UNIVERSITEIT ANTWERPEN
Computational Linguistics & Psycholinguistics Research Center
Faculteit Letteren en Wijsbegeerte
Departement Taalkunde
Antwerpen, 2018

Computational Mechanisms for Bootstrapping in Language Development: Discovering Categories in Speech

Proefschrift voorgelegd tot het behalen
van de graad van doctor in de taalkunde
aan de Universiteit Antwerpen
te verdedigen door

Robert Grimm

Promotor: Prof. Dr. Walter Daelemans
Promotor: Prof. Dr. Steven Gillis

Computational Mechanisms for Bootstrapping in Language Development: Discovering Categories in Speech

Dutch title: *Computationele Mechanismen voor Bootstrapping in Taalontwikkeling: Categorieën Ontdekken in Spraak*

The research presented in this thesis was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

Abstract

When humans acquire the capacity to comprehend and produce language, they need to solve many interconnected tasks – including but not limited to the segmentation of continuous speech into discrete units, the assignment of meaning to those units, and the production of novel utterances by combining segmented units. The available evidence indicates that these processes do not unfold in isolation. Instead, knowledge acquired through progress on one task appears to sustain and facilitate – or *bootstrap* – progress in other areas.

Here, we construct computationally explicit accounts of bootstrapping processes in language development, which involve the usage of information from one domain in order to solve tasks from a different domain. In doing so, we investigate how learners discover linguistically relevant categories in speech input: Over the course of three studies, we first model how children use knowledge from the perceptual domain in order to discover linguistic units in unsegmented speech; followed by a final study wherein we model how adult listeners use resources from the domain of normal hearing to break into category perception with cochlear implants.

To examine the discovery of word-like sequences in unsegmented input, we argue that children store multi-word and multi-syllable chunks as hypothesized linguistic units. We then select various chunks from large corpora of transcribed child-directed speech, and we predict the age at which children first produce words based on the number of chunks containing each word. This approach assumes that if a particular word is contained in many stored chunks, children should easily identify it as a shared, independent sub-unit – and should subsequently begin to use it early in development. The reverse should hold for words included in fewer chunks, which should be harder to identify, and which children should begin to use at later stages.

Having developed this method in two separate studies, we apply it to evaluate which types of chunks children extract and store during early speech segmentation. It will emerge that short syllable chunks, in contrast to frequent or internally predictable sequences, are the most well-suited for predicting the time course of word learning. In addition, short syllable sequences are also the most likely to correspond to words – suggesting that children's early proto-lexica contain short, word-like chunks.

Models of speech segmentation should be constrained to reflect this bias. Beyond that, our results have implications for theories of formulaic multi-word sequences in older language users, often conceptualized as frequent word combinations. Since speech segmentation appears to be biased towards word-like instead of frequent chunks, formulaic sequences are likely discovered through usage patterns within fully segmented input – and do not originate as holistic chunks during segmentation.

Following the chapters concerned with chunks, we describe a final study that simulates speech processing in adults with cochlear implants (CIs) – neural prostheses used to partially replace a damaged inner ear. CIs can restore hearing in listeners with sensorineural hearing loss, but implant-delivered signals have poor spectral resolution and lead to degraded speech recognition performance.

Many CI users are implanted after a period of normal hearing. As a result, these listeners transition from processing high-resolution signals, delivered through the inner ear, to lower-resolution signals delivered through the implant. During the transition period, the brain presumably fine-tunes existing neural circuits, acquired during normal hearing, for use with the implant. We model this in deep neural networks, which we train first on high-resolution input, followed by additional training on low-resolution data.

Concretely, we use this design to evaluate the effect of *channel interaction*, caused by interference between neighboring channels within CIs. It turns out that neural networks which are first trained on high-resolution speech, prior to training on low-resolution data, learn more slowly if simulated channel interaction is present in low-resolution input. This effect is absent, however, if the networks are directly trained on low-resolution data. The spectral degradation caused by channel interaction may thus require additional fine-tuning of existing neural circuits, slowing the transition to CIs after normal hearing. Accordingly, a reduction of channel interaction is expected to accelerate the transition period.

Acknowledgements

Conducting the research that culminated in this thesis has been a rewarding and, at times, a rather arduous process. I would not have been able to complete it without support from supervisors, colleagues, friends, and family.

First, I would like to express my sincere gratitude for my supervisors, Prof. Walter Daelemans and Prof. Steven Gillis, for advising and mentoring me throughout the four years of my PhD. Their guidance and feedback during various meetings and personal conversations was absolutely crucial. I would also like to thank the other members of my thesis committee, Prof. Antal van den Bosch and Prof. Dominiek Sandra, for their encouragement, insightful comments, and helpful questions; as well as Prof. Niels Schiller, for agreeing to be on my PhD jury.

For stimulating discussions, as well as for all the good times we have enjoyed, I am thankful to my former and current lab mates, especially Ben, Chris, Enrique, Giovanni, Guy, Janneke, Lisa, Madhumita, Michèle, Pieter, Pietro, Simon, Stéphan, and Tim. Special thanks go to Giovanni Cassani, for many fruitful collaborations.

Research never happens in isolation; and in some way or another, all of the chapters in this thesis draw from discussions with other researchers. In particular, I am grateful to Prof. Mike Kestemont for feedback on early work that eventually resulted in chapter 2. Chapter 5 would not exist without Michèle Pettinato. Thanks also to Etienne Gaudrain, for clarifying comments on an earlier version of that chapter.

