153*, 79–88. https://doi.org/10.1016/j.cognition.2016.04.017

Appendix S13: Controlling for Baseline Segmentation Performance

An unexpected finding of our study was that, when we used syllabified input, no model was able to outperform the baseline in developmental measures. Providing a model with the input syllabic structure likely represents a strong facilitation which makes it difficult to compare competing models. First, given that models' input contains 81% of monosyllabic tokens, syllabifying the input (i.e., avoiding oversegmentation of syllables) allows a model to discover—by chance—a large proportion of word types. For example, although models were exposed to limited input compared to what children receive, when processing syllabified input they discovered more word types, M = 7,223, min = 5903, max = 8,047, than did Thomas (the child with the largest production vocabulary; N = 5,899). The models also learned more low-frequency words than did children when run on a syllabified input (see Figure S8.1 in the Appendix S8), and this may have been for the same reason.

Furthermore, previous computational work has shown that providing chunking models with the input syllabic structure might not be necessary, as models that are run on phonemic input only commit a small proportion of intrasyllabic segmentation error (Goldwater et al., 2009). To confirm this, more work that compares models and infants' actual segmentation performance is needed. For example, future work could investigate whether the issue with the syllabic baseline applies cross-linguistically or is only present in languages such as English that have a large number of monosyllabic words.

References

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

Appendix S14: Morphological Analysis Excluding Words with Multiple Morpheme Segmentations

Figure S14.1 and S14.2 display the number of morpheme tokens and types discovered by each segmentation model in the study, when excluding words that can have alternative morphological segmentations based on part-of-speech. The models' performance is similar to that found when models were evaluated on all the words available in the input (see Figure 8b and Figure 12 in the thesis).

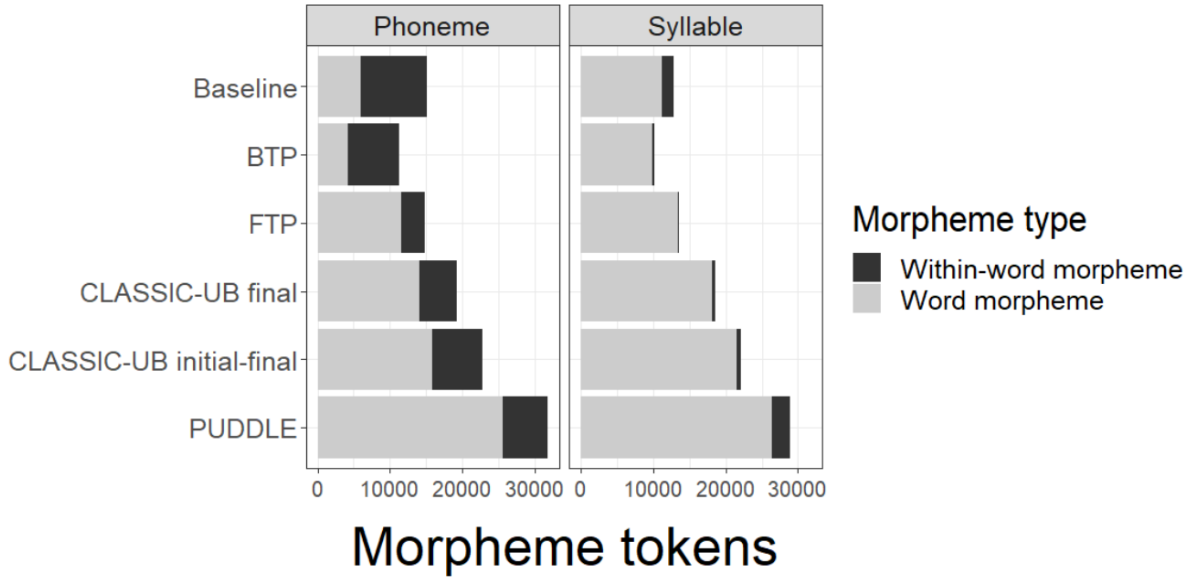


Figure S14.1 Number of morpheme tokens segmented by Baseline, backward transitional probability (BTP), forward transitional probability (FTP), CLASSIC-UB final, CLASSIC-UB initial-final, and PUDDLE. Tokens are grouped by morphemes that appear within words (Within-word morpheme) and morphemes that correspond to words (Word morpheme). Morpheme tokens are shown for models run on phonemic or syllabic input.

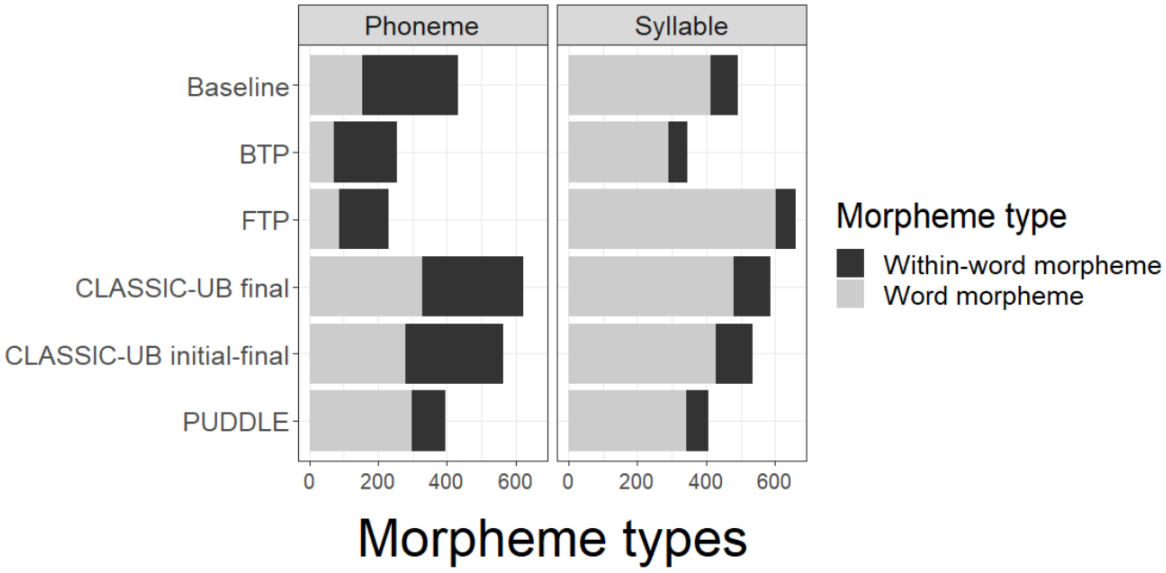


Figure S14.2 Number of morpheme types segmented by Baseline, backward transitional probability (BTP), forward transitional probability (FTP), CLASSIC-UB final, CLASSIC-UB initial-final, and PUDDLE. Types are grouped by morphemes that appear within words (Within-word morpheme) and morphemes that correspond to words (Word morpheme). Morpheme types are shown for models run on phonemic or syllabic input.

Appendix S15: Age-based Age of First Production Measure

In this section, we report results of the age of first production analysis but computing word stage of first production as the lowest age in months (instead mean length of utterance, MLU) of a transcript in which a target word type appeared. Differently from the measure included in Chapter 3, this age-based measure focuses on age rather than on linguistic competence. As shown in Table S15.1 and S15.2, results obtained with this age-based measure are similar to that found with the age of first production measure based on linguistic competence (see Table 3a in the thesis, and Table S17.1 in Appendix S17). This suggests that differences obtained between the CDI-based age of acquisition measure and the MLU-based age of first production measure in Chapter 3 are likely not determined by the fact that the two measures focus on age and linguistic competence respectively.

Table S15.1 Adjusted R^2 for linear regression models predicting *age-based* word age of first production as a function of weighted log10 number of times a word was correctly segmented by each model

	Phonemic input			Syllabified input		
		95% CI			95% CI	
Model	$R^2_{adjusted}$	*LL*	*UL*	$R^2_{adjusted}$	*LL*	*UL*
Baseline	.049	.021	.084	.041	.016	.074
Backward transitional probability	.028	.012	.048	.017	.006	.034
Forward transitional probability	.054	.028	.088	.003	0	.011
CLASSIC-UB final	**.113**	.074	.162	.07	.039	.105
CLASSIC-UB initial/final	.108	.069	.155	**.066**	.039	.1
PUDDLE	.043	.019	.071	.041	.019	.069

Note. Heteroskedasticity-robust standard errors were computed using a HC2 estimator. The 95% confidence intervals indicate lower and upper limits of bootstrap confidence intervals around the estimate (based on 1,000 iterations). Holm's correction was applied to the confidence intervals.

Table S15.2 Pairwise differences between adjusted R^2 of weighted *age-based* age of first production models

Model comparison	Input type	ΔR^2	95% CI	
			LL	*UL*
Baseline vs. BTP	Phoneme	.021	-.007	.052
Baseline vs. PUDDLE	Phoneme	.006	-.031	.047
FTP vs. BTP	Phoneme	.026	-.001	.061
FTP vs. PUDDLE	Phoneme	.011	-.031	.044
FTP vs. Baseline	Phoneme	.005	-.028	.041
CLASSIC-UB final vs. BTP	Phoneme	.085	.041	.131
CLASSIC-UB final vs. PUDDLE	Phoneme	.07	.026	.128
CLASSIC-UB final vs. Baseline	Phoneme	.064	.01	.131
CLASSIC-UB final vs. FTP	Phoneme	.059	.00	.113
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.005	-.014	.023
CLASSIC-UB initial/final vs. BTP	Phoneme	.08	.038	.133
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.065	.019	.113
CLASSIC-UB initial/final vs. Baseline	Phoneme	.059	.004	.122
CLASSIC-UB initial/final vs. FTP	Phoneme	.054	.005	.111
PUDDLE vs. BTP	Phoneme	.015	-.012	.05
Baseline vs. FTP	Syllable	.038	.009	.069
Baseline vs. BTP	Syllable	.024	-.006	.059
Baseline vs. PUDDLE	Syllable	.00	-.024	.025

BTP vs. FTP	Syllable	.014	-.004	.037
CLASSIC-UB final vs. FTP	Syllable	.067	.035	.111
CLASSIC-UB final vs. BTP	Syllable	.053	.018	.093
CLASSIC-UB final vs. Baseline	Syllable	.029	-.021	.074
CLASSIC-UB final vs. PUDDLE	Syllable	.029	-.013	.063
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.004	-.01	.018
CLASSIC-UB initial/final vs. FTP	Syllable	.063	.031	.098
CLASSIC-UB initial/final vs. BTP	Syllable	.049	.011	.092
CLASSIC-UB initial/final vs. Baseline	Syllable	.025	-.02	.069
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.025	-.013	.06
PUDDLE vs. FTP	Syllable	.038	.012	.07
PUDDLE vs. BTP	Syllable	.024	-.007	.057

Note. ΔR^2 = difference between adjusted R^2 values. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S16: Comparison of Precision and Recall Measures

We have included a narrative account of the findings in Table S16.1, S16.2, and S16.3 in Chapter 3, in the section Results / Precision and Recall.

Table S16.1 Comparison of models for the precision and recall measures for phonemic input

Model comparison	Measure	M1[a]	M2[b]	ΔM[c]	*t*	*p*	*df*	95% CI	
								LL	*UL*
BTP vs. PUDDLE	Recall	.11	.58	-.48	-120	<.001	18,612.8	-.488	-.462
BTP vs. PUDDLE	Precision	.09	.49	-.4	-98.92	<.001	17,504	-.411	-.386
BTP vs. CLASSIC-UB initial/final	Precision	.09	.44	-.35	-81.51	<.001	16,803	-.357	-.333
BTP vs. CLASSIC-UB initial/final	Recall	.11	.45	-.34	-80.04	<.001	17,435.3	-.357	-.331
BTP vs. CLASSIC-UB final	Precision	.09	.43	-.34	-79.26	<.001	16,628	-.352	-.326
BTP vs. CLASSIC-UB final	Recall	.11	.43	-.32	-72.85	<.001	17,245.9	-.33	-.304
BTP vs. FTP	Recall	.11	.31	-.2	-52.17	<.001	19,129.7	-.21	-.188
BTP vs. FTP	Precision	.09	.24	-.15	-41.77	<.001	19,231	-.159	-.139
BTP vs. Baseline	Recall	.11	.13	-.02	-5.33	<.001	19,984.9	-.026	-.009
BTP vs. Baseline	Precision	.09	.09	0	.8	.617	19,827	-.004	.009

FTP vs. PUDDLE	Recall	.31	.58	-.28	-63.84	<.001	19,917.9	-.291	-.263
FTP vs. PUDDLE	Precision	.24	.49	-.25	-57.76	<.001	19,285	-.265	-.236
FTP vs. Baseline	Recall	.31	.13	.18	47.97	<.001	18,931.6	.169	.194
FTP vs. Baseline	Precision	.24	.09	.15	43.98	<.001	18,473	.14	.162
CLASSIC-UB final vs. Baseline	Precision	.43	.09	.34	81.76	<.001	15,724	.329	.357
CLASSIC-UB final vs. Baseline	Recall	.43	.13	.3	69.29	<.001	16,990.2	.287	.315
CLASSIC-UB final vs. FTP	Precision	.43	.24	.19	41.86	<.001	18,589	.178	.205
CLASSIC-UB final vs. PUDDLE	Recall	.43	.58	-.16	-32.7	<.001	19,601.8	-.173	-.143
CLASSIC-UB final vs. FTP	Recall	.43	.31	.12	25.3	<.001	19,200.8	.106	.132
CLASSIC-UB final vs. PUDDLE	Precision	.43	.49	-.06	-11.86	<.001	19,846	-.073	-.045
CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.43	.45	-.03	-5.24	<.001	19,990.8	-.04	-.014
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.43	.44	-.01	-1.02	.617	19,992	-.016	.005
CLASSIC-UB initial/final vs. Baseline	Precision	.44	.09	.35	84.12	<.001	15,890	.335	.361
CLASSIC-UB initial/final vs. Baseline	Recall	.45	.13	.33	76.5	<.001	17,180.2	.314	.341
CLASSIC-UB initial/final vs. FTP	Precision	.44	.24	.2	43.49	<.001	18,741	.184	.209
CLASSIC-UB initial/final vs. FTP	Recall	.45	.31	.15	31.34	<.001	19,332.9	.133	.159

CLASSIC-UB initial/final vs. PUDDLE	Recall	.45	.58	-.13	-27.44	<.001	19,697.6	-.145	-.117
CLASSIC-UB initial/final vs. PUDDLE	Precision	.44	.49	-.05	-10.91	<.001	19,900	-.068	-.04
PUDDLE vs. Baseline	Recall	.58	.13	.46	116.56	<.001	18,383.9	.445	.47
PUDDLE vs. Baseline	Precision	.49	.09	.4	102.27	<.001	16,574	.388	.414

Note. Pairwise comparisons via Welch's *t* test for unequal variances; *p* values and bootstrap 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

[a]M1 = first model mean. [b]M2 = second model mean. [c]ΔM = mean difference.

Table S16.2 Comparison of precision and recall measures for syllabified input

								95% CI	
Model comparison	Measure	M1[a]	M2[b]	ΔM[c]	*t*	*p*	*df*	*LL*	*UL*
BTP vs. PUDDLE	Recall	.37	.64	-.27	-56.73	<.001	19,387.74	-.285	-.255
BTP vs. CLASSIC-UB initial/final	Recall	.37	.6	-.24	-46.6	<.001	19,979.91	-.253	-.222
BTP vs. CLASSIC-UB initial/final	Precision	.42	.63	-.21	-40.82	<.001	19,940.213	-.222	-.19
BTP vs. CLASSIC-UB final	Recall	.37	.55	-.18	-34.66	<.001	19,997.43	-.195	-.164
BTP vs. CLASSIC-UB final	Precision	.42	.59	-.18	-34.27	<.001	19,983.493	-.192	-.16
BTP vs. FTP	Precision	.42	.57	-.15	-29.02	<.001	19,964.685	-.163	-.132
BTP vs. FTP	Recall	.37	.5	-.13	-25.81	<.001	19,997.92	-.15	-.119

BTP vs. PUDDLE	Precision	.42	.55	-.13	-25.8	<.001	19,799.583	-.14	-.115
BTP vs. Baseline	Precision	.42	.3	.12	24.29	<.001	19,863.693	.108	.136
BTP vs. Baseline	Recall	.37	.3	.06	12.89	<.001	19,898.38	.052	.078
FTP vs. Baseline	Precision	.57	.3	.27	55.06	<.001	19,963.588	.254	.283
FTP vs. Baseline	Recall	.5	.3	.2	39.61	<.001	19,904.01	.184	.212
FTP vs. PUDDLE	Recall	.5	.64	-.14	-28.79	<.001	19,400.90	-.149	-.121
FTP vs. PUDDLE	Precision	.57	.55	.02	4.12	<.001	19,927.452	.008	.031
CLASSIC-UB final vs. Baseline	Precision	.59	.3	.3	60.33	<.001	19,936.794	.284	.311
CLASSIC-UB final vs. Baseline	Recall	.55	.3	.24	48.71	<.001	19,882.90	.23	.258
CLASSIC-UB final vs. PUDDLE	Recall	.55	.64	-.09	-18.96	<.001	19,352.72	-.104	-.078
CLASSIC-UB final vs. CLASSIC-UB initial/final	Recall	.55	.6	-.06	-11.3	<.001	19,972.91	-.071	-.044
CLASSIC-UB final vs. PUDDLE	Precision	.59	.55	.05	9.81	<.001	19,890.950	.036	.061
CLASSIC-UB final vs. FTP	Recall	.55	.5	.05	8.95	<.001	19,996.90	.033	.062
CLASSIC-UB final vs. CLASSIC-UB initial/final	Precision	.59	.63	-.03	-6.19	<.001	19,983.508	-.043	-.019
CLASSIC-UB final vs. FTP	Precision	.59	.57	.03	5.56	<.001	19,994.125	.015	.041
CLASSIC-UB initial/final vs. Baseline	Precision	.63	.3	.33	67.55	<.001	19,981.746	.313	.341
CLASSIC-UB initial/final vs. Baseline	Recall	.6	.3	.3	61.36	<.001	19,964.85	.286	.316

CLASSIC-UB initial/final vs. FTP	Recall	.6	.5	.1	20.45	<.001	19,982.28	.089	.119
CLASSIC-UB initial/final vs. PUDDLE	Precision	.63	.55	.08	16.37	<.001	19,954.809	.064	.092
CLASSIC-UB initial/final vs. FTP	Precision	.63	.57	.06	11.87	<.001	19,994.615	.044	.072
CLASSIC-UB initial/final vs. PUDDLE	Recall	.6	.64	-.03	-7.03	<.001	19,568.70	-.045	-.02
PUDDLE vs. Baseline	Recall	.64	.3	.33	73.15	<.001	19,767.36	.322	.348
PUDDLE vs. Baseline	Precision	.55	.3	.25	52.53	<.001	19,991.504	.235	.264

Note. Pairwise comparisons via Welch's *t* test for unequal variances; *p* values and bootstrap 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

[a]M1 = first model mean. [b]M2 = second model mean. [c]ΔM = mean difference.

Table S16.3 For each model, comparison of phonemic vs. syllabic model implementation (Δt), in Precision and Recall

			95% CI	
Measure	Model	Δt	*LL*	*UL*
Precision	BTP	-23.49	-27.17	-19.32
Recall	BTP	-18.22	-21.69	-14.43
Precision	FTP	-11.08	-14.32	-7.59
Recall	FTP	8.36	5.59	11.25
Precision	CLASSIC-UB final	21.43	15.94	25.98
Recall	CLASSIC-UB final	20.58	16.06	24.70

Precision	CLASSIC-UB initial-final	16.58	12.29	20.91
Recall	CLASSIC-UB initial-final	15.13	11.17	19.10
Precision	PUDDLE	49.74	45.50	54.34
Recall	PUDDLE	43.42	38.90	48.73

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S17: Age of First Production and Age of Acquisition Analyses: Pairwise Differences Between Models' Adjusted R^2

A narrative account of Table S17.1, S17.2, and S17.3 is available in section Results / Word Age of Acquisition and Production of Chapter 3.

Table S17.1 Age of first production analyses: Pairwise differences between adjusted R^2 values of models when phonemic or syllabified input was used

			95% CI	
Model comparison	Input type	ΔR^2	*LL*	*UL*
Baseline vs. BTP	Phoneme	.006	-.012	.026
FTP vs. BTP	Phoneme	.026	.002	.052
FTP vs. Baseline	Phoneme	.02	-.008	.047
FTP vs. PUDDLE	Phoneme	.016	-.013	.043
CLASSIC-UB final vs. BTP	Phoneme	.065	.027	.106
CLASSIC-UB final vs. Baseline	Phoneme	.059	.013	.102
CLASSIC-UB final vs. PUDDLE	Phoneme	.055	.012	.095
CLASSIC-UB final vs. FTP	Phoneme	.039	-.004	.081
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.011	-.006	.027
CLASSIC-UB initial/final vs. BTP	Phoneme	.054	.018	.094
CLASSIC-UB initial/final vs. Baseline	Phoneme	.048	.009	.093
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.044	.01	.081
CLASSIC-UB initial/final vs. FTP	Phoneme	.028	-.009	.07

PUDDLE vs. BTP	Phoneme	.01	-.015	.038
PUDDLE vs. Baseline	Phoneme	.004	-.02	.026
Baseline vs. FTP	Syllable	.032	.008	.062
Baseline vs. BTP	Syllable	.025	-.005	.058
Baseline vs. PUDDLE	Syllable	.01	-.019	.042
BTP vs. FTP	Syllable	.007	-.006	.024
CLASSIC-UB final vs. FTP	Syllable	.048	.024	.079
CLASSIC-UB final vs. BTP	Syllable	.041	.012	.078
CLASSIC-UB final vs. PUDDLE	Syllable	.026	-.009	.072
CLASSIC-UB final vs. Baseline	Syllable	.016	-.022	.053
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.005	-.006	.016
CLASSIC-UB initial/final vs. FTP	Syllable	.043	.019	.075
CLASSIC-UB initial/final vs. BTP	Syllable	.036	.006	.075
CLASSIC-UB initial/final vs. PUDDLE	Syllable	.021	-.015	.055
CLASSIC-UB initial/final vs. Baseline	Syllable	.011	-.023	.043
PUDDLE vs. FTP	Syllable	.022	.003	.047
PUDDLE vs. BTP	Syllable	.015	-.01	.042

Note. ΔR^2 = difference between adjusted R^2 values. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. FTP = forward transitional probability; BTP = backward transitional probability.

Table S17.2 Age of acquisition analyses: Pairwise differences between adjusted R^2 values of models when phonemic or syllabified input was used

			95% CI	
Model comparison	Input type	ΔR^2	*LL*	*UL*
Baseline vs. BTP	Phoneme	.022	-.027	.069
Baseline vs. CLASSIC-UB final	Phoneme	.021	-.044	.084
Baseline vs. CLASSIC-UB initial/final	Phoneme	.015	-.053	.08
Baseline vs. PUDDLE	Phoneme	.015	-.062	.075
Baseline vs. FTP	Phoneme	.01	-.035	.065
FTP vs. BTP	Phoneme	.012	-.029	.054
FTP vs. CLASSIC-UB final	Phoneme	.011	-.056	.071
FTP vs. CLASSIC-UB initial/final	Phoneme	.005	-.055	.063
FTP vs. PUDDLE	Phoneme	.005	-.065	.06
CLASSIC-UB final vs. BTP	Phoneme	.001	-.032	.04
CLASSIC-UB initial/final vs. BTP	Phoneme	.007	-.032	.077
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.006	-.01	.038
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.00	-.036	.039
PUDDLE vs. BTP	Phoneme	.007	-.029	.068
PUDDLE vs. CLASSIC-UB final	Phoneme	.006	-.042	.06
Baseline vs. FTP	Syllable	.00	-.014	.018
BTP vs. Baseline	Syllable	.003	-.021	.049
BTP vs. FTP	Syllable	.003	-.022	.052

CLASSIC-UB final vs. Baseline	Syllable	.006	-.025	.048
CLASSIC-UB final vs. FTP	Syllable	.006	-.02	.055
CLASSIC-UB final vs. BTP	Syllable	.003	-.036	.04
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.002	-.015	.022
CLASSIC-UB initial/final vs. Baseline	Syllable	.004	-.025	.041
CLASSIC-UB initial/final vs. FTP	Syllable	.004	-.021	.045
CLASSIC-UB initial/final vs. BTP	Syllable	.001	-.034	.031
PUDDLE vs. Baseline	Syllable	.013	-.033	.071
PUDDLE vs. FTP	Syllable	.013	-.023	.074
PUDDLE vs. BTP	Syllable	.01	-.055	.079
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.009	-.036	.066
PUDDLE vs. CLASSIC-UB final	Syllable	.007	-.047	.073

Note. ΔR^2 = difference between adjusted R^2 values. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. FTP = forward transitional probability; BTP = backward transitional probability.

Table S17.3 For each model, comparison of phonemic vs. syllabic model implementation (ΔR^2), in the age of first production measure. For each model implementation, the analysis controls for chance levels by initially subtracting a baseline AdjR^2 from a model AdjR^2 (see Chapter 3 for a detailed explanation). For each model, positive and negative ΔR^2 values indicate higher contribution of phonemic or syllabic input respectively. This analysis is carried out in Italian and English. English data are taken from Cabiddu et al. (2023)

			95% CI	

Model	Language	ΔR^2	*LL*	*UL*
BTP	Italian	.019	-.01	.05
FTP	Italian	.053	.016	.086
CLASSIC-UB final	Italian	.043	-.004	.096
CLASSIC-UB initial-final	Italian	.038	-.006	.085
PUDDLE	Italian	.014	-.011	.037
BTP	English	.049	.028	.069
FTP	English	.038	.016	.059
CLASSIC-UB final	English	.064	.038	.087
CLASSIC-UB initial-final	English	.051	.026	.075
PUDDLE	English	.022	.004	.039

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S18: Approximation of Child Production Vocabulary by Phonemic Length

Word types

We ran analyses on both phonemic and syllabified input. We have given a narrative account of the phonemic-input analysis in the section Results / Word-Level Characteristics of Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S18.1 Child-model comparison by phonemic length.

Comparison	Input type	χ^2	*df*	*p*	95% CI	
					LL	*UL*
Children vs. Baseline	Phoneme	286.02	9	<.001	198.62	479.33
Children vs. BTP	Phoneme	501.91	9	<.001	332.17	914.91
Children vs. FTP	Phoneme	694.18	9	<.001	513.31	1031.51
Children vs. CLASSIC-UB final	Phoneme	**17.35**	9	.13	9.06	78.77
Children vs. CLASSIC-UB initial/final	Phoneme	32.32	9	.001	13.19	111.01
Children vs. PUDDLE	Phoneme	127.97	9	<.001	59.51	235.55
Children vs. Baseline	Syllable	**7.22**	9	.614	7.22	50.86
Children vs. BTP	Syllable	26.23	9	.009	14.06	113.91
Children vs. FTP	Syllable	162.29	9	<.001	89.77	307.94
Children vs. CLASSIC-UB final	Syllable	25.76	9	.009	11.68	109.95

Children vs. CLASSIC-UB initial/final	Syllable	16.48	9	.13	10.16	86.05
Children vs. PUDDLE	Syllable	102.03	9	<.001	57.32	204.77

Note. We compared the probability of observing words of different phonemic lengths in the models' vocabularies against the expected probability of words being of a given phonemic length in children's vocabularies. Comparisons were tested via a chi-square goodness of fit test. The chi-square statistic always compares the distance of a model's distribution from children's. The table shows the type of comparison, the input type used, the chi-square statistic, degrees of freedom, *p* value and cut-offs of 95% bootstrap confidence interval of the statistic. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S18.2 Pairwise differences between the chi-square statistics reported in Table S18.1, comparing how well two models' observed probabilities of phonemic lengths fit children's expected probabilities, when phonemic or syllabified input is used

			95% CI	
Model comparison	Input type	$\Delta\chi^2$	*LL*	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	268.66	156.91	424.43
Baseline vs. CLASSIC-UB initial/final	Phoneme	253.69	140.96	425.89
Baseline vs. PUDDLE	Phoneme	158.05	5.93	404.26
BTP vs. CLASSIC-UB final	Phoneme	484.56	296.67	820.98
BTP vs. CLASSIC-UB initial/final	Phoneme	469.59	264.08	847.97
BTP vs. PUDDLE	Phoneme	373.95	172.61	763.75
BTP vs. Baseline	Phoneme	215.90	-6.81	551.59

FTP vs. CLASSIC-UB final	Phoneme	676.83	455.95	1049.33
FTP vs. CLASSIC-UB initial/final	Phoneme	661.86	441.92	964.69
FTP vs. PUDDLE	Phoneme	566.22	338.80	886.49
FTP vs. Baseline	Phoneme	408.17	167.05	752.40
FTP vs. BTP	Phoneme	192.27	-205.68	541.46
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	14.97	-27.99	79.39
PUDDLE vs. CLASSIC-UB final	Phoneme	110.62	21.97	201.48
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	95.65	-0.83	179.67
BTP vs. Baseline	Syllable	19.01	-40.18	90.20
BTP vs. CLASSIC-UB initial/final	Syllable	9.74	-52.91	67.60
BTP vs. CLASSIC-UB final	Syllable	0.47	-55.05	51.66
FTP vs. Baseline	Syllable	155.07	47.26	320.30
FTP vs. CLASSIC-UB initial/final	Syllable	145.81	42.18	293.69
FTP vs. CLASSIC-UB final	Syllable	136.54	35.28	274.75
FTP vs. BTP	Syllable	136.07	16.76	305.11
FTP vs. PUDDLE	Syllable	60.26	-57.63	203.29
CLASSIC-UB final vs. Baseline	Syllable	18.54	-24.02	83.59
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	9.27	-46.80	66.01
CLASSIC-UB initial/final vs. Baseline	Syllable	9.26	-22.84	63.60
PUDDLE vs. Baseline	Syllable	94.81	20.30	161.78
PUDDLE vs. CLASSIC-UB initial/final	Syllable	85.55	0.42	167.71
PUDDLE vs. CLASSIC-UB final	Syllable	76.28	-18.75	151.19

PUDDLE vs. BTP	Syllable	75.81	-16.85	179.44

Note. The $\Delta\chi^2$ measure examined whether two models' distributions were at the same distance from children's expected probabilities. The order of each pairwise difference was set as in the column Comparison (e.g., in Baseline vs. CLASSIC-UB final, the CLASSIC-UB final χ^2 estimate is subtracted from the Baseline χ^2 estimate). Lower and upper limits of bootstrap 95% confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

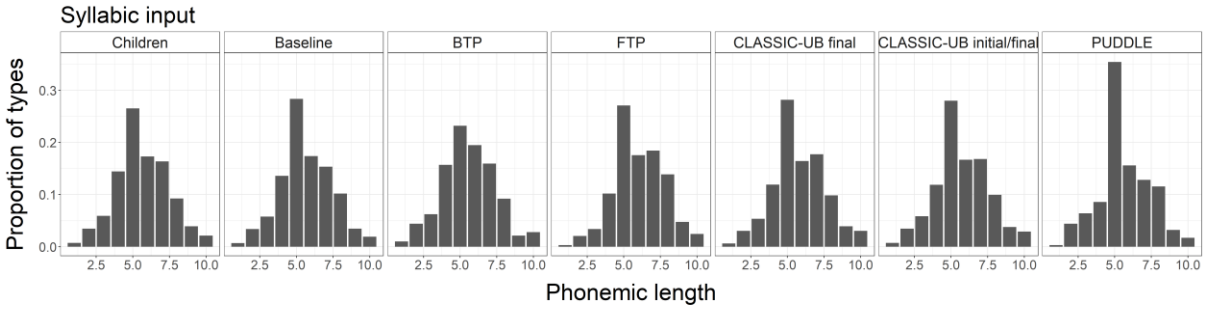


Figure S18.1 Proportion of unique words (types) produced by children and discovered by each model by phonemic length, when syllabified input was used.