Finally, I am lucky to have parents who have always supported me in my endeavors, and who have encouraged me to pursue meaningful work, even if the outcome was uncertain. I am also lucky to be able to count on my two siblings and on my step father. I have especially fond memories of many excellent times, in spite of moving countries for study and research opportunities, with old friends from back home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This PhD thesis investigates bootstrapping operations in language development, broadly defined as cognitive processes that involve the repurposing of resources or knowledge from one domain in order to (partially) solve a task from a different domain. Bootstrapping thus concerns the manner in which a learning system uses resources or knowledge. The central idea is that this can be done in one of two possible ways: Given some task X, the learner can acquire resources that are only relevant for solving X; or it can exploit resources from a different domain, previously or currently used for some other task, and use those to solve X. The latter (but not the former) strategy constitutes bootstrapping. We use this notion of bootstrapping as a way to conceptualize language development.

Humans begin to acquire the capacity to comprehend and produce language in utero, where the fetus develops the ability to discriminate speech properties characteristic of the mother's language (DeCasper and Fifer, 1980; Mehler et al., 1988). Subsequently, infants and young children need to solve a number of language-related tasks, including the segmentation of continuous speech into smaller units such as words and morphemes; the assignment of meaning to those units; and the production of novel utterances through the combination of segmented units (Clark, 2009).

The available evidence indicates that some of these processes overlap, with progress in one area leading to advances in other areas. Language development, that is, does not unfold in neatly delineated stages, with synergistic interactions arising between different processes. The present PhD thesis informs this emerging view by

constructing computationally explicit models of bootstrapping processes that lead to empirically testable predictions, e.g. in psycholinguistic experiments.

In doing so, we focus on two areas: (1) bootstrapping speech segmentation through reliance on perceptual knowledge about syllables; and (2) bootstrapping category perception with cochlear implants (a type of neural prosthesis) through the reuse of neural circuits developed for normal hearing. In the former case, we model how children utilize existing knowledge from the perceptual domain in order to break into the linguistic domain; and in the latter case, we model how adults use existing neural resources from the domain of normal hearing to break into the domain of implant-assisted hearing.

The modeling tools used are (i) a mixture of corpus analyses and statistical methods, applied to chunks extracted from transcribed child-directed speech; and (ii) deep neural networks, trained to classify speech into abstract categories. Before providing a summary of the two modeling approaches, we first review relevant portions of the literature, summarizing examples of bootstrapping in language development and discussing bootstrapping at the neural level.

1.1 Boostrapping in language development

The literature suggests that bootstrapping operations are common in language development, with dependencies arising across what may have traditionally been considered distinct stages or modules, and progress on one task supported by and feeding back into progress on other tasks. An example for such a synergy between different processes is the relationship between speech segmentation and the assignment of meaning to speech sounds.

Consider first the speech segmentation process. Without knowledge of the linguistic units in their target language, infants and children need to extract discrete units (such as words or morphemes) from continuous speech input. In computational models, one by-product of this process is the emergence of unanalyzed units that may appear word-like but are, in fact, composed of words whose boundaries have not yet been identified (Goldwater et al., 2009). Such undersegmented chunks also emerge in children, who sometimes use multi-word sequences without knowledge

of the smaller units contained within them (Peters, 1983; Pine and Lieven, 1993; Lieven et al., 2009a).

Since one can only verbally communicate if one has assigned meaning to the sounds one produces, it is reasonable to assume that children assign some kind of meaning to undersegmented chunks. These early meanings might well be idiosyncratic and restricted to particular communicative purposes, with further segmentation and broader usage occurring over the course of development (Tomasello, 1992, 2000, 2009). But the fact remains that a linking of meaning and chunks is likely to occur before those chunks are segmented into smaller units.

The developmental processes of speech segmentation and meaning assignment could thus overlap, raising the possibility of interaction between them. Suppose, for instance, that a word which has not yet been identified as an independent unit often co-occurs with a particular object in the environment; and that the word, when encountered by the learner, is embedded in a variety of different speech sequences. Under such circumstances, the word in question might be more reliably linked to the co-occurring object than the (unsegmented) speech sounds surrounding it, prompting learners to split the word from adjacent speech material.

This scenario is supported by converging evidence from different methodologies, including computational simulations showing that co-occurrences between referents and unsegmented speech can help to constrain the formation of an early protolexicon (Räsänen and Rasilo, 2015); behavioral results showing that novel words embedded in unsegmented speech are more easily identified when presented together with a meaningful object (Cunillera et al., 2010); and neurophysiological evidence showing that the simultaneous presentation of objects and novel words, embedded in unsegmented speech, elicits neural responses indicative of both semantic and segmentation-related processes (François et al., 2017).

It thus appears that the mapping of objects to words can facilitate the segmentation of speech. More generally, we could say that semantic information can be used to (partially) solve the speech segmentation task – a clear case of bootstrapping. The literature offers various other examples of bootstrapping operations. In some cases, it has even become established nomenclature to refer to synergistic interactions between information from different domains as particular types of bootstrapping.

For example, the term *phonological bootstrapping* refers to reliance on prosodic and

phonological information in order to assign words to abstract lexical categories, e.g. *verb* or *noun*. Durieux and Gillis (2001) have shown that features such as lexical stress, vowel height, or number of phonological segments can increase the accuracy with which machine learners assign words to a noun or verb category. Importantly, human subjects likewise appear to utilize phonological cues, with English-speaking 6- to 8-year-olds preferentially mapping noun-like non-words (*haps*, *galv*) to objects and verb-like non-words (*sig*, *sming*) to actions (Fitneva et al., 2009). Thus, while we should certainly not claim that phonology is the primary driving force of lexical category formation, the cited studies nevertheless indicate that children can use phonological knowledge to inform this process.

Another source of information for lexical category formation has to do with the co-occurrence patterns of words, with initial proposals going back to Bloomfield (1933) and Harris (1954). Here, the idea is that children form lexical categories by grouping words that occur in a similar context. The utility of this insight is underscored by developments in computational linguistics, where it led to the creation of vector representations that encode a word's meaning by tracking which other words tend to occur next to it within a small window, e.g. sentence-internally (Turney and Pantel, 2010; Mikolov et al., 2013; Pennington et al., 2014). In a vector space populated with many such representations, words that are used in similar contexts cluster closely, sometimes forming groups that we might intuitively recognize as lexical categories.

With respect to language development, early corpus analyses showed that distributional patterns in child-directed speech can be exploited to group words into lexical categories (Cartwright and Brent, 1997; Redington et al., 1998). Building on this, Mintz (2003) classified words according to the *frames* they occur in. A *frame* consists of two co-occurring context words, with a variable target word in-between. Frames such as *the_and* often contain target words that we would recognize as nouns, whereas frames such as *you_it* occur mostly with verbs. Psycholinguistic studies indicate that children can indeed exploit this kind of information in order to assign words to lexical categories (Gerken et al., 2005; Mintz, 2006). In the literature, the notion that children rely on word co-occurrences as a first step towards solving lexical category formation is often referred to as *distributional bootstrapping*.

In *syntactic bootstrapping*, a term coined by Gleitman (1990), knowledge about lexical categories and how they can be combined feeds back into meaning assignment.

The idea is that, upon observing an utterance with a particular internal structure, children can use that structure to constrain the possible meanings of verbs. Evidence for this proposal comes from experiments wherein children are exposed to nonsense words embedded in differently structured utterances. For instance, Naigles and Kako (1993) had 2-year-olds listen to nonsense verbs such as *seb*, either in isolation ("Look! *Sebbing*!") or embedded within a transitive utterance ("The frog is *sebbing* the duck"). When presented with the transitive construction, but not in isolation, children preferentially matched the verb to an action where one character affects another. Or consider a more recent study by Papafragou et al. (2007), who found that 4-year-olds tend to assign a mental or cognitive meaning to nonsense verbs appearing with a sentential complement (e.g. "Matt *gorps* that his grandmother is under the covers!"). These and other similar studies – see Fisher et al. (2010) for a review – demonstrate that children can use information about how verbs combine with their arguments to constrain possible action-verb mappings.

Another aid for meaning assignment comes in the form of inflectional morphology. Consider the English verb suffix *-ing*, which signals that verbs refer to ongoing actions or events; or the plural suffix *-s*, which signals that a given noun refers to more than one object. Jolly and Plunkett (2008) reasoned that children should be able to use this knowledge in order to constrain the possible referents of a novel noun, suffixed with *-s*, to multiple rather than individual objects. In an experiment that involved the simultaneous presentation of auditory nonsense nouns and images of animals, they subsequently found that 30-month-olds associated nouns presented with a plural suffix with *groups* of animals, while nouns without the suffix were associated with *individual* animals. It would seem, then, that children can use knowledge of inflectional morphology to assist with the assignment of referents to novel word forms, a process often referred to as *inflectional bootstrapping*.

The reviewed bootstrapping processes suggest that many of the tasks involved in language development rely on bootstrapping to some extent. In the current PhD thesis, we add to this emerging picture by constructing computationally explicit proposals for bootstrapping operations in two different areas. The first concerns the manner in which infants and children break into speech segmentation. Building on evidence that proto-syllables emerge as perceptual primitives early in development (Bertoncini and Mehler, 1981; Bijeljac-Babic et al., 1993; Räsänen et al., 2018), we suggest that learners extract larger chunks by identifying particular types of

syllable sequences. Much like the studies and theories we have reviewed so far, the learning mechanisms are defined on the level of psychological mechanisms, with scant reference to neural processes.

Our second area of inquiry, bootstrapping category perception with cochlear implants, is somewhat different in that we explicitly define the learning mechanism at the level of neural processing, using deep neural networks to model the reuse and fine-tuning of existing neural circuits. In the following section, we discuss bootstrapping at the neural level, and we review a previous study that employed neural networks to model it.

1.2 Bootstrapping at the neural level

There is no reason to expect bootstrapping processes only in language development, and not in other areas of cognitive development. In fact, a recent proposal – termed *neural reuse* (Anderson, 2010) – suggests that the brain constantly repurposes neural circuits from various different domains to advance developmental processes. Two of the core ideas are (1) Anderson (2010)'s "massive redeployment" hypothesis, which states that for new functionality to emerge on an evolutionary timescale, nature avoids developing new circuits, recruiting existing neural machinery instead; and (2) Dehaene (2005)'s and Dehaene and Cohen (2011)'s "neuronal recycling" hypothesis, according to which the brain develops culturally acquired higher-level cognitive functions – e.g. reading or tool use – by reusing existing structures. Together, the two hypotheses suggest that whenever a novel cognitive function develops (in normal development or through evolution over successive generations), the brain tends to recruit existing circuitry. As a result, synergistic interactions emerge between regions that need not be topographically coherent, and that can each partially support functions from several different domains.

For example, there currently exist thousands of brain imaging studies correlating performance on a behavioral task (e.g. reading, or listening to a sequence of tones) with measures of neural activation in some specific brain area. Through simple calculations based on a data base of such results (Anderson et al., 2010), Anderson (2010) found that most areas are in fact involved in tasks from different domains (e.g. vision, attention, mathematics). This is also true for areas that were histori-

cally thought to be highly specialized – e.g. Broca's area, originally associated with language processing, is additionally involved in action- and imagery-related tasks. Pervasive reuse of resources across tasks from different domains thus appears to be the norm rather than the exception. The reuse of neural circuits from different domains is a direct correlate of bootstrapping at the psychological level – i.e., it exactly fits our definition of bootstrapping, except that it is defined at the neural rather than the psychological level.