Word tokens

This analysis is a repetition of the analysis done on word types above, but considering the distributions of word tokens in children and models. We have given a narrative account of this analysis in the section Exploratory Analysis of Word-Level Properties at the Token Level in Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S18.3 Child-model comparison by phonemic length.

			95% CI	

Comparison	Input type	χ^2	*df*	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	11724.67	9	<.001	10702.54	12888
Children vs. BTP	Phoneme	6858.66	9	<.001	6460.14	7305.29
Children vs. FTP	Phoneme	18378.04	9	<.001	17504.42	19470.99
Children vs. CLASSIC-UB final	Phoneme	2741.4	9	<.001	2455.01	3104.63
Children vs. CLASSIC-UB initial/final	Phoneme	**2549.08**	9	<.001	2243.2	2884.7
Children vs. PUDDLE	Phoneme	5451.78	9	<.001	4804.53	6297.6
Children vs. Baseline	Syllable	2572.99	9	<.001	2241.48	3001.58
Children vs. BTP	Syllable	4909.25	9	<.001	4320.07	5584.04
Children vs. FTP	Syllable	10276.23	9	<.001	9492.15	11293.92
Children vs. CLASSIC-UB final	Syllable	2761.34	9	<.001	2345.02	3319.55
Children vs. CLASSIC-UB initial/final	Syllable	**1107.39**	9	<.001	920.28	1411.46
Children vs. PUDDLE	Syllable	2977.77	9	<.001	2449.03	3582.16

Note. We compared the probability of observing words of different phonemic lengths in the models' vocabularies against the expected probability of words being of a

given phonemic length in children's vocabularies. Comparisons were tested via a chi-square goodness of fit test. The chi-square statistic always compares the distance of a model's distribution from children's. The table shows the type of comparison, the input type used, the chi-square statistic, degrees of freedom, *p* value and cut-offs of 95% bootstrap confidence interval of the statistic. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S18.4 Pairwise differences between the chi-square statistics reported in Table S18.3, comparing how well two models' observed probabilities of phonemic lengths fit children's expected probabilities, when phonemic or syllabified input is used

			95% CI	
Model comparison	Input type	$\Delta\chi^2$	*LL*	*UL*
Baseline vs. CLASSIC-UB initial/final	Phoneme	9175.60	7744.24	10593.40
Baseline vs. CLASSIC-UB final	Phoneme	8983.28	7606.19	10438.68
Baseline vs. PUDDLE	Phoneme	6272.89	4958.73	7517.70
Baseline vs. BTP	Phoneme	4866.01	3665.33	6120.34
BTP vs. CLASSIC-UB initial/final	Phoneme	4309.59	3719.21	4809.03
BTP vs. CLASSIC-UB final	Phoneme	4117.27	3523.86	4664.88
BTP vs. PUDDLE	Phoneme	1406.88	882.14	1991.45
FTP vs. CLASSIC-UB initial/final	Phoneme	15828.96	14589.53	16802.34
FTP vs. CLASSIC-UB final	Phoneme	15636.64	14481.46	16955.05
FTP vs. PUDDLE	Phoneme	12926.26	12091.99	13853.69
FTP vs. BTP	Phoneme	11519.37	10699.81	12438.44

FTP vs. Baseline	Phoneme	6653.36	5211.40	8126.65
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	192.32	-89.40	478.79
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	2902.70	2191.08	3621.60
PUDDLE vs. CLASSIC-UB final	Phoneme	2710.38	1845.95	3592.87
Baseline vs. CLASSIC-UB initial/final	Syllable	1465.61	811.95	2014.36
BTP vs. CLASSIC-UB initial/final	Syllable	3801.86	3158.87	4490.74
BTP vs. Baseline	Syllable	2336.25	1359.24	3261.62
BTP vs. CLASSIC-UB final	Syllable	2147.90	1422.09	2786.18
BTP vs. PUDDLE	Syllable	1931.48	838.41	2975.82
FTP vs. CLASSIC-UB initial/final	Syllable	9168.84	8314.60	10126.70
FTP vs. Baseline	Syllable	7703.23	6537.68	9045.58
FTP vs. CLASSIC-UB final	Syllable	7514.88	6826.85	8263.64
FTP vs. PUDDLE	Syllable	7298.46	6004.19	8603.68
FTP vs. BTP	Syllable	5366.98	4405.97	6167.21
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	1653.96	1268.49	2104.18
CLASSIC-UB final vs. Baseline	Syllable	188.35	-413.27	864.51
PUDDLE vs. CLASSIC-UB initial/final	Syllable	1870.38	1110.38	2564.54
PUDDLE vs. Baseline	Syllable	404.77	-92.92	951.54
PUDDLE vs. CLASSIC-UB final	Syllable	216.42	-611.89	996.58

Note. The $\Delta\chi^2$ measure examined whether two models' distributions were at the same distance from children's expected probabilities. The order of each pairwise difference was set as in the column Comparison (e.g., in Baseline vs. CLASSIC-UB final*,* the CLASSIC-UB final χ^2 estimate is subtracted from the Baseline χ^2 estimate). Lower and upper limits of bootstrap 95% confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S18.5 For each model, comparison of phonemic vs. syllabic model implementation (ΔX^2), in the phonemic length measure. For each model implementation, the analysis controls for chance levels by initially subtracting a baseline X^2 from a model X^2 (see Chapter 3 for a detailed explanation). For each model, positive and negative ΔX^2 values indicate higher contribution of phonemic or syllabic input respectively.

		95% CI	
Model	ΔX^2	*LL*	*UL*
BTP	7202.26	6081.08	8363.48
FTP	1049.87	-319.49	2384.48
CLASSIC-UB final	9171.62	8214.39	10223.04
CLASSIC-UB initial/final	7709.99	6738.81	8700.69
PUDDLE	6677.67	5789.53	7585.13

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S19: Approximation of Child Production Vocabulary by weighted log10 frequency

Word types

We ran analyses on both phonemic and syllabified input. We have given a narrative account of the phonemic-input analysis in the section Results / Word-Level Characteristics of Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S19.1 Child-model comparison by weighted log10 frequency.

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.4	<.001	.33	.46
Children vs. BTP	Phoneme	.49	<.001	.42	.58
Children vs. FTP	Phoneme	.31	<.001	.24	.39
Children vs. CLASSIC-UB final	Phoneme	.12	<.001	.1	.16
Children vs. CLASSIC-UB initial/final	Phoneme	.14	<.001	.11	.18
Children vs. PUDDLE	Phoneme	**.09**	<.001	.06	.12
Children vs. Baseline	Syllable	**.04**	.089	.02	.08
Children vs. BTP	Syllable	.09	<.001	.06	.15
Children vs. FTP	Syllable	.17	<.001	.13	.22
Children vs. CLASSIC-UB final	Syllable	.09	<.001	.07	.12
Children vs. CLASSIC-UB initial/final	Syllable	.09	<.001	.07	.11
Children vs. PUDDLE	Syllable	.11	<.001	.07	.15

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted log10 word frequency were compared to

child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S19.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S19.1

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. PUDDLE	Phoneme	.314	.228	.392
Baseline vs. CLASSIC-UB final	Phoneme	.276	.2	.349
Baseline vs. CLASSIC-UB initial/final	Phoneme	.263	.187	.332
Baseline vs. FTP	Phoneme	.094	-.009	.174
BTP vs. PUDDLE	Phoneme	.405	.289	.507
BTP vs. CLASSIC-UB final	Phoneme	.367	.27	.458
BTP vs. CLASSIC-UB initial/final	Phoneme	.354	.259	.447
BTP vs. FTP	Phoneme	.185	.075	.299
BTP vs. Baseline	Phoneme	.091	-.007	.199
FTP vs. PUDDLE	Phoneme	.22	.133	.328
FTP vs. CLASSIC-UB final	Phoneme	.182	.099	.267
FTP vs. CLASSIC-UB initial/final	Phoneme	.169	.08	.276
CLASSIC-UB final vs. PUDDLE	Phoneme	.038	-.026	.108
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.051	-.017	.127
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.013	-.021	.054
BTP vs. Baseline	Syllable	.051	-.008	.098

BTP vs. CLASSIC-UB initial/final	Syllable	.003	-.033	.052
BTP vs. CLASSIC-UB final	Syllable	0	-.028	.041
FTP vs. Baseline	Syllable	.128	.036	.186
FTP vs. CLASSIC-UB initial/final	Syllable	.08	.02	.141
FTP vs. CLASSIC-UB final	Syllable	.077	.018	.132
FTP vs. BTP	Syllable	.077	-.005	.143
FTP vs. PUDDLE	Syllable	.062	.019	.109
CLASSIC-UB final vs. Baseline	Syllable	.051	.003	.079
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.003	-.01	.018
CLASSIC-UB initial/final vs. Baseline	Syllable	.048	.001	.076
PUDDLE vs. Baseline	Syllable	.065	-.009	.13
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.018	-.036	.079
PUDDLE vs. CLASSIC-UB final	Syllable	.014	-.044	.075
PUDDLE vs. BTP	Syllable	.014	-.06	.073

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted log10 word frequency when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

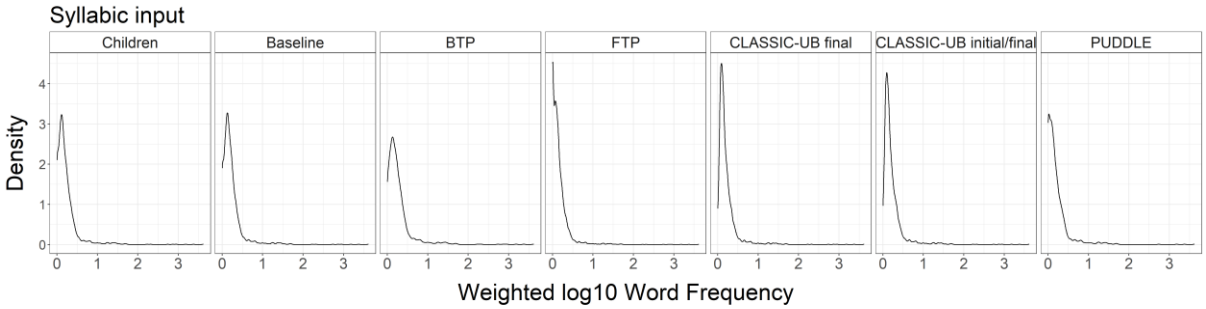


Figure S19.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted log10 word frequency. Syllabified input was used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

Word tokens

This analysis is a repetition of the analysis done on word types above, but considering the distributions of word tokens in children and models. We have given a narrative account of this analysis in the section Exploratory Analysis of Word-Level Properties at the Token Level in Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S19.3 Child-model comparison by weighted log10 frequency.

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.42	<.001	.41	.44
Children vs. BTP	Phoneme	.46	<.001	.45	.48
Children vs. FTP	Phoneme	.43	<.001	.42	.44
Children vs. CLASSIC-UB final	Phoneme	**.08**	<.001	.07	.09

Children vs. CLASSIC-UB initial/final	Phoneme	**.08**	<.001	.07	.09
Children vs. PUDDLE	Phoneme	.16	<.001	.15	.17
Children vs. Baseline	Syllable	.18	<.001	.17	.2
Children vs. BTP	Syllable	.13	<.001	.12	.14
Children vs. FTP	Syllable	.23	<.001	.22	.24
Children vs. CLASSIC-UB final	Syllable	.08	<.001	.07	.09
Children vs. CLASSIC-UB initial/final	Syllable	**.05**	<.001	.04	.06
Children vs. PUDDLE	Syllable	.16	<.001	.15	.17

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted log10 word frequency were compared to child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S19.4 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S19.3

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	.344	.323	.364
Baseline vs. CLASSIC-UB initial/final	Phoneme	.343	.327	.358
Baseline vs. PUDDLE	Phoneme	.262	.25	.277
BTP vs. CLASSIC-UB final	Phoneme	.385	.363	.406
BTP vs. CLASSIC-UB initial/final	Phoneme	.383	.364	.399
BTP vs. PUDDLE	Phoneme	.302	.289	.314

BTP vs. Baseline	Phoneme	.04	.023	.055
BTP vs. FTP	Phoneme	.035	.023	.049
FTP vs. CLASSIC-UB final	Phoneme	.35	.333	.366
FTP vs. CLASSIC-UB initial/final	Phoneme	.348	.334	.358
FTP vs. PUDDLE	Phoneme	.267	.258	.275
FTP vs. Baseline	Phoneme	.005	-.006	.016
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.001	-.007	.012
PUDDLE vs. CLASSIC-UB final	Phoneme	.082	.066	.102
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.081	.069	.091
Baseline vs. CLASSIC-UB initial/final	Syllable	.136	.12	.147
Baseline vs. CLASSIC-UB final	Syllable	.103	.081	.126
Baseline vs. BTP	Syllable	.049	.028	.074
Baseline vs. PUDDLE	Syllable	.024	.015	.032
BTP vs. CLASSIC-UB initial/final	Syllable	.087	.067	.103
BTP vs. CLASSIC-UB final	Syllable	.054	.044	.063
FTP vs. CLASSIC-UB initial/final	Syllable	.18	.162	.195
FTP vs. CLASSIC-UB final	Syllable	.147	.139	.157
FTP vs. BTP	Syllable	.094	.085	.105
FTP vs. PUDDLE	Syllable	.069	.05	.089
FTP vs. Baseline	Syllable	.045	.025	.064
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.033	.015	.048
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.112	.1	.12

PUDDLE vs. CLASSIC-UB final	Syllable	.078	.061	.097
PUDDLE vs. BTP	Syllable	.025	.008	.043

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted log10 word frequency when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S19.5 For each model, comparison of phonemic vs. syllabic model implementation (ΔD), in the word frequency measure. For each model implementation, the analysis controls for chance levels by initially subtracting a baseline D from a model D (see Chapter 3 for a detailed explanation). For each model, positive and negative ΔD values indicate higher contribution of phonemic or syllabic input respectively.