Neural bootstrapping can be modeled in deep neural networks (DNNs), which have dramatically improved state-of-the-art performance in machine learning areas such as automatic speech recognition, object detection, or natural language processing (LeCun et al., 2015). In a DNN, the input (e.g. an image or a sound wave) is processed in several layers, with each layer processing the output of the previous layer. Often, the representations retained at later layers becomes more and more abstract. For example, intermediate layers in networks trained to map images (e.g. images of digits) to categories (e.g. integers) often encode visual primitives such as lines, edges, or corners – similar to the receptive fields of neurons in visual cortex (Lee et al., 2007; LeCun et al., 2015).

Abstract representations from DNNs lend themselves to *transfer learning*: the usage, in machine learning, of representations acquired as a result of learning to perform a task X in order to solve some other task Y (Bengio, 2012). Transfer learning, in other words, is analogous to bootstrapping in a machine learning context. This was recognized by Testolin et al. (2017), who recently used transfer learning in DNNs to model the brain's reuse of general visual primitives in order to obtain abstract letter representations. To this end, the authors trained a Restricted Boltzmann Machine (RBM) to re-produce images of natural scenes. RBMs belong to a class of neural networks that learn to transform data (in this case, images) into a low-dimensional internal representation, from which the original high-dimensional input can then be reconstructed. At each time step, the parameters of the network are adjusted to increase the similarity between the original and the re-constructed data – i.e., training proceeds in an unsupervised manner, without using labeled data.[1]

[1]Annotating large amounts of data with labels is expensive and cognitively implausible: Most data points in the world do not come with labels indicating what category they belong to. For instance, speech sounds are not labeled with the words they correspond to, and images are not labeled with the objects they depict.

As a side effect of learning to reconstruct training examples, RBMs acquire internal representations that only retain the features which are most predictive of the input. These abstract representations can boost performance on supervised machine learning tasks (which rely on labeled data), especially if the amount of labeled input is restricted (Bengio, 2012). Utilizing this, Testolin et al. (2017) first show that, unsurprisingly, an RBM trained to reconstruct images of natural scenes represents such images in terms of visual primitives (edges, lines, curves, and so on). Next, they show that the same RBM can also represent letters in terms of such features, without requiring letter-specific training. Moreover, the general letter representations derived from the RBM lead to better performance on a supervised letter recognition task than features specifically tailored to letters. This demonstrates that domain-general visual primitives, derived from natural scenes, can be used to bootstrap letter recognition – supporting Dehaene and Cohen (2011)'s proposal that culturally acquired cognitive functions (such as reading and writing letters) recycle neural resources evolved for object and face recognition.

In a later chapter of this thesis, we will follow a similar path of inquiry by using DNNs to model neural bootstrapping of category perception in cochlear implant users. There, our focus will be on how adults with cochlear implants can reuse existing speech representations, developed during early language development, in order to quickly adapt to the implant after hearing loss.

1.3 Thesis overview

This thesis investigates bootstrapping in two areas: In chapters 2 – 4, we ask how children break into the speech segmentation process, without knowledge of the linguistic categories in their target language(s); and in chapter 5, we examine how adults can use neural circuits, formed during early language development, to regain hearing with a cochlear implant. Below, we briefly summarize the two modeling approaches and main findings.

1.3.1 Speech segmentation

In chapters 2 – 4, we consider the following question: Without knowledge of the linguistic units in their target language(s), what are the first hypothesized linguistic units that children extract from speech, and what is the psychological mechanism they use to achieve this? One hypothesis, which can be traced back at least to Peters (1983), is that children break into speech segmentation by storing larger chunks before discovering the words contained within them. We explore and add to this hypothesis through a mix of corpus analyses and statistical modeling, ending with a concrete and testable prediction for psycholinguistic research.

Beginning in chapter 2, we use multi-word units (MWUs), extracted from transcribed corpora of English child-directed speech, to predict the time course of word learning. This is done by using an existing computational model (McCauley and Christiansen, 2011, 2014) that selects MWUs whose component words are particularly predictive of one another. For thousands of individual words, we then count how many MWUs contain each word, and we correlate the resulting statistic with the age at which children learn to produce the words. Our results show that early-learned words tend to be contained in a large number of MWUs, even when controlling for word frequency.

Following Peters (1983), we suggest that this is best explained as a result of early speech segmentation, with children storing undersegmented MWUs in long-term memory before discovering the words contained within them. Given a repertoire of such undersegmented chunks, children could discover smaller sub-sequences (including words) by comparing stored MWUs to one another; and if a word is contained in a large number of stored MWUs, it should be easier to discover or segment than words contained in comparatively fewer MWUs. All else being equal, children should thus learn to use easily segmentable words (found in many MWUs) before less easily segmentable words (found in fewer MWUs), explaining the observed tendency of early-learned words to be contained in many corpus-extracted MWUs.

In chapter 3, we will see that this result is fairly robust: It also emerges with a different model of MWU formation (Brooke et al., 2014), and it remains even when controlling for a number of statistical confounds. For instance, words contained in many MWUs also tend to be frequent. And since frequency is associated with

early word learning (Ambridge et al., 2015), we cannot rule out that it is really the *frequency* of words which causes them to be early-learned. We control for this by including frequency and various other psycholinguistically relevant variables as co-variates in multiple linear regression models.

Finally, in chapter 4 we define the key statistics on the syllable rather than the word level, following evidence that syllables emerge as perceptual primitives in infants (Bertoncini and Mehler, 1981; Bijeljac-Babic et al., 1993; Räsänen et al., 2018). Comparing (a) short utterances, (b) frequent utterances, and (c) utterances whose syllables are highly predictive of one another, we find that short utterances are best-suited for predicting the time course of word learning. This implies that children rely on knowledge from the perceptual domain, pertaining to syllable-like perceptual primitives, to break into the linguistic domain – i.e., that they bootstrap speech segmentation by storing short syllable sequences as hypothesized linguistic units. Our results also suggest that, on the whole, short utterances are most likely to correspond to individual words. Thus, not only do short utterances perform best at predicting when their component words are learned, but they also happen to be the most word-like.