		95% CI	
Model	ΔD	*LL*	*UL*
BTP	-.09	-.11	-.07
FTP	.04	.02	.06
CLASSIC-UB final	.24	.22	.26
CLASSIC-UB initial/final	.21	.19	.22
PUDDLE	.24	.22	.25

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S20: Approximation of Child Production Vocabulary by weighted neighbourhood density

Word types

We ran analyses on both phonemic and syllabified input. We have given a narrative account of the phonemic-input analysis in the section Results / Word-Level Characteristics of Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S20.1 Child-model comparison by weighted neighbourhood density

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.26	<.001	.2	.33
Children vs. BTP	Phoneme	.46	<.001	.4	.54
Children vs. FTP	Phoneme	.47	<.001	.4	.54
Children vs. CLASSIC-UB final	Phoneme	.06	.144	.02	.1
Children vs. CLASSIC-UB initial/final	Phoneme	.07	.022	.04	.12
Children vs. PUDDLE	Phoneme	**.03**	1	.02	.08
Children vs. Baseline	Syllable	**.01**	1	.01	.05
Children vs. BTP	Syllable	.01	1	.02	.06
Children vs. FTP	Syllable	.09	<.001	.06	.14
Children vs. CLASSIC-UB final	Syllable	.04	.869	.02	.08
Children vs. CLASSIC-UB initial/final	Syllable	.03	1	.01	.07
Children vs. PUDDLE	Syllable	.04	1	.02	.08

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted neighbourhood density were compared to

child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S20.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S20.1

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. PUDDLE	Phoneme	.229	.152	.286
Baseline vs. CLASSIC-UB final	Phoneme	.206	.129	.265
Baseline vs. CLASSIC-UB initial/final	Phoneme	.191	.112	.254
BTP vs. PUDDLE	Phoneme	.427	.336	.513
BTP vs. CLASSIC-UB final	Phoneme	.404	.317	.497
BTP vs. CLASSIC-UB initial/final	Phoneme	.389	.297	.49
BTP vs. Baseline	Phoneme	.198	.108	.31
FTP vs. PUDDLE	Phoneme	.434	.331	.512
FTP vs. CLASSIC-UB final	Phoneme	.411	.325	.495
FTP vs. CLASSIC-UB initial/final	Phoneme	.396	.307	.476
FTP vs. Baseline	Phoneme	.206	.112	.303
FTP vs. BTP	Phoneme	.008	-.092	.086
CLASSIC-UB final vs. PUDDLE	Phoneme	.023	-.027	.061
CLASSIC-UB initial/final vs. PUDDLE	Phoneme	.038	-.021	.082
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.015	-.026	.065
Baseline vs. BTP	Syllable	.001	-.029	.022

FTP vs. BTP	Syllable	.081	.009	.11
FTP vs. Baseline	Syllable	.081	.011	.113
FTP vs. CLASSIC-UB initial/final	Syllable	.064	.021	.094
FTP vs. PUDDLE	Syllable	.059	.004	.091
FTP vs. CLASSIC-UB final	Syllable	.057	.014	.088
CLASSIC-UB final vs. BTP	Syllable	.025	-.039	.051
CLASSIC-UB final vs. Baseline	Syllable	.024	-.04	.054
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.008	-.028	.043
CLASSIC-UB final vs. PUDDLE	Syllable	.002	-.034	.032
CLASSIC-UB initial/final vs. BTP	Syllable	.017	-.037	.049
CLASSIC-UB initial/final vs. Baseline	Syllable	.016	-.031	.048
PUDDLE vs. BTP	Syllable	.023	-.033	.058
PUDDLE vs. Baseline	Syllable	.022	-.031	.06
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.006	-.028	.042

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted neighbourhood density when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

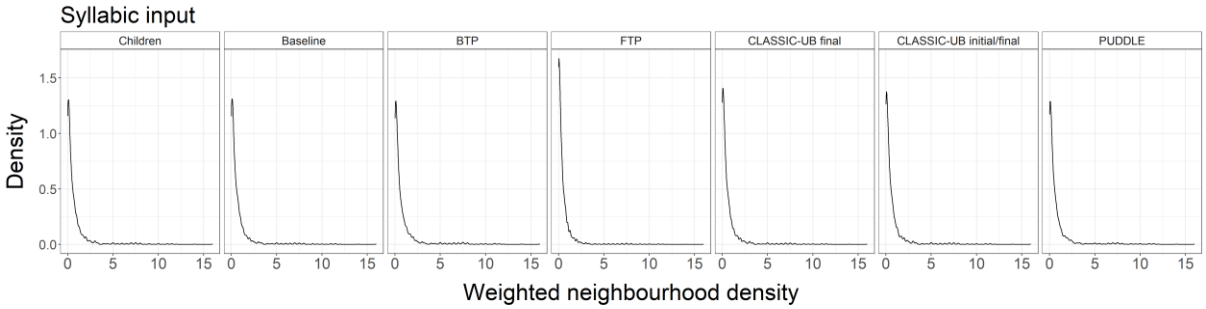


Figure S20.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted neighbourhood density. Syllabified input was used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

Word tokens

This analysis is a repetition of the analysis done on word types above, but considering the distributions of word tokens in children and models. We have given a narrative account of this analysis in the section Exploratory Analysis of Word-Level Properties at the Token Level in Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S20.3 Child-model comparison by weighted neighbourhood density

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.41	<.001	.4	.42
Children vs. BTP	Phoneme	.38	<.001	.36	.39
Children vs. FTP	Phoneme	.32	<.001	.31	.34
Children vs. CLASSIC-UB final	Phoneme	.08	<.001	.07	.09

Children vs. CLASSIC-UB initial/final	Phoneme	**.03**	<.001	.03	.04
Children vs. PUDDLE	Phoneme	.14	<.001	.13	.15
Children vs. Baseline	Syllable	.14	<.001	.13	.16
Children vs. BTP	Syllable	.14	<.001	.12	.15
Children vs. FTP	Syllable	.2	<.001	.19	.22
Children vs. CLASSIC-UB final	Syllable	.09	<.001	.08	.1
Children vs. CLASSIC-UB initial/final	Syllable	**.06**	<.001	.05	.06
Children vs. PUDDLE	Syllable	.11	<.001	.1	.12

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted neighbourhood density were compared to child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S20.4 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S20.3

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. CLASSIC-UB initial/final	Phoneme	.377	.353	.395
Baseline vs. CLASSIC-UB final	Phoneme	.331	.307	.35
Baseline vs. PUDDLE	Phoneme	.274	.257	.291
Baseline vs. FTP	Phoneme	.086	.071	.1
Baseline vs. BTP	Phoneme	.034	.019	.05
BTP vs. CLASSIC-UB initial/final	Phoneme	.343	.319	.367

BTP vs. CLASSIC-UB final	Phoneme	.297	.271	.318
BTP vs. PUDDLE	Phoneme	.24	.223	.257
BTP vs. FTP	Phoneme	.052	.036	.067
FTP vs. CLASSIC-UB initial/final	Phoneme	.291	.269	.31
FTP vs. CLASSIC-UB final	Phoneme	.244	.223	.267
FTP vs. PUDDLE	Phoneme	.187	.174	.2
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.046	.037	.055
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.104	.086	.12
PUDDLE vs. CLASSIC-UB final	Phoneme	.057	.041	.071
Baseline vs. CLASSIC-UB initial/final	Syllable	.085	.065	.107
Baseline vs. CLASSIC-UB final	Syllable	.052	.03	.075
Baseline vs. PUDDLE	Syllable	.03	.021	.039
Baseline vs. BTP	Syllable	.007	-.01	.021
BTP vs. CLASSIC-UB initial/final	Syllable	.079	.066	.091
BTP vs. CLASSIC-UB final	Syllable	.045	.034	.057
BTP vs. PUDDLE	Syllable	.023	.006	.039
FTP vs. CLASSIC-UB initial/final	Syllable	.148	.139	.159
FTP vs. CLASSIC-UB final	Syllable	.115	.104	.127
FTP vs. PUDDLE	Syllable	.093	.072	.11
FTP vs. BTP	Syllable	.07	.055	.081
FTP vs. Baseline	Syllable	.063	.043	.083
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.033	.026	.041

PUDDLE vs. CLASSIC-UB initial/final	Syllable	.055	.036	.075
PUDDLE vs. CLASSIC-UB final	Syllable	.022	.007	.038

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted neighbourhood density when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S20.5 For each model, comparison of phonemic vs. syllabic model implementation (ΔD), in the neighbourhood density measure. For each model implementation, the analysis controls for chance levels by initially subtracting a baseline D from a model D (see Chapter 3 for a detailed explanation). For each model, positive and negative ΔD values indicate higher contribution of phonemic or syllabic input respectively.

		95% CI	
Model	ΔD	*LL*	*UL*
BTP	.03	.01	.05
FTP	.15	.13	.17
CLASSIC-UB final	.28	.26	.29
CLASSIC-UB initial/final	.29	.27	.31
PUDDLE	.24	.23	.26

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S21: Approximation of Child Production Vocabulary by weighted phonotactic probability

Word types

We ran analyses on both phonemic and syllabified input. We have given a narrative account of the phonemic-input analysis in the section Results / Word-Level Characteristics of Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S21.1 Child-model comparison by weighted phonotactic probability

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.1	.003	.06	.17
Children vs. BTP	Phoneme	.28	<.001	.2	.38
Children vs. FTP	Phoneme	.21	<.001	.14	.3
Children vs. CLASSIC-UB final	Phoneme	**.02**	1	.02	.07
Children vs. CLASSIC-UB initial/final	Phoneme	.06	.075	.03	.12
Children vs. PUDDLE	Phoneme	.12	<.001	.08	.18
Children vs. Baseline	Syllable	**.03**	1	.02	.07
Children vs. BTP	Syllable	.07	.006	.04	.13
Children vs. FTP	Syllable	.07	.001	.04	.12
Children vs. CLASSIC-UB final	Syllable	.04	.394	.02	.09
Children vs. CLASSIC-UB initial/final	Syllable	.03	.803	.02	.08
Children vs. PUDDLE	Syllable	.11	<.001	.07	.17

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted phonotactic probability were compared to

child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S21.2 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S8.1

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. CLASSIC-UB final	Phoneme	.079	.01	.136
Baseline vs. CLASSIC-UB initial/final	Phoneme	.042	-.03	.114
BTP vs. CLASSIC-UB final	Phoneme	.253	.139	.339
BTP vs. CLASSIC-UB initial/final	Phoneme	.217	.107	.315
BTP vs. Baseline	Phoneme	.174	.067	.279
BTP vs. PUDDLE	Phoneme	.155	.04	.28
BTP vs. FTP	Phoneme	.068	-.052	.183
FTP vs. CLASSIC-UB final	Phoneme	.185	.087	.272
FTP vs. CLASSIC-UB initial/final	Phoneme	.149	.071	.236
FTP vs. Baseline	Phoneme	.106	.004	.216
FTP vs. PUDDLE	Phoneme	.087	.005	.196
CLASSIC-UB initial/final vs. CLASSIC-UB final	Phoneme	.036	-.014	.077
PUDDLE vs. CLASSIC-UB final	Phoneme	.098	.01	.144
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.062	-.02	.124
PUDDLE vs. Baseline	Phoneme	.02	-.053	.089
BTP vs. Baseline	Syllable	.048	-.002	.087

BTP vs. CLASSIC-UB initial/final	Syllable	.041	-.006	.083
BTP vs. CLASSIC-UB final	Syllable	.034	-.013	.08
BTP vs. FTP	Syllable	.004	-.028	.042
FTP vs. Baseline	Syllable	.043	0	.073
FTP vs. CLASSIC-UB initial/final	Syllable	.037	-.005	.072
FTP vs. CLASSIC-UB final	Syllable	.03	-.002	.065
CLASSIC-UB final vs. Baseline	Syllable	.013	-.025	.041
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.006	-.025	.035
CLASSIC-UB initial/final vs. Baseline	Syllable	.007	-.027	.039
PUDDLE vs. Baseline	Syllable	.081	.027	.124
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.074	.018	.118
PUDDLE vs. CLASSIC-UB final	Syllable	.068	.018	.111
PUDDLE vs. FTP	Syllable	.038	-.005	.085
PUDDLE vs. BTP	Syllable	.033	-.013	.08

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted phonotactic probability when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

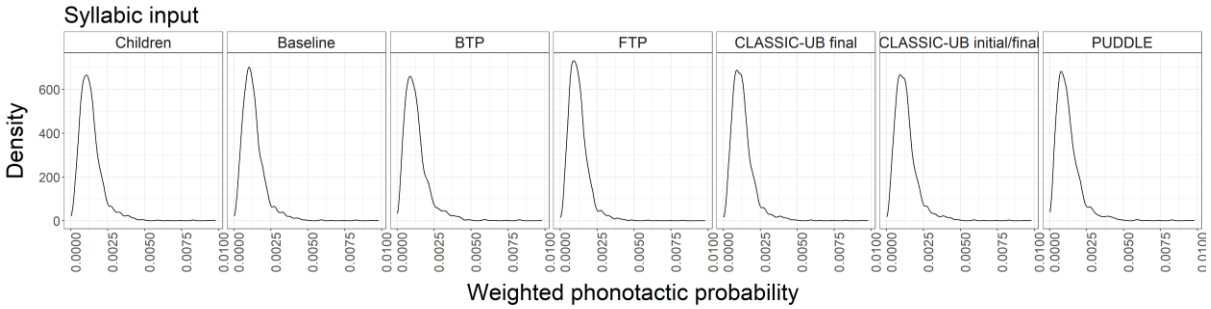


Figure S21.1 Gaussian kernel density estimate of the distribution of unique words in children's speech (Children) and discovered by each model, by weighted phonotactic probability. Syllabified input was used. The area under each curve represents 100% of data points. Curve peaks represent the mode of each distribution.

Word tokens

This analysis is a repetition of the analysis done on word types above, but considering the distributions of word tokens in children and models. We have given a narrative account of this analysis in the section Exploratory Analysis of Word-Level Properties at the Token Level in Chapter 3; below, we include statistical comparisons for both phonemic input and syllabic input.

Table S21.3 Child-model comparison by weighted phonotactic probability

				95% CI	
Comparison	Input type	D	*p*	*LL*	*UL*
Children vs. Baseline	Phoneme	.14	<.001	.13	.16
Children vs. BTP	Phoneme	.13	<.001	.11	.14
Children vs. FTP	Phoneme	.19	<.001	.18	.21
Children vs. CLASSIC-UB final	Phoneme	.07	<.001	.06	.08

Children vs. CLASSIC-UB initial/final	Phoneme	**.06**	<.001	.05	.07
Children vs. PUDDLE	Phoneme	.09	<.001	.08	.1
Children vs. Baseline	Syllable	.07	<.001	.06	.08
Children vs. BTP	Syllable	.11	<.001	.1	.12
Children vs. FTP	Syllable	.13	<.001	.12	.14
Children vs. CLASSIC-UB final	Syllable	.07	<.001	.06	.08
Children vs. CLASSIC-UB initial/final	Syllable	**.06**	<.001	.05	.07
Children vs. PUDDLE	Syllable	.08	<.001	.07	.09

Note. Comparisons were tested via Kolmogorov–Smirnov test statistic. Models distributions of unique words by weighted phonotactic probability were compared to child distribution. Holm's correction was applied to *p* values and confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Table S21.4 Pairwise differences between the Kolmogorov–Smirnov statistics reported in Table S21.3

			95% CI	
Model comparison	Input type	ΔD	*LL*	*UL*
Baseline vs. CLASSIC-UB initial/final	Phoneme	.084	.064	.112
Baseline vs. CLASSIC-UB final	Phoneme	.076	.054	.104
Baseline vs. PUDDLE	Phoneme	.056	.042	.073
Baseline vs. BTP	Phoneme	.015	-.006	.037
BTP vs. CLASSIC-UB initial/final	Phoneme	.069	.048	.094
BTP vs. CLASSIC-UB final	Phoneme	.061	.039	.081

BTP vs. PUDDLE	Phoneme	.041	.026	.057
FTP vs. CLASSIC-UB initial/final	Phoneme	.134	.115	.156
FTP vs. CLASSIC-UB final	Phoneme	.126	.104	.145
FTP vs. PUDDLE	Phoneme	.106	.093	.116
FTP vs. BTP	Phoneme	.065	.047	.085
FTP vs. Baseline	Phoneme	.05	.028	.069
CLASSIC-UB final vs. CLASSIC-UB initial/final	Phoneme	.008	-.001	.019
PUDDLE vs. CLASSIC-UB initial/final	Phoneme	.029	.008	.046
PUDDLE vs. CLASSIC-UB final	Phoneme	.02	.003	.036
Baseline vs. CLASSIC-UB initial/final	Syllable	.006	-.008	.02
BTP vs. CLASSIC-UB initial/final	Syllable	.05	.036	.061
BTP vs. Baseline	Syllable	.044	.024	.063
BTP vs. CLASSIC-UB final	Syllable	.043	.031	.055
BTP vs. PUDDLE	Syllable	.034	.015	.052
FTP vs. CLASSIC-UB initial/final	Syllable	.064	.055	.076
FTP vs. Baseline	Syllable	.058	.039	.076
FTP vs. CLASSIC-UB final	Syllable	.057	.046	.068
FTP vs. PUDDLE	Syllable	.048	.029	.068
FTP vs. BTP	Syllable	.014	.004	.026
CLASSIC-UB final vs. CLASSIC-UB initial/final	Syllable	.008	0	.016
CLASSIC-UB final vs. Baseline	Syllable	.001	-.01	.014
PUDDLE vs. CLASSIC-UB initial/final	Syllable	.016	0	.035

PUDDLE vs. Baseline	Syllable	.01	0	.021
PUDDLE vs. CLASSIC-UB final	Syllable	.008	-.008	.026

Note. Comparison of how closely two models' distributions of unique words were to children's productions by weighted phonotactic probability when phonemic or syllabified input was used. Lower and upper limits of bootstrap confidence intervals were based on 1,000 iterations and corrected using Holm's correction. BTP = backward transitional probability; FTP = forward transitional probability.