This pattern supports accounts wherein infants extract undersegmented chunks based on sequence length, but not on the basis of frequency or syllabic predictability. Further evidence for this proposal could be collected with an artificial segmentation task, similar to the one used in Saffran et al. (1996)'s and Aslin et al. (1998)'s seminal work. These studies showed that infants rely on conditional probabilities between syllables to identify possible words. But our results suggest that, all else being equal, infants rely more strongly on sequence length.

Apart from that, the results have implications for a recent debate on the origin of cognitive representations for frequent multi-word sequences (Arnon and Christiansen, 2017). Various studies have found processing advantages for such sequences, arguing that the effect is a result of whole-sequence frequency, and that it cannot be reduced to the frequencies of included words (Bannard and Matthews, 2008; Arnon and Clark, 2011; Arnon and Snider, 2010; Arnon and Priva, 2014). This is typically taken to imply that language users store (parts of) frequent word sequences in memory.

Such representations could result from two different processes (Arnon and Chris-

tiansen, 2017): storage of frequent undersegmented chunks, which persist past the segmentation stage; or storage of frequent usage patterns *after* children have segmented the input into discrete units. Our results imply that children store word-like sequences as undersegmented chunks – which tend to be short, not frequent. We thus suggest that representations for frequent multi-word sequences emerge after the segmentation process has run its course, and that they cannot be traced back to undersegmented chunks.

1.3.2 Category perception with cochlear implants

Whereas chapters 2 – 4 cover bootstrapping operations defined at the level of psychological mechanisms (pertaining to the extraction and storage of chunks), chapter 5 considers bootstrapping at the neural level. Specifically, we examine the reuse of speech representations in individuals outfitted with cochlear implants (CIs) – neural prostheses that can partially replace damaged sensory cells within the inner ear.

For this purpose, we train deep neural networks to categorize speech sounds into abstract categories, using (a) high-resolution input modeled after the signal transmitted through the intact ear in normally hearing people, and (b) lower-resolution input that approximates the degraded implant-delivered signal. The question of interest then is to what extent the networks can reuse representations acquired from (a) in order to process (b).

We generally find that networks trained on high-resolution speech can reuse most of their representations – achieving high initial accuracy without any additional training, and improving slightly with further fine-tuning on CI-like input. Similar patterns emerge with human subjects, who quickly adapt to CIs when implanted after a period of normal hearing (Oh et al., 2003). Our results suggest that this can be understood as an instance of neural bootstrapping: Instead of completely re-learning speech processing, CI users recycle neural circuits developed for the domain of normal hearing, with some additional fine-tuning to accommodate signals from the domain of implant-assisted hearing (reflected in a transition period after implantation).

In further experiments, we explore how this bootstrapping process is affected by *channel interaction*, a phenomenon in CIs whereby information spreads across neigh-

boring frequency channels. Channel interaction degrades the spectral resolution of the CI-delivered signal and is generally associated with poor speech recognition performance (Stickney et al., 2006). It is thus not surprising that we obtain poor classification performance when simulating it in the CI-like data fed to the networks. However, in addition to a general performance degradation, we also find that networks pre-trained on high-resolution speech require additional finetuning in order to adjust their internal representations to the presence of channel interaction.

This suggests that the neural representations acquired from high-resolution input generalize poorly when channel interaction is present – unless the networks are given time to adjust to the reduced spectral resolution caused by channel interaction. In human subjects who are implanted after a period of normal hearing, we expect channel interaction to cause a similarly lengthened transition period; and we predict that in addition to improving general speech recognition performance, methods for reducing channel interaction (Landsberger and Srinivasan, 2009) should also result in speedier adaptation to CIs.

Chapter 2

Facilitatory effects of multi-word units in lexical processing and word learning

Previous studies have suggested that children and adults form cognitive representations of co-occurring word sequences. In this chapter, we propose (1) that the formation of such multi-word unit (MWU) representations precedes and facilitates the formation of single-word representations in children and thus benefits word learning, and (2) that MWU representations facilitate adult word recognition and thus benefit lexical processing. Using a modified version of an existing computational model, we extract MWUs from a corpus of child-directed speech and a corpus of conversations among adults. We then correlate the number of MWUs within which each word appears with (1) age of first production and (2) adult reaction times on a word recognition task. In doing so, we take care to control for the effect of word frequency, as frequent words will naturally tend to occur in many MWUs. We also compare results to a baseline model which randomly groups words into sequences – and find that MWUs have a unique facilitatory effect on both response variables. We discus possible underlying mechanisms and suggest, in particular, that children store MWUs as undersegmented chunks during early speech segmentation. By comparing stored chunks to one another and to incoming speech input, children could then identify smaller sequences such as words – leading to earlier identification and, eventually, earlier usage of words contained in a large number of chunks.

2.1 Introduction

In this chapter, we examine the role of lexicalized word combinations in (1) child word learning and (2) adult lexical processing. Consider, for example, sequences whose meanings cannot be derived from the meaning of their constituent words (e.g. *leave of absence*, *high five*, or *kick the bucket*). Due to their semantic opacity, such expressions are likely to be stored wholesale in long-term memory. But even non-compositional sequences such as *don't have to worry* or *I want to go* appear to be represented as units in their own right (Arnon and Snider, 2010). Here, we use the term *multi-word unit* (MWU) to refer to any sequence of words – semantically opaque or not – which is likely to be lexicalized; and by using a modified version of an existing computational model (McCauley and Christiansen, 2014) which forms MWUs by relying on transitional probabilities between words, we operationalize MWUs as particularly internally coherent word sequences.