Table S21.5 For each model, comparison of phonemic vs. syllabic model implementation (ΔD), in the phonotactic probability measure. For each model implementation, the analysis controls for chance levels by initially subtracting a baseline D from a model D (see Chapter 3 for a detailed explanation). For each model, positive and negative ΔD values indicate higher contribution of phonemic or syllabic input respectively.

		95% CI	
Model	ΔD	*LL*	*UL*
BTP	.06	.04	.08
FTP	.01	-.01	.03
CLASSIC-UB final	.08	.06	.1
CLASSIC-UB initial/final	.08	.06	.1
PUDDLE	.07	.05	.08

Note. 95% confidence intervals are corrected for multiple comparisons (using Holm's correction). FTP = forward transitional probability; BTP = backward transitional probability.

Appendix S22: Size of Noun Advantage

In this section, we report an analysis that compares the size of the noun advantage over verbs in children and models, by phonemic or syllabic input. A detailed description of how this analysis was carried out is included in the method section of Chapter 3. Here, we report the statistical results of the analysis in Table S22.1. A narrative account of the results is included in the section Results / Word-Level Characteristics of Chapter 3.

Table S22.1 Child-model noun advantage comparisons, by input type. A negative ΔP indicates a that a model noun advantage is smaller than that shown in children's productions.

			95% CI	
Model comparison	Input type	ΔP	*LL*	*UL*
Children vs. Baseline	Phoneme	-.177	-.285	-.068
Children vs. BTP	Phoneme	-.238	-.39	-.074
Children vs. FTP	Phoneme	-.124	-.243	-.001
Children vs. CLASSIC-UB final	Phoneme	-.133	-.219	-.047
Children vs. CLASSIC-UB initial/final	Phoneme	-.136	-.229	-.051
Children vs. PUDDLE	Phoneme	-.095	-.162	-.025
Children vs. Baseline	Syllable	-.095	-.169	-.023
Children vs. BTP	Syllable	-.167	-.274	-.072
Children vs. FTP	Syllable	-.176	-.257	-.096
Children vs. CLASSIC-UB final	Syllable	-.11	-.182	-.044
Children vs. CLASSIC-UB initial/final	Syllable	-.14	-.219	-.058
Children vs. PUDDLE	Syllable	-.111	-.196	-.024

Note. ΔP is the difference between the proportional noun advantage in children and model: for example, ΔP = phonemic Baseline's P (29% nouns – 31% verbs = -2%) - Children's P (47% nouns – 31% verbs = 16%) = -18% . Holm's correction was applied to confidence intervals. BTP = backward transitional probability; FTP = forward transitional probability.

Appendix S23: Input Changes in Neighbourhood Density and Phonotactic Probability as a Function of Word Phonemic Length

In this section, we include plots that show a non-linear relation of phonemic length with neighbourhood density (Figure S23.1a) and phonotactic probability (Figure S23.1b). Such non-linear relations mean that in Italian, given its high average word length, input words have a low number of neighbours in the language and are composed of biphone sequences that are frequent in the language.

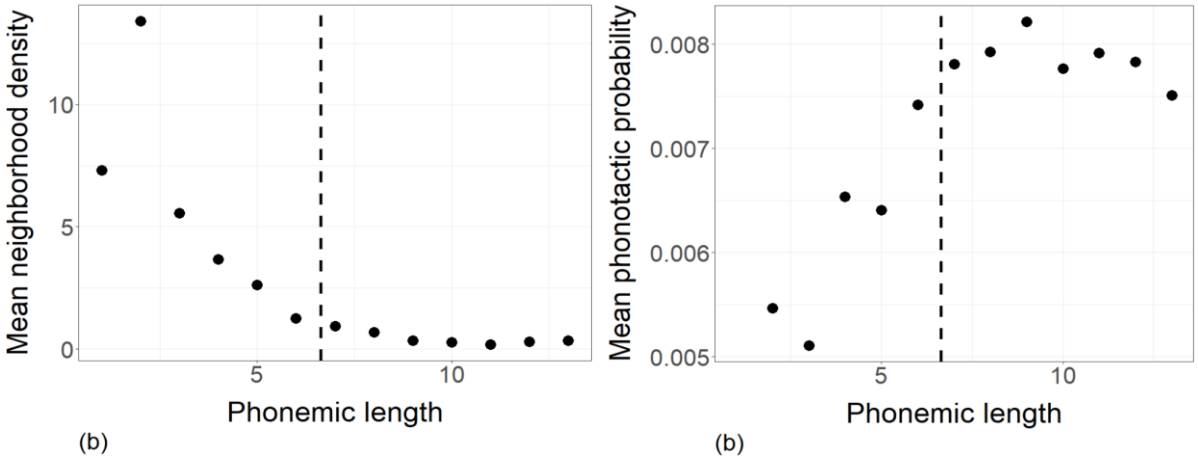


Figure S23.1 Average neighbourhood density (a) and average phonotactic probability (b) by word phonemic length, in the Italian child-directed speech input. Vertical dashed lines indicate the average word length in the child-directed speech input.

Appendix S24: Frequency Match of Target Senses and Distractor Words' Distributions

Distractor words appeared in the input corpora and were chosen (specifically considering their stems) to match the targets' input frequency distribution and to be semantically associated to target senses. Where possible (i.e., given the frequency constraint was satisfied), semantic association between distractors and targets was based on the Florida association norms (Nelson et al., 2004). Input frequencies of target senses and corresponding distractors are shown in Table S24.1, while Figure S24.1 shows how the frequency distributions of distractors and targets compare to the frequency distribution of common nouns in the input corpora.

Table S24.1 Input frequency per million of target senses and corresponding distractor words.

Target	Frequency Target	Distractor	Frequency Distractor
band: object	14.38	sock	204.13
band: music group	4.69	team	9.23
bat: animal	19.96	owl	80.99
bat: object	10.50	sword	11.30
bow: knot	18.58	dress	135.64
bow: weapon	2.18	target	2.63
button: tech	45.89	bell	64.92
button: clothing	23.03	zip	11.85
chicken: animal	118.20	crow	5.25
chicken: food	75.71	biscuit	138.90
flower/flour: flower	284.48	leaf	130.39
flower/flour: flour	28.28	salad	43.99
glasses: eye	55.18	scarf	21.72

Target	Frequency Target	Distractor	Frequency Distractor
glasses: drinking	50.09	jug	14.08
letter: alphabet	116.83	number	240.57
letter: mail	76.43	box	717.73
line: geometry	38.05	circle	150.67
line: order	19.47	square	64.92
moose/mousse: moose	14.38	gorilla	21.96
moose/mousse: mousse	3.39	donut	10.82
nail: body part	37.17	carrot	119.97
nail: object	8.56	screwdriver	28.68
sun/son: sun	163.93	tree	421.63
sun/son: son	29.49	pirate	19.57

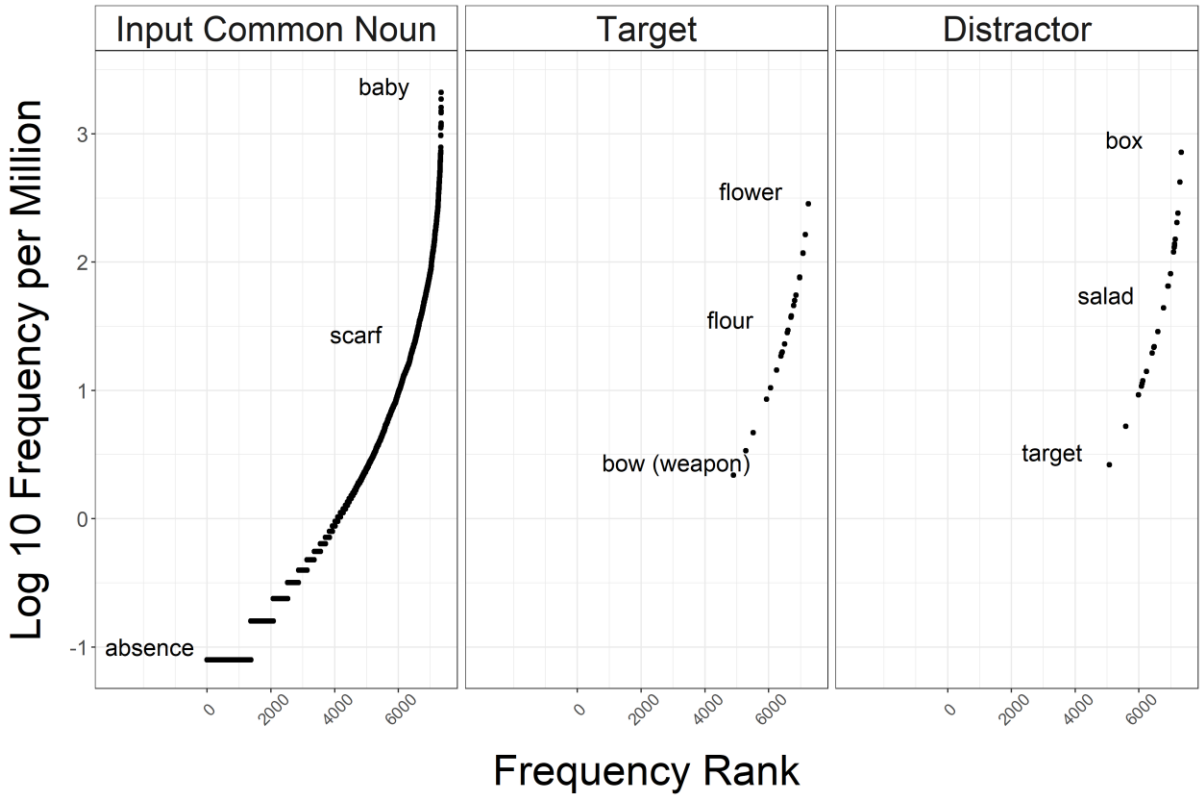


Figure S24.1 Input frequency distribution of all common nouns in the corpora (left panel), target nouns (middle panel) and distractors (right panel). No significant

difference was found between distractors and targets' log10 frequency distributions (Welch t = -1.26, p = 0.215).

References

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Appendix S25: Socio-Demographic Characteristics of the Child Sample

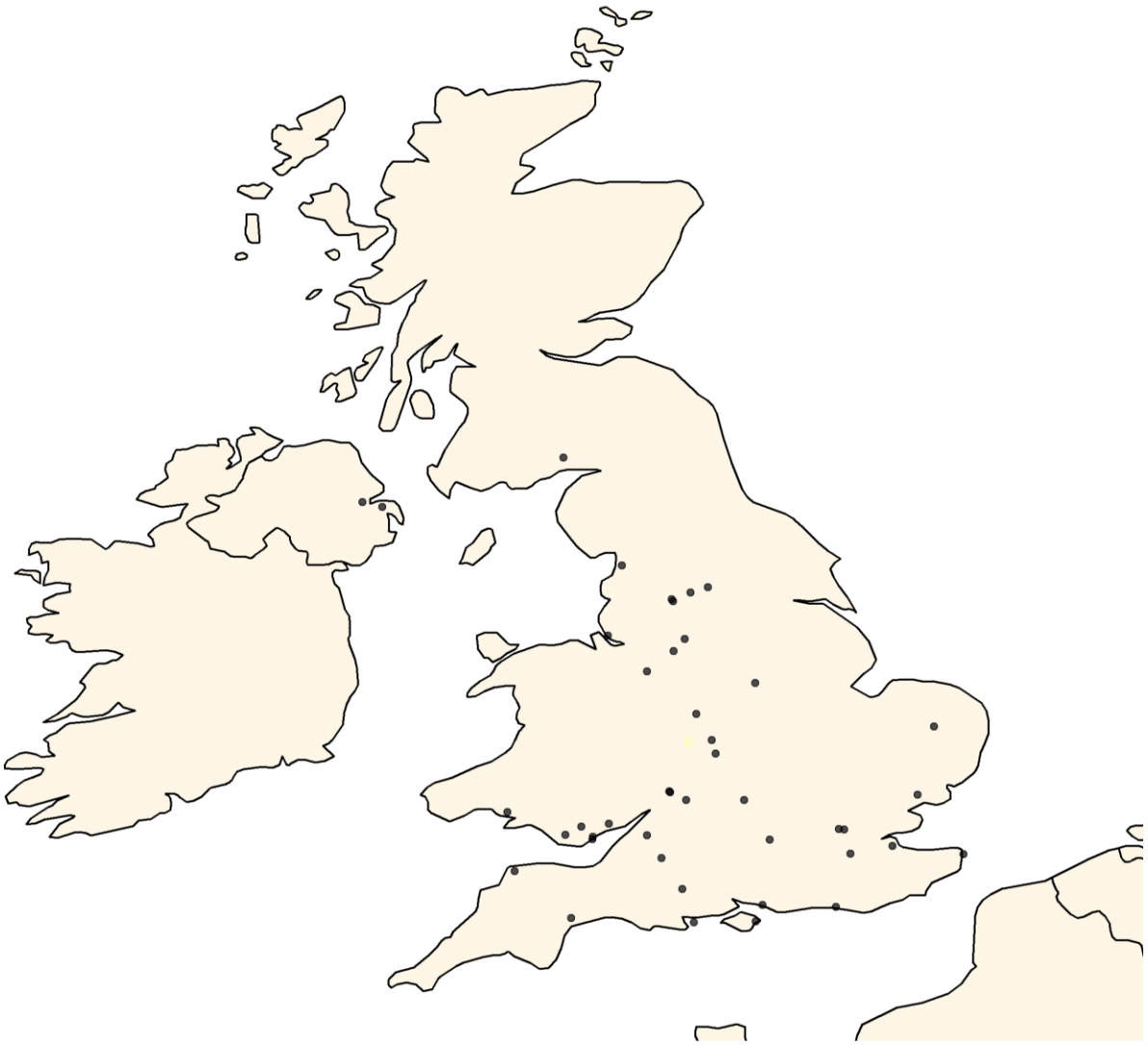


Figure S25.1 Place of residence of children's families in the UK.

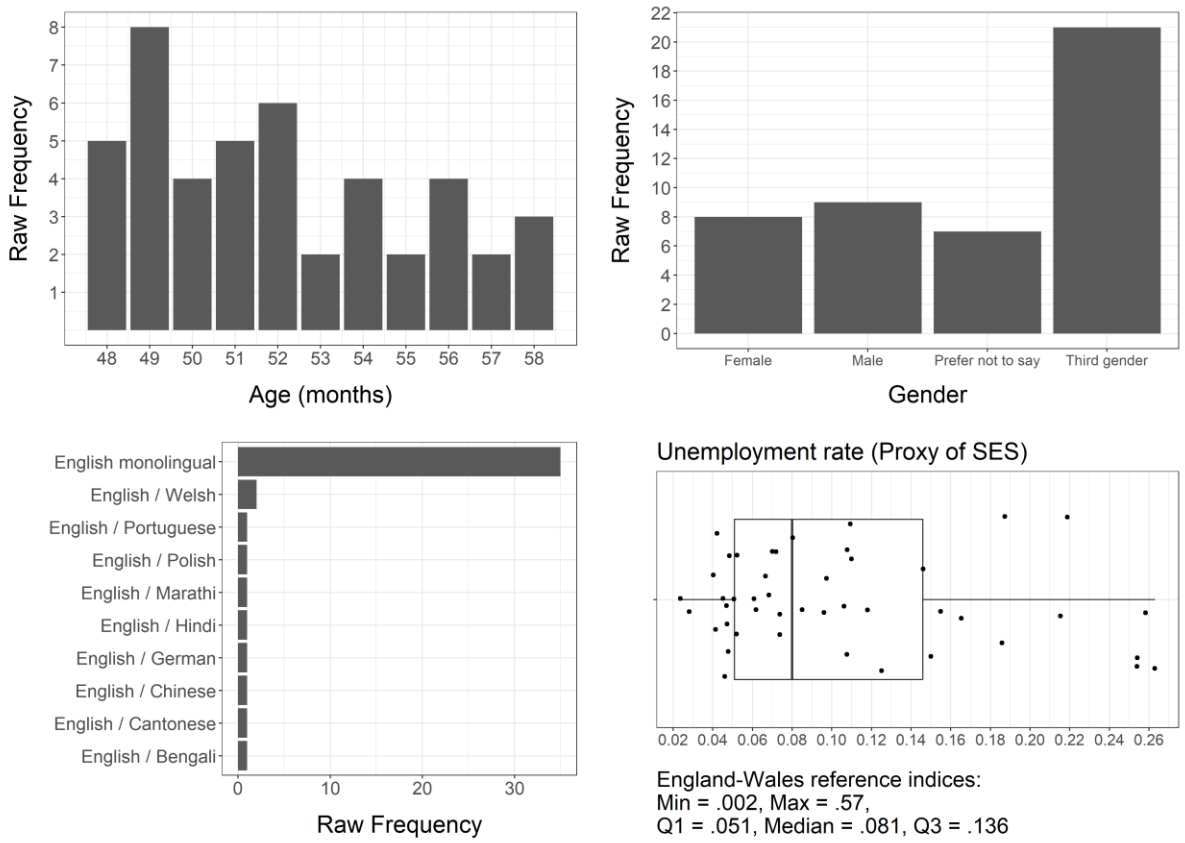


Figure S25.2 Socio-Demographic characteristics of the child sample ($N = 45$). Postcode unemployment rate was chosen as a proxy of socio-economic status because each country (England-Wales, Scotland, and Northern Ireland) had different indices of deprivation. Instead, the proportion of population that is unemployed was a measure consistently used across countries' deprivation reports.

Appendix S26: Experimental Stories

Table S26.1 Stimuli divided by 3 counterbalancing blocks. Each story has a prior context and a following control, verb-lexical or verb-event context. Verb-sense associations for subordinate and dominant senses are reported. These were computed as the raw frequency of verb-sense occurrence weighted by the number of times a sense appeared in ChiSense-12 as a verb object. The first three stories are training trials, presented to participants in random order.

Training Trials		
Emily went to the shop. Then, she bought a banana.		
John and Mary were at the zoo. Then, they heard the tiger.		
Charlie was waiting for his friend. Then, he opened the door.		
Block A		
Prior context	***Control context***	***Verb-sense association (subordinate/dominant)***
Sophia listened to some music	Then, she played in a band	.029 / .000
John threw the ball	Then, Mary swung the bat	.078 / .000
Wendy bought some tools and a piece of wood	Then, she got a nail	.056 / .014
George had an apple for breakfast	Then, he ate a mousse	.077 / .000
Prior context	***Verb-lexical context***	
The teacher said goodbye to the daughter	Then, she looked at the sun	.000 / .041
Harry got eggs, milk and sugar	Then, he held the flower	.000 / .006
Olivia had some chips	Then, she saw the chicken	.005 / 053
Jack got some arrows	Then, he made a bow	.000 / .054
Prior context	***Verb-event context***	
Julia and Beth wanted some milk	Then, Julia folded the glasses	.000 / .000
Leo and Mark were waiting for the bus	Then, Mark rubbed out the line	.000 / .000
John was putting on a shirt	Then, he rang the button	.000 / .000
Charlie got some stamps this morning	Then, he sang the letters	.000 / .000
Block B		
Prior context	***Control context***	***Verb-sense association (subordinate/dominant)***
The teacher said goodbye to the daughter	Then, she talked to the son	.023 / .000
Harry got eggs, milk and sugar	Then, he mixed the flour	.015 / .000
Olivia had some chips	Then, she ate the chicken	.116 / .007
Jack got some arrows	Then, he shot a bow	.111 / .000
Prior context	***Verb-lexical context***	

Julia and Beth wanted some milk	Then, Julia found the glasses	.007 / .016
Leo and Mark were waiting for the bus	Then, Mark followed the line	.005 / .055
John was putting on a shirt	Then, he touched the button	.000 / .014
Charlie got some stamps this morning	Then, he looked for the letters	.004 / .011
Prior context	***Verb-event context***	
Sophia listened to some music	Then, she twisted a band	.000 / .000
John threw the ball	Then, Mary got bitten by the bat	.000 / .000
Wendy bought some tools and a piece of wood	Then, she chewed on a nail	.000 / .000
George had an apple for breakfast	Then, he met a moose	.000 / .000
Block C		
Prior context	***Control context***	***Verb-sense association (subordinate/dominant)***
Julia and Beth wanted some milk	Then, Julia filled the glasses	.007 / .000
Leo and Mark were waiting for the bus	Then, Mark stood in the line	.044 / .000
John was putting on a shirt	Then, he undid the button	.071 / .000
Charlie got some stamps this morning	Then, he posted the letters	.185 / .000
Prior context	***Verb-lexical context***	
Sophia listened to some music	Then, she got a band	.000 / .065
John threw the ball	Then, Mary liked the bat	.000 / .033
Wendy bought some tools and a piece of wood	Then, she drew a nail	.000 / .014
George had an apple for breakfast	Then, he saw a moose	.038 / .085
Prior context	***Verb-event context***	
The teacher said goodbye to the daughter	Then, she relaxed under the sun	.000 / .000
Harry got eggs, milk and sugar	Then, he trimmed the flower	.000 / .000
Olivia had some chips	Then, she rescued the chicken	.000 / .001
Jack got some arrows	Then, he ironed a bow	.000 / .000

Appendix S27: Children's Reported Knowledge of Target Senses and Verbs

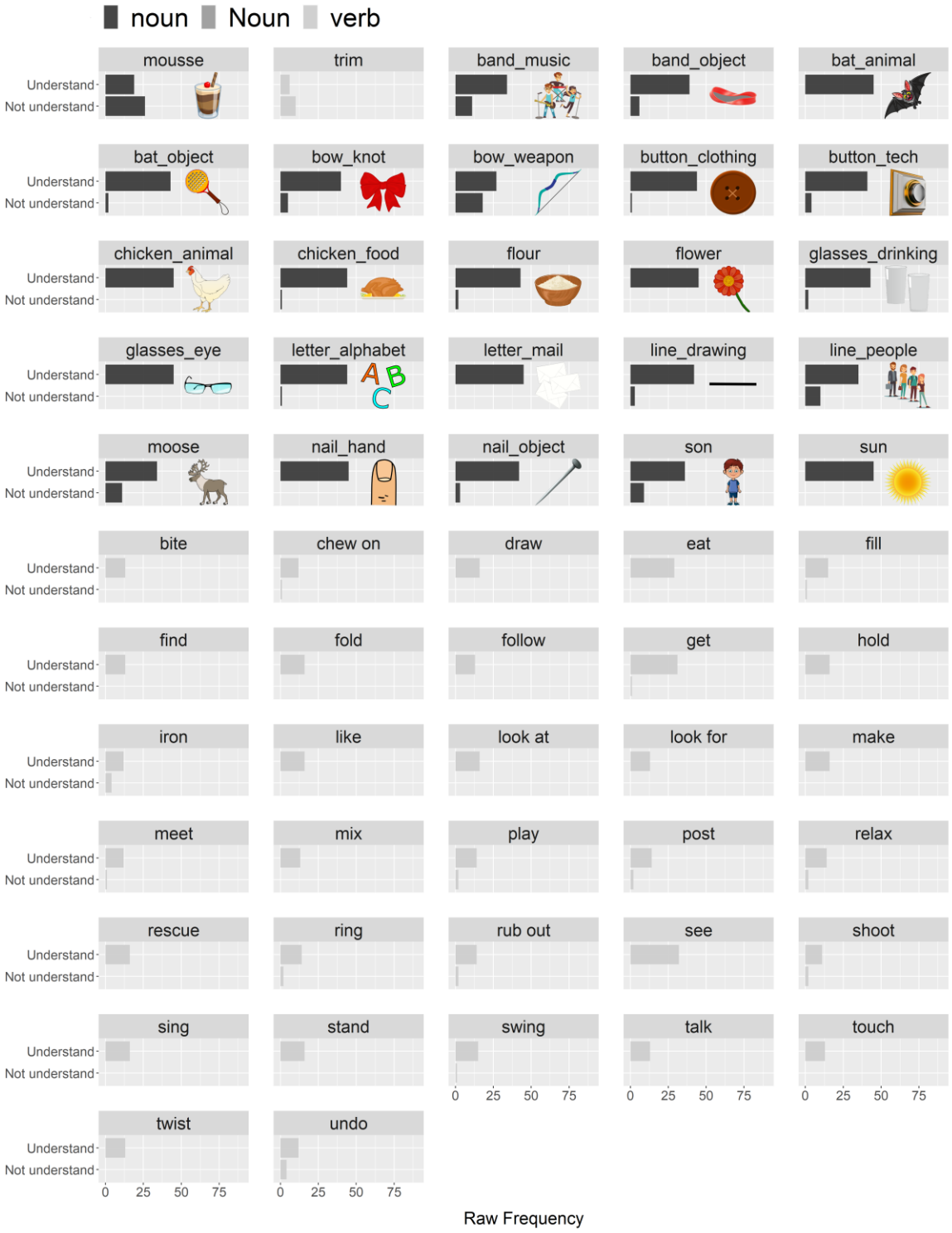


Figure S27.1 Caregivers' responses indicating whether children understood (i.e., "Understand Only" or "Understand & Use") or did not understand a target noun sense or verb. An item (e.g., "band") was excluded from the analyses if the caregiver reported that the child did not understand either the verb used in that

context (e.g., "to play in" in "played in a band") or any of the target senses presented in the task (music band or elastic band). Note that item exclusion was implemented on an individual basis. For instance, even if most children did not comprehend "mousse", the item "moose/mousse" was included for certain children because the caregiver reported that the child understood both target senses and the verb. As depicted in the figure, the only stimuli that seemed to be notably challenging due to a significant proportion of children not knowing them were the sense "mousse" and the verb "to trim".

Appendix S28: Statistical Models of Pre-Registered Analyses

This section presents the reporting of all statistical analyses related to the pre-registered hypotheses investigating the impact of verb-lexical associations and verb-event structures on lexical disambiguation.

Table S28.1 presents the statistical model using adult data. The assumptions of this model are confirmed and examined in Figure S28.1.

Subsequently, Table S28.2 introduces two alternative statistical models for children's data, with differences in their random effect structures. The models use data that have been filtered based on the knowledge of target nouns and verbs reported by parents for their children. Figure S28.2 confirms that the assumptions of these models are met. Following this, Figure S28.3 displays various plots to examine whether a desirable level of power is achieved for the observed effect sizes in the children's data.

Finally, Table S28.3 shows consistent results when applying the same statistical models to the raw data from children, which are not filtered for reported noun and verb knowledge.

Table S28.1

Mixed-effects logistic regression model on adult data. The model employs sense choice (dominant or subordinate) as the dependent variable, while the condition (control, verb-lexical, or verb-event) is used as the independent variable. Two contrasts were analysed: control versus verb-lexical, and control versus verb-event. The model's random effect structure includes random intercepts for participants and items, and random slopes of condition per participant and item. The model omits any estimated correlations between item intercepts and slopes to ensure model convergence.

		Adult Model	
Predictors	*Odds Ratios*	*CI*	*p*

(Intercept)	0.02	0.01 – 0.06	**<0.001**
Control vs. Verb-Lexical	25.29	9.00 – 71.05	**<0.001**
Control vs. Verb-Event	759.56	231.61 – 2490.92	**<0.001**

Random Effects

σ^2	3.29
$T_{00\ id}$	0.06
$T_{00\ Item}$	1.69
$T_{11\ id.VerbLexical}$	0.34
$T_{11\ id.VerbEvent}$	2.21
$T_{11\ Item.VerbLexical}$	1.27
$T_{11\ Item.VerbEvent}$	1.39
$\rho_{01\ id.VerbLexical}$	-0.17
$\rho_{01\ id.VerbEvent}$	-0.99
ICC	0.42
$N_{\ ptid}$	83
$N_{\ ItemNumber}$	12
Observations	995
Marginal R^2 / Conditional R^2	0.566 / 0.746

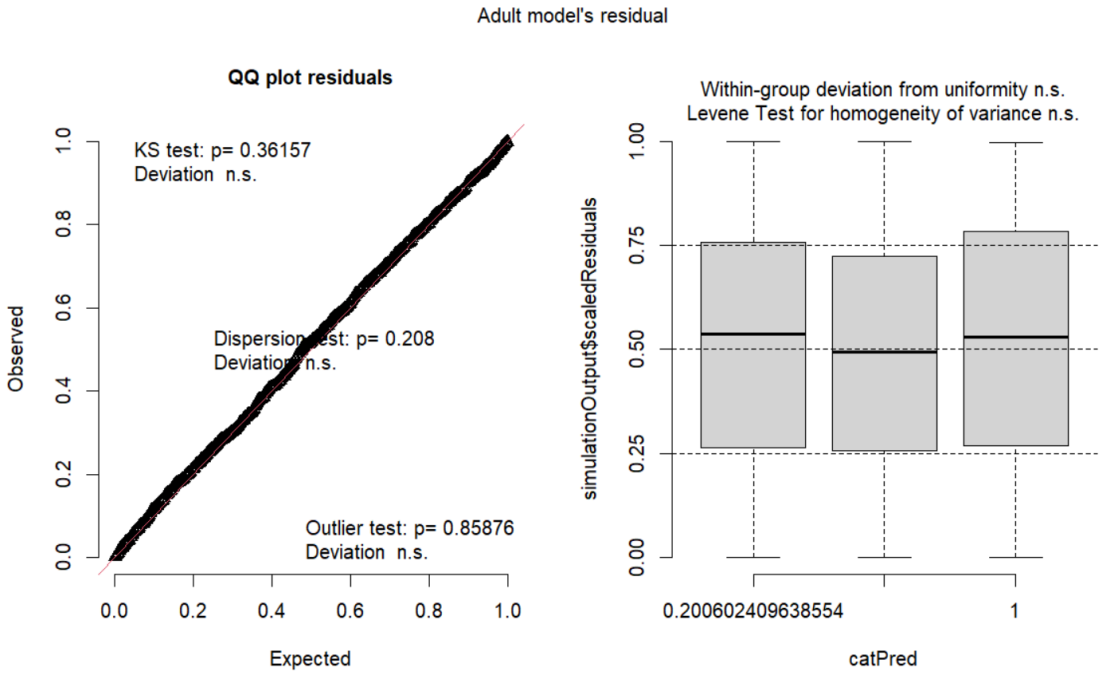


Figure S28.1 Examination of assumptions for the mixed-effects logistic model fitted on adult data, using the DHARMa package in R (Hartig, 2022). The left panel displays a QQ-plot with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right panel presents a plot of residuals against the predicted value, with the absence of clear patterns in this plot indicating a lack of heteroscedasticity issues.