To date, MWUs have been investigated in studies on first language acquisition (Bannard and Matthews, 2008; Arnon and Clark, 2011; McCauley and Christiansen, 2014) as well as in work concerned with adult processing (Arnon and Snider, 2010; Arnon and Priva, 2014). Findings in the two areas suggest that both adults and children possess cognitive representations of MWUs. In addition, Arnon and Clark (2011) provide experimental evidence that MWUs facilitate the acquisition of smaller linguistic units contained within them. Together, the available evidence suggests a developmental pattern from MWU to single-word representations, with a beneficial effect of the former on the acquisition of the latter. Based on this, we hypothesize that children sometimes form MWU representations before they form representations of the words contained within them, and that these MWU representations then facilitate the acquisition of single-word representations. We dub this the *MWU acquisition hypothesis*.

We furthermore propose that MWU representations interact not just with the acquisition of individual words in children, but also with the processing of individual words in adult cognition. This proposal is motivated by a strand of research concerned with the contextual distribution of words (McDonald and Shillcock, 2001; Adelman et al., 2006; Jones et al., 2012; Johns et al., 2012, 2014). Generally, increased contextual diversity (measured in terms of documents or co-occurring context words)

is associated with faster word recognition in adults. We suggest a link between findings relating to MWUs and to contextual diversity: High contextual diversity of words will lead to the formation MWU representations containing such words. Therefore, just like contextual diversity, MWUs are expected to be associated with faster lexical processing in adults. Thus, we hypothesize that MWU representations facilitate recognition of the individual words contained within them – a proposal which we refer to as the *MWU processing hypothesis*. In the following, we describe in more detail the findings on which we base the two hypotheses as well as how we evaluate them in this study.

2.1.1 Multi-word units in language acquisition

In the language acquisition literature, MWUs have emerged as a key theoretical concept of usage-based approaches (Tomasello, 2009; Behrens, 2009). Within this broad theoretical framework, learners' linguistic representations are conceived of as continually complexifying entities, with the developed cognitive system containing both lexically specific and more abstract patterns. At early stages in development, most representations are lexically specific, and child language is "(partially) formulaic and item-based" (Behrens, 2009, p. 393). In other words, child language development is thought to involve representations which are lexically specific and span multiple words.

Observations to this effect have been made by several researchers. Peters (1983) surveyed various examples, concluding that many of the early linguistic units acquired by children consist of more than one word and are often not yet analyzed in terms of their constituent parts. For example, Clark (1974) reported child utterances such as *I don't know where's Emma one*, which appear to consist of two previously heard utterances (*I don't know* and *where's Emma one*) – the implication being that the child must have treated each of these utterances as a single unit. Similarly, Tomasello (1992) reported that his daughter first began using the verb *find* as part of the utterance *find-it* during her 17th month, apparently to express a desire for an absent object (e.g. a particular toy). It was only at later stages that she started to generalize usage: First, she began to use *find-it* in combination with particular object names – as in *find-it bird*; and finally, at 20 to 24 months, she began to use *find* together with function words like pronouns and articles.

Tomasello (2000) reviews studies which suggest that a gradual development from lexically specific to more general language use is the norm. In a frequently used paradigm, young children are taught novel verbs – e.g. *tam*, as in *Jim is tamming*. Later, they are given the opportunity to use the verb in novel syntactic constructions, such as the transitive sentence *Jim is tamming the car*. Aggregating findings from several such studies shows that the proportion of children who generalize usage of novel verbs from intransitive to transitive sentences increases with age, with around 10% of children generalizing at 2 years and close to 100% generalizing at 8 years (cf. Tomasello, 2000, p. 223).

There is thus evidence that children's early utterances are lexically specific, whereas adults appear to be more easily capable of productive language use. This in turn suggests that some early representations are *fossilized MWUs*: representations which span multiple words, with usage restricted to particular situations and to particular communicative purposes. It is only at later stages in development that children begin to form single-word representations, which then leads to more productive language use.

Experimental evidence for the existence of children's MWU representations is provided by Bannard and Matthews (2008), who presented 2 and 3 year-olds with frequent MWUs like *a drink of tea* and matched infrequent MWUs like *a drink of milk* that differed in the last word. 2 and 3 year-olds were faster to repeat frequent MWUs, and 3 year-olds were also faster to repeat the first three words if they formed a frequent MWU with the fourth word. Since the final word and the final bigram (e.g. *of tea* and *of milk*) were matched for frequency, the processing advantage for frequent MWUs can only be attributed to the frequency of the entire MWU, rather than to the frequencies of its component words, suggesting that children have access to cognitive representations of MWUs. Bannard and Matthews (2008) argue, furthermore, that their subjects were likely familiar with the words contained in the MWUs, which implies the co-existence of MWU and single-word representations. The same argument can be made for adults, who are faster to recognize and produce frequent four-word MWUs in similar experiments (Arnon and Snider, 2010; Arnon and Priva, 2014).

One of the emerging patterns in language acquisition, then, is that children's early lexical representations span multiple words. In addition, Arnon and Clark (2011)

found that MWUs interact with the acquisition of morphemes. In their study, 4;6 year-olds produced more correct irregular plurals after familiar lexically specific frames than after general questions. Subjects were presented with depictions of several object instances. The object name was elicited either with a labeling question or with a lexically specific frame. For example, on one particular trial the objects were sheep, the lexically specific frame was *Count some –*, and the labeling question was *What are all these called?* 4;6 year-olds were more likely to complete the lexically specific frame with *sheep* and would provide relatively more incorrect plural forms – like the over-regularized *sheeps* – in response to the labeling question. This suggests that MWUs like *count some sheep* affect the way in which some of the smaller units contained within them are learned.

Given the evidence, it seems natural to suggest not only that children's early lexical representations often span several words, but also that such MWU representations facilitate the language acquisition process (cf. Arnon, 2009). In particular, we propose the *MWU acquisition hypothesis*, according to which the formation of MWU representations precedes and facilitates the formation of single-word representations.

2.1.2 Multi-word units in adult processing

Since adults, like children, appear to possess MWU representations (Arnon and Snider, 2010; Arnon and Priva, 2014), we suggest that MWUs also facilitate the processing of individual words in adult cognition (*MWU processing hypothesis*). We do not have experimental evidence indicative of a facilitatory effect of MWUs on adult lexical processing, but we can nevertheless derive indirect evidence from a strand of research concerned with the effect of contextual diversity on word recognition.

Several studies have investigated the effect of contextual diversity (henceforth CD) on adult lexical processing. In a corpus-based analysis, Adelman et al. (2006) counted the number of documents in which each target word occurred and found the resultant measure of CD to be more predictive of reaction times in word naming and lexical decision tasks than raw frequency counts. Their approach has been refined by Jones et al. (2012), who weighted document counts relative to semantic overlap among documents and achieved an even better fit. Since both studies relied

on naming and lexical decision data collected via *visual* word naming and recognition tasks, it is possible that the results are an artifact of modality. Johns et al. (2012) addressed this caveat by using data from an auditory word recognition task and found similar effects of CD.

Experimental evidence for a facilitatory effect of CD was collected by Johns et al. (2014). Adult subjects were presented with reading passages, each containing a low-frequency word which was replaced by a novel word form (the target). In a low-CD condition, targets were embedded in reading passages taken from a single discourse topic. In a corresponding high-CD condition, targets appeared across passages from different topics. After the reading phase, subjects performed a pseudo-lexical decision task, wherein targets presented in the high-CD condition were recognized faster and more accurately.

There is thus evidence that CD, defined on a paragraph or document level, increases the speed with which adults recognize written and spoken word forms. This is mirrored by the effect of more locally defined contextual diversity. McDonald and Shillcock (2001) counted co-occurring context words, within a small window to the left and right of each target word, and measured the divergence (relative entropy) between each target's context word distribution and a baseline frequency distribution. Target words where this divergence is large tend to be associated with longer lexical decision latencies, which suggests that words which appear in relatively specific local contexts are harder to recognize. Put differently, words whose context of use is relatively limited are hard to recognize, whereas words that can be used together with a broad range of context words are easy to recognize. This implies, to borrow McDonald and Shillcock (2001)'s phrasing, that "exposure to the context in which a given word is spoken contributes to aspects of that context being encoded in the word's mental representation." (p. 301) In the present study, we would say that co-occurrence with context words – or high CD – leads to the formation of MWU representations.

Further evidence that such a process could unfold in the human mind was collected by Hills et al. (2010). Their study takes as a starting point the previous observation that age of acquisition and adult-generated free associates are negatively correlated (Hills et al., 2009). Associates are generated by presenting a cue word (e.g. *cat*) to adult subjects, who then give back the first word that comes to mind (the target,

e.g. *mouse*). The number of different cues for which a target is provided (the *indegree* of the target) is negatively correlated with age of acquisition – i.e., words with a high indegree tend to be acquired at relatively early ages. Hills et al. (2010) show that indegree is positively correlated with the number of different context words that co-occur with the target in a corpus of child-directed speech – presumably because children link words to one another if they co-occur in the input. That is, the latter correlation is probably responsible for the former: CD likely leads to the internal linking-together of co-occurring words, which appears to facilitate the acquisition of individual words.

We can directly connect this result to Arnon and Clark (2011)'s study: Arnon and Clark (2011) found that MWUs affect the acquisition of irregular plural morphemes, while Hills et al. (2010)'s results suggest that linking words to one another – i.e. the formation of MWU representations – is likely to affect the acquisition of individual words. The formation of MWU representations, in other words, appears to affect the acquisition of smaller linguistic units (e.g. words or morphemes) contained within them. It is reasonable, then, to expect an effect of MWUs not just on word learning in children but *also* on adult word recognition. After all, a range of studies have demonstrated an effect of CD on the speed with which adults recognize words; hence, if CD leads to the formation of MWU representations, we should expect MWU representations to facilitate word recognition in adults.

2.1.3 The current study

Based on the reviewed findings, we have proposed two hypotheses: according to the *MWU acquisition hypothesis*, the formation of MWU representations precedes and facilitates the formation of single-word representations in children; and according to the *MWU processing hypothesis*, MWU representations facilitate the processing of individual words in adults. In this study, our primary objective is the evaluation of the two hypotheses via correlational analysis.

Concretely, we use an existing computational model (McCauley and Christiansen, 2011, 2014), with minor modifications designed to make the output more noise-resistant, to extract MWUs from a corpus. The kinds of MWUs the model discovers have previously been used (cf. McCauley and Christiansen, 2014) to match results

from Arnon and Clark (2011) and Bannard and Matthews (2008), which gives credence to their suitability as approximations of the types of MWUs human learners might discover. After running the model on two different corpora, we use the number of MWUs within which a given word is contained as an independent variable. If the *MWU acquisition hypothesis* is true, words contained in many different MWUs should be easier to acquire than words contained in fewer MWUs. Likewise, if the *MWU processing hypothesis* is true, such words should also be easier to process. To see why, suppose that the model discovers a large number of different MWUs which each contain a particular target word X. We do not know if human learners, given similar input, would discover the exact same MWUs; but our expectation is that the more MWUs containing X are discovered by the model, the more likely human learners would be to form cognitive representations of MWUs that also contain X. And if the *MWU acquisition hypothesis* is true, the formation of such MWU representations should facilitate the acquisition of words contained within them. Thus, X should be easier to acquire than words which appear in fewer MWUs. Similarly, if the *MWU processing hypothesis* is true, representations of MWUs containing X should facilitate processing of X – and hence, X should be easier to process than words contained in fewer MWUs.