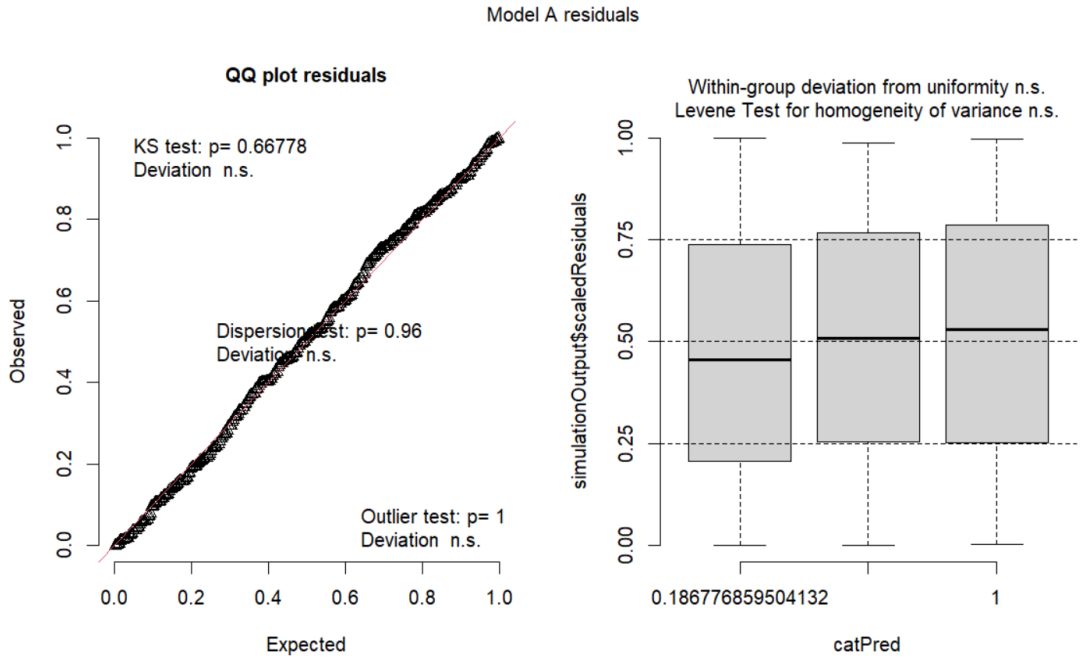
Table S28.2 presents the outputs of two statistical models applied to children's data. Model A incorporates a random effects structure that includes only random intercepts for participant and item, while Model B includes random slopes for participant per condition.

Our power stopping rule indicated that a sample size of 45 would be sufficient for fitting the more complex Model B. This model demonstrated convergence on the child data, and the standard deviation values of random slopes for participant per condition did not exceed twice those of adults (refer to our OSF for how the stopping rule was dependent on these criteria). However, we also observed that the effect sizes for children were smaller than those estimated from previous studies (see expected effect sizes in our OSF). Therefore, we examined whether sufficient power was still achieved based on the observed effect sizes in the child models.

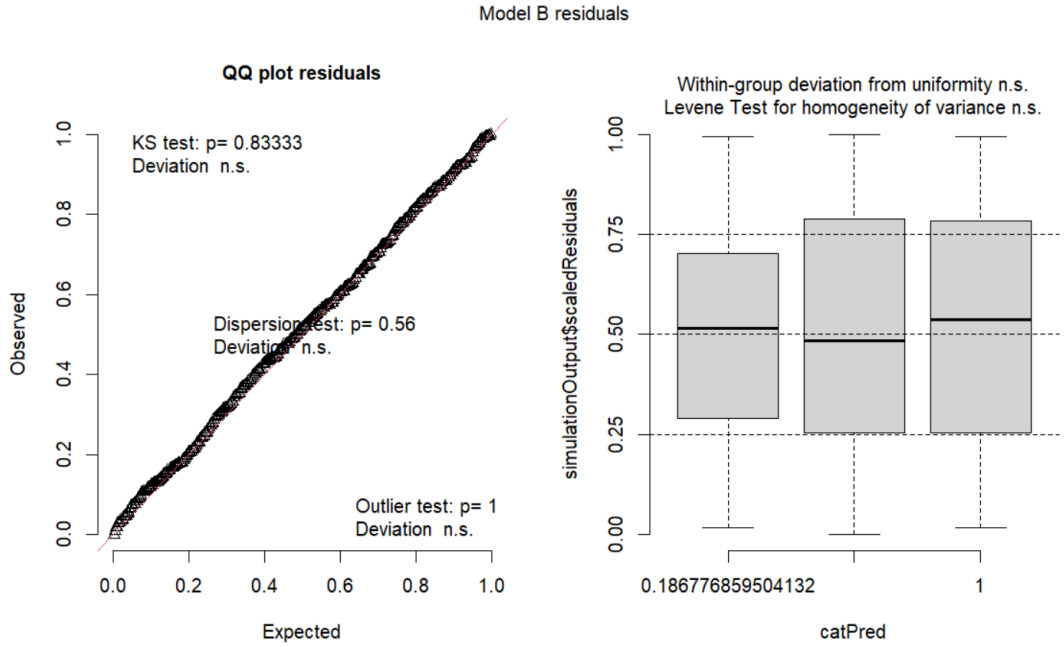
Table S28.2

Mixed-effects logistic regression models applied to children's data, filtered based on reported knowledge of target nouns and verbs. The dependent variable is sense choice (dominant, subordinate), and the independent variable is condition (control, verb-lexical, verb-semantic). Two contrasts were analysed: Control versus verb-lexical, and control versus verb-event. Model A incorporates random effect intercepts for participant and item, whereas Model B additionally includes random slopes for participant per condition.

	Child models (vocabulary knowledge filtered data)							
	Model A				**Model B**			
Predictors	*Odds Ratios*	*SE*	*CI*	*p*	*Odds Ratios*	*SE*	*CI*	*p*
(Intercept)	0.35	0.11	0.19 – 0.65	**0.001**	0.38	0.13	0.20 – 0.73	**0.004**
Control vs. Verb-Lexical	5.30	1.64	2.89 – 9.71	**<0.001**	4.68	1.69	2.31 – 9.49	**<0.001**
Control vs. Verb-Event	8.36	2.77	4.37 – 16.00	**<0.001**	7.39	2.70	3.62 – 15.11	**<0.001**
Random Effects								
σ^2	3.29				3.29			
τ_{00}	0.14 $_{id}$				0.53 $_{id}$			
	0.51 $_{Item}$				0.54 $_{Item}$			
τ_{11}					1.59 $_{id.VerbLexical}$			
					1.31 $_{id.VerbEvent}$			
ρ_{01}					-0.70 $_{id.VerbLexical}$			
					-0.98 $_{id.verbEvent}$			
ICC	0.17				0.24			
N	45 $_{id}$				45 $_{id}$			
	12 $_{Item}$				12 $_{Item}$			
Observations	362				362			
Marginal R^2 / Conditional R^2	0.168 / 0.306				0.139 / 0.350			



(a)



(b)

Figure S28.2 Examination of assumptions for the mixed-effects logistic model fitted on child data, for Model A (random intercepts only) in panel (a), and Model B (random intercepts and slopes) in panel (b). The left plots display QQ-plots with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right plots present plots of residuals against the predicted value, with the absence of clear patterns in these plots indicating a lack of heteroscedasticity issues.

For Model B (the most complex model that allowed for convergence), we found that a power of .8 was still achieved in the verb-event condition, even when considering the smaller effect size observed. However, in the verb-lexical condition, we found a power of .67 when relating the observed child effect size to the simulated power curve (see Figure S28.3, bottom row). We verified whether the loss of simulated power impacted the estimates of the model. We, therefore, also fitted Model A to the child data—a less complex model including only random intercepts—

for which a power of .8 was achieved in both the verb-lexical and verb-event conditions. As depicted in Figure S28.2, odds ratios, standard errors, and confidence intervals are similar across Models A and B, suggesting that the simulated loss of power did not influence the parameter estimation in Model B.

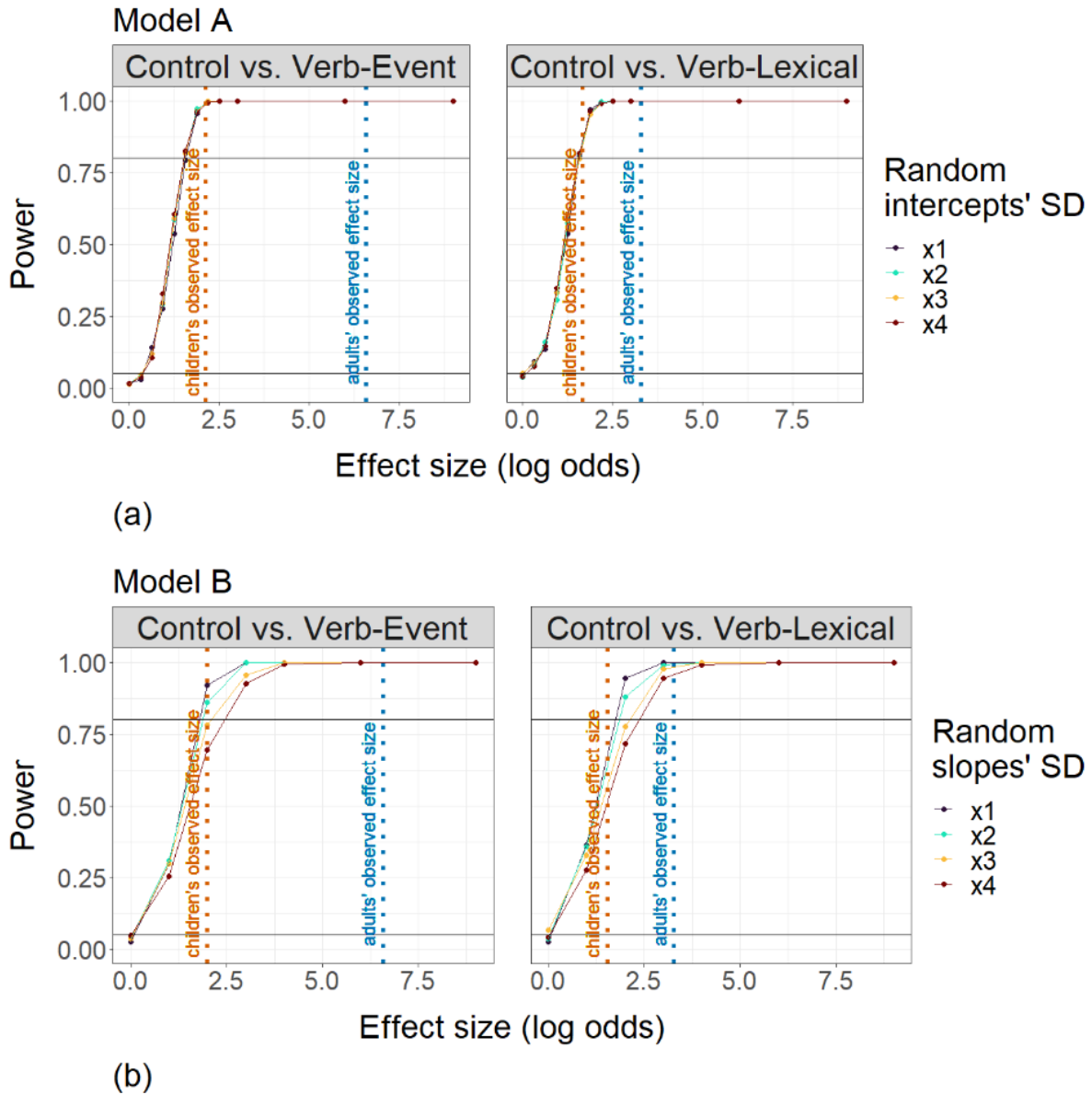


Figure S28.3 Simulated power curves as a function of effect size and standard deviations of random effects intercepts or slopes of participant per condition. The top panel (a) refers to power simulated for model A, while the bottom panel (b) refers to power simulated for model B. The top black horizontal line indicates power

of .8. The bottom black horizontal line indicates the type I error cut-off at alpha = .05. Blue dotted lines indicate observed estimates for adults, while the orange dotted lines refer to children's observed estimates for vocabulary filtered data.

The concordance in parameters between Models A and B, combined with the statistically significant result found in the verb-lexical condition, is reassuring. These findings may suggest that the parameter estimates are relatively stable, irrespective of the model's complexity, and that even with somewhat reduced simulated power, the effect was strong enough to be detected in the verb-lexical condition. However, it is important to emphasize that these aspects do not directly resolve the power issue. The lower power in the verb-lexical condition for Model B still constitutes a limitation of the study. Future research should address this by adjusting the stopping rule and power simulations based on our observed effect sizes, which will likely necessitate collecting more data from children.

Table S28.3

Refitting of the mixed-effects logistic regression models using raw data from children, showing significant effects even when not filtering children's data based on reported target nouns and verbs' knowledge.

	Child models (raw data)							
	Model A				**Model B**			
Predictors	*Odds Ratios*	*SE*	*CI*	*p*	*Odds Ratios*	*SE*	*CI*	*p*
(Intercept)	0.56	0.15	0.33 – 0.94	**0.029**	0.58	0.17	0.33 – 1.01	0.054
Control vs. Verb-Lexical	3.67	0.95	2.22 – 6.09	**<0.001**	3.44	1.09	1.85 – 6.41	**<0.001**
Control vs. Verb-Event	4.92	1.31	2.92 – 8.31	**<0.001**	4.73	1.61	2.42 – 9.21	**<0.001**
Random Effects								
σ^2	3.29				3.29			
T_{00}	0.04 $_{id}$				0.51 $_{id}$			
	0.44 $_{Item}$				0.44 $_{Item}$			

τ_{11}		1.47 $_{id.VerbLexical}$ 1.97 $_{id.VerbEvent}$
ρ_{01}		-0.66 $_{id.VerbLexical}$ -1.00 $_{id.VerbEvent}$
ICC	0.13	
N	45 $_{id}$ 12 $_{Item}$	45 $_{id}$ 12 $_{Itemr}$
Observations	460	460
Marginal R^2 / Conditional R^2	0.112 / 0.225	0.119 / -

References

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. https://CRAN.R-project.org/package=DHARMa

Appendix S29: Statistical Models of Exploratory Analyses

This section presents the results of all statistical analyses related to the exploratory investigation of various predictors' impact on adults and children's performance in the experimental task.

Table S29.1 provides the statistical model examining the predictors of performance in the verb-lexical condition for adults and children. Figure S29.1 confirms and examines the assumptions of this model.

Table S29.2 provides the statistical model that investigates the predictors of performance in the verb-event condition for adults and children. Figure S29.2 confirms and examines the assumptions of this model.

Table S29.3 provides the statistical model that investigates the predictors of performance in the verb-event condition for children exclusively. This model includes the child-reported knowledge of target nouns and verbs for each experimental item as one of the predictors. Figure S29.3 confirms and examines the assumptions of this model.

Table S29.1

Mixed-effects logistic model considering sense choice in the verb-lexical condition (dominant/subordinate) as the outcome and using age group (adult/child), relative frequency of the dominant sense (dominance), verb-sense association, and prior context associations as predictors. The model includes two-way interactions among predictors, and three-way interactions between the age group and each pair of continuous predictors. To allow for model convergence, only the participant random effect intercept is included.

	Adult-Child Model (Verb-Lexical Condition)		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.46	0.33 – 0.63	**<0.001**
Age [adult vs. child]	4.20	2.41 – 7.32	**<0.001**

Prior association	0.68	0.51 – 0.92	**0.011**
Dominance	1.89	1.37 – 2.59	**<0.001**
Verb-sense association	1.78	1.25 – 2.55	**0.001**
Age × prior association	1.06	0.64 – 1.75	0.832
Age × dominance	0.38	0.21 – 0.67	**0.001**
Age × verb-sense association	0.74	0.40 – 1.37	0.336
Prior association × dominance	0.55	0.38 – 0.80	**0.002**
Dominance × verb-sense association	1.22	0.84 – 1.78	0.300
Prior association × verb-sense association	0.70	0.42 – 1.16	0.162
(Age × prior association) × dominance	1.13	0.58 – 2.18	0.719
(Age × dominance) × verb-sense association	0.67	0.35 – 1.29	0.232
(Age] × prior association) × verb-sense association	1.60	0.68 – 3.76	0.279
Random Effects			
σ^2	3.29		
$\tau_{00\ id}$	0.46		
ICC	0.12		
$N_{\ id}$	128		
Observations	462		
Marginal R^2 / Conditional R^2	0.335 / 0.417		

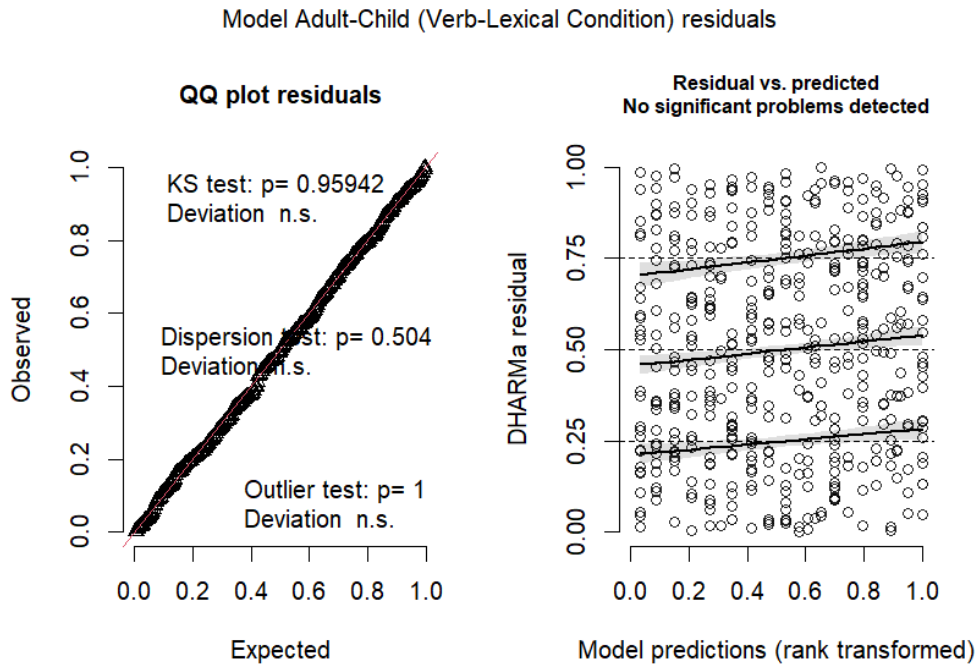


Figure S29.1 Examination of assumptions for the mixed-effects logistic model fitted on adult and child data in the verb-lexical condition, using the DHARMa package in R (Hartig, 2022). The left plot displays a QQ-plot with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right plot presents residuals against the predicted value, with the absence of clear patterns in these plots indicating a lack of heteroscedasticity issues.

Table S29.2

Mixed-effects logistic model considering sense choice in the verb-event condition (dominant/subordinate) as the outcome and using age group (adult/child), relative frequency of the dominant sense (dominance), and prior context associations as predictors. The model includes two-way and three-way interactions. Participant and item random effect intercepts are included.

	Adult-Child Model (Verb-Event Condition)		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	13.11	5.55 – 30.94	**<0.001**

Age [adult vs. child]	0.32	0.15 – 0.70	**0.004**
Prior association	0.87	0.36 – 2.13	0.763
Dominance	0.87	0.37 – 2.05	0.758
Age × prior association	1.49	0.67 – 3.31	0.332
Age × dominance	1.25	0.58 – 2.69	0.563
Prior association × dominance	0.62	0.26 – 1.48	0.279
(Age × prior association) × dominance	0.74	0.36 – 1.55	0.432
Random Effects			
σ^2	3.29		
$T_{00\ id}$	0.89		
$T_{00\ Item}$	1.44		
ICC	0.41		
N_{id}	124		
N_{Item}	12		
Observations	451		
Marginal R^2 / Conditional R^2	0.104 / 0.475		

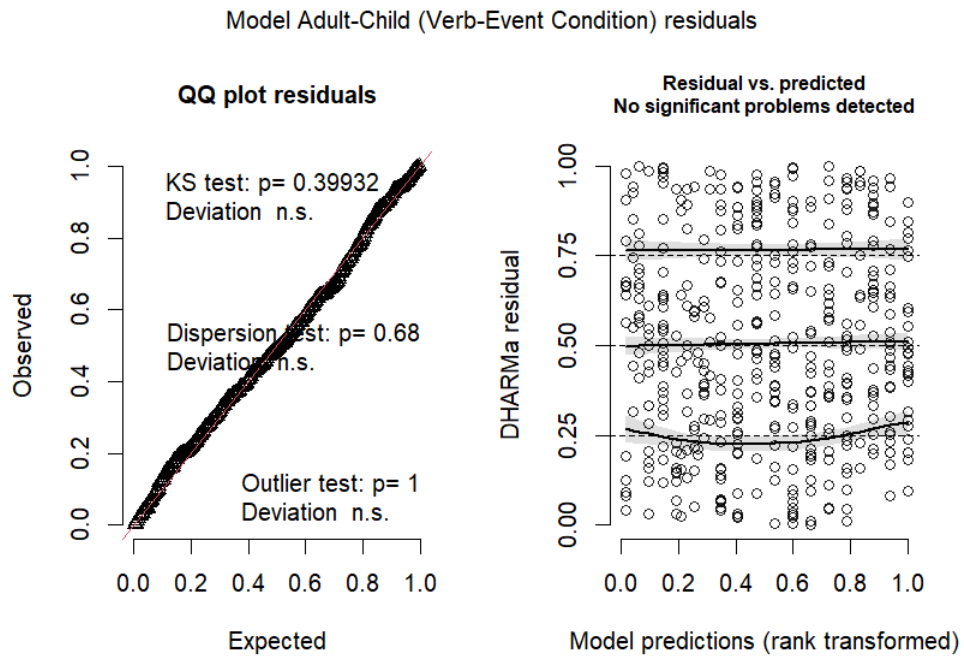


Figure S29.2 Examination of assumptions for the mixed-effects logistic model fitted on adult and child data in the verb-event condition. The left plot displays a QQ-plot with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right plot presents residuals against the predicted value, with the absence of clear patterns in these plots indicating a lack of heteroscedasticity issues.

Table S29.3

Mixed-effects logistic model considering children's sense choice in the verb-event condition (dominant/subordinate) as the outcome and using verb reported comprehension (produced/not produced), relative frequency of the dominant sense (dominance), and prior context associations as predictors. The model includes two-way interactions. Participant and item random effect intercepts are included.

	Child Model (Verb-Event Condition)		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.59	0.45 – 5.61	0.474
Comprehension [produced vs. not produced]	3.36	1.08 – 10.49	**0.037**
Prior association	1.93	0.47 – 7.88	0.361
Dominance	1.74	0.49 – 6.20	0.393
Comprehension × prior association	0.55	0.15 – 2.02	0.371
Comprehension × dominance	0.64	0.19 – 2.19	0.476
Prior association × dominance	0.47	0.18 – 1.25	0.129
Random Effects			
σ^2	3.29		
$T_{00\ id}$	0.25		
$T_{00\ Item}$	1.47		
ICC	0.34		

N_{id}	45
N_{Item}	12
Observations	155
Marginal R^2 / Conditional R^2	0.115 / 0.419

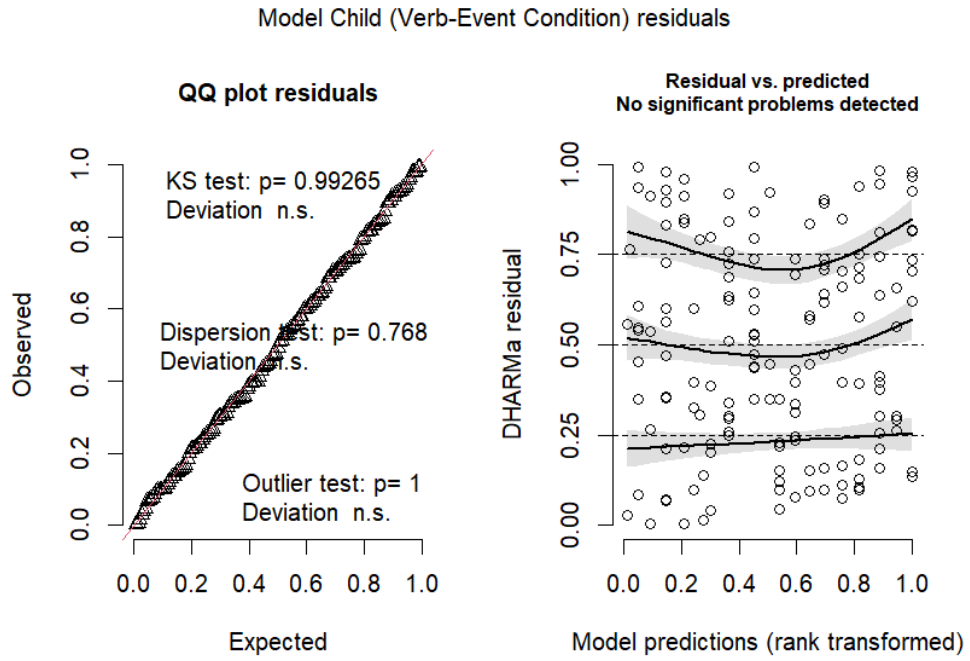


Figure S29.3 Examination of assumptions for the mixed-effects logistic model fitted on child data in the verb-event condition. The left plot displays a QQ-plot with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right plot presents residuals against the predicted value, with the absence of clear patterns in these plots indicating a lack of heteroscedasticity issues.

References

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. https://CRAN.R-project.org/package=DHARMa

Appendix S30: Age Group Differences in Sense Switching Selection

This section presents the reporting of an additional exploratory statistical model. This model has the same structure as the models fitted in the pre-registered analyses section (Appendix S28), but it includes the age group (Adults vs. Children) as a predictor.

Table S30.1 presents the statistical model using data from both children and adults. The assumptions of this model are confirmed and examined in Figure S30.1.

The results of this additional model show that adults switched from the subordinate to the dominant sense more frequently than children. This pattern was observed when transitioning both from the control to the verb-lexical condition, as well as from the control to the verb-event condition.

Table S30.1

Mixed-effects logistic regression model on adult and child data. The model employs sense choice (dominant or subordinate) as the dependent variable, while the condition (control, verb-lexical, or verb-event) is used as the independent variable. Two contrasts were analysed: control versus verb-lexical, and control versus verb-event. An additional predictor variable of age group (Adult vs. Child) has been included in the model, as well as its interaction with the variables from the two contrasts. The model's random effect structure includes random intercepts for participants and items, and random slopes of condition per participant and item.

		Adult-Child Model	
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.03	0.01 – 0.06	**<0.001**
Age [Adult vs. Child]	17.26	7.39 – 40.35	**<0.001**
Control vs. Verb-Lexical	19.73	8.13 – 47.91	**<0.001**
Control vs. Verb-Event	457.39	159.25 – 1313.67	**<0.001**

Age × [Control vs. Verb-Lexical]	0.26	0.10 – 0.68	**0.006**
Age × [Control vs. Verb-Event]	0.02	0.01 – 0.05	**<0.001**

Random Effects

σ^2	3.29
$T_{00\ id}$	0.44
$T_{00\ Item}$	0.92
$T_{11\ id.VerbLexical}$	0.50
$T_{11\ id.VerbEvent}$	2.02
$T_{11\ Item.VerbLexical}$	0.79
$T_{11\ Item.VerbEvent}$	1.26
ρ_{01}	-0.62
	-0.91
	-0.07
	-0.26
ICC	0.36
N_{id}	128
N_{Item}	12
Observations	1357
Marginal R^2 / Conditional R^2	0.492 / 0.678

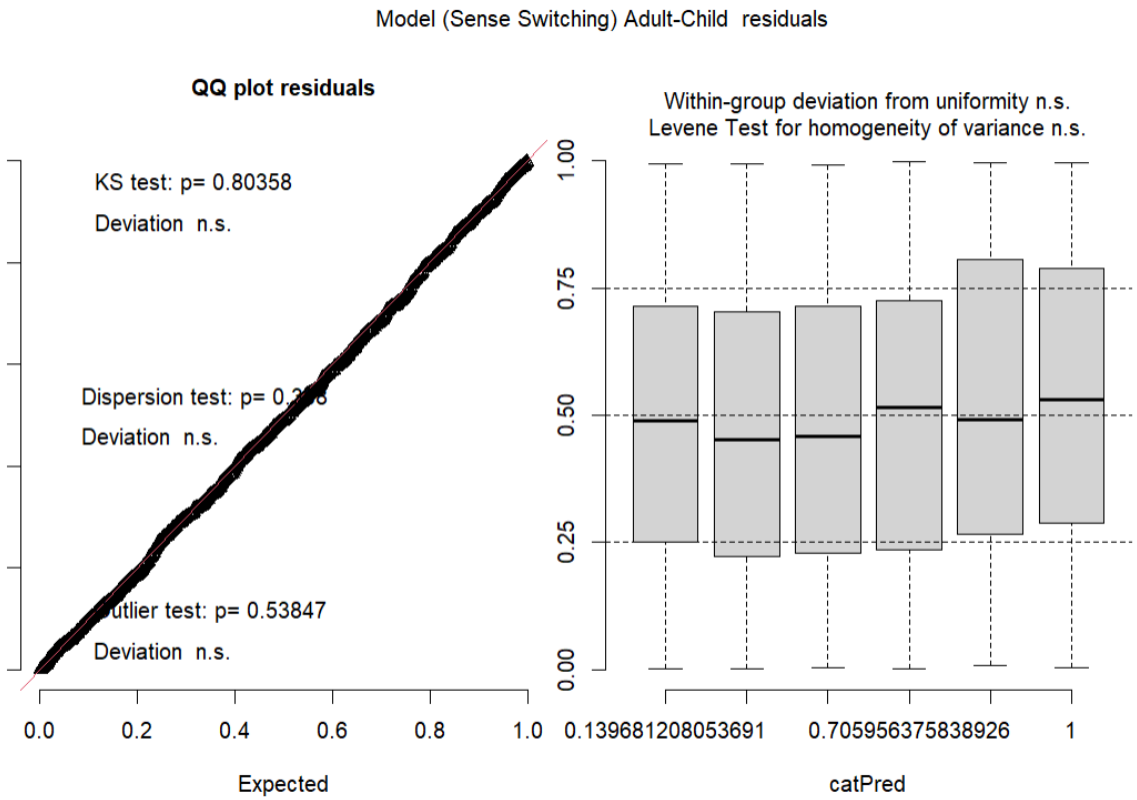


Figure S30.1 Examination of assumptions for the mixed-effects logistic model fitted on adult and child data to examine sense switching behaviour. The left plot displays a QQ-plot with no observed deviations from the expected distribution, as confirmed by Kolmogorov-Smirnov (KS) test for distribution correctness, as well as additional dispersion and outlier tests. The right plot presents residuals against the predicted value, with the absence of clear patterns in these plots indicating a lack of heteroscedasticity issues.