To track word learning in children and lexical processing in adults, we use two response variables: age of first production (AoFP) and adult reaction times (RTs) from a lexical decision task. AoFP serves as an index of word learning: If a word is first produced relatively early in development, we assume that this is in part because it is easy to learn when and how to use it. Likewise, if first production occurs comparatively late, we assume that this reflects difficulties in establishing when and how to use the word. Next to AoFP, we use RTs from a lexical decision task to measure word recognition in adults: Words with fast RTs are easier to recognize, relative to words with slow RTs. Correlating the number of different MWUs per target word with AoFP and adult RTs thus allows us to measure (a) the potential impact of MWUs on child word learning and (b) their potential impact on adult word recognition. In line with our two hypotheses, we expect that words contained in many MWUs will be first produced at relatively early stages in development and will be recognized relatively quickly in a lexical decision task. In other words, we expect the independent variable to correlate negatively with both RTs and AoFP.

Our first and primary goal is to test this prediction via correlational analysis. In

doing so, we attempt to control for the frequency of target words, since frequent target words will also tend to appear within many MWUs. Beyond that, we aim to compare the effect of MWUs across word recognition and word learning – i.e. we ask which of the two areas is potentially more strongly affected by MWUs. Here, we have no a priori reason to expect a stronger effect on one over the other area: Given that children's early utterances are lexically specific MWUs, it could be that language acquisition interacts particularly strongly with MWUs. But it is also possible that MWU representations become more entrenched over the course of development and thus become even more central to adult processing.

2.2 Analysis I: Extracting multi-word units from child- and adult-directed speech

In this first analysis, we use a modified version of an existing computational model (McCauley and Christiansen, 2011, 2014) to extract MWUs from a corpus of transcribed child-directed speech and a size-matched corpus of transcribed informal conversations among adults. The types of MWUs the model discovers – sequences with particularly strong transitional probabilities between constituent words – have previously been used to model results with respect to the role of MWUs in child language acquisition (McCauley and Christiansen, 2014), providing empirical support for their cognitive relevance. By running the model on a corpus of child-directed speech, we aim to approximate the types of MWU representations that children would discover; and by running the model on a corpus of transcribed speech exchanged among adults, we aim to approximate the types of MWU representations that adults might possess. The two sets of MWUs then serve as the basis for calculating the independent variable used in the subsequent correlational analyses: the number of MWUs per target word.

2.2.1 Method

Model

The computational mechanism we use to discover MWUs is a modified version of a model developed by McCauley and Christiansen (2011, 2014). In a first phase, their model – called Chunk-Based Learner (CBL) – extracts MWUs from a corpus of child-directed speech. In a second phase, it generates child-produced utterances based on discovered MWUs. The full model is described in McCauley and Christiansen (2011, 2014), along with how it can be used to generate child productions and model results from Bannard and Matthews (2008) and Arnon and Clark (2011). Here, we provide a brief description of the component responsible for the discovery of MWUs, as well as how we modified it in order to reduce the impact of noisy input.

The CBL is psychologically motivated in that (1) it processes a given corpus in an incremental fashion – i.e., utterance by utterance and word by word –, and (2) it relies on backward transitional probabilities (BTPs), which human learners are sensitive to (Pelucchi et al., 2009). In addition, it does not require parameters governing MWU length or frequency. For example, consider the selection of common word sequences as a possible way of extracting MWUs from a corpus. With such a method, we would have to define both a maximum MWU length as well as an arbitrary frequency threshold for a word sequence to count as an MWU. The CBL, in contrast, utilizes BTPs between words as the only criterion for inclusion into MWUs. We can conceptualize the model as a psychologically grounded method for segmenting a corpus into MWUs which are, by virtue of the BTPs between component words, more internally coherent than randomly selected word sequences.

More formally, processing an utterance u is initiated by incrementing the frequency count of the first word w_1 by 1 and creating a new MWU with w_1 as its only member. For each subsequent word w_i at utterance position $1 < i \leq length(u)$, the model keeps track of the number of times w_i has been encountered so far, as well as how often the immediately preceding word w_{i-1} has occurred one position to the left of w. The model then calculates the BTP of w_i and w_{i-1}: $p(w_{i-1}|w_i)$. If this conditional probability is larger than the average BTP, across all words which have occurred one position to the left of w in all utterances so far considered, w_i is added to the current MWU. Else, the current MWU is added to a set M, and a new MWU is created –

again with w_i as its only member. Once the model has formed a first set of MWUs, it uses them as a resource to constrain the formation of future MWUs: If a sequence of words w_{i-1}, w_i constitutes part of an existing MWU, future occurrences of w_{i-1} and w_i are grouped into an MWU regardless of the BTP between the two words. In this way, the model discovers MWUs of size 2 or larger, as well as single-word units, collected in M.

As mentioned, we introduce a minor modification to the CBL. In the original version, two given words form part of an MWU if the BTP between them is larger than average. However, for words which the model has not yet encountered very often, BTP may be quite noisy. This is a matter of sample size: Statistics estimated from small samples can be strongly influenced by aberrations in the data, and BTPs calculated on the basis of very low frequency counts could be biased by a number of possible peculiarities (e.g. a particular topic of conversation, a non-standard dialect, transcription errors, and so on). As words are encountered more often, the effect of noise will diminish, and BTPs will become more representative of general language use. To guard against noise at early stages of learning, when BTPs may be unstable, we weigh the decision to group words into MWUs by the amount of prior experience: A given word w_i and the immediately preceding word w_{i-1} are included in an MWU only if the BTP between them is larger than the mean BTP *plus* the reciprocal of the frequency count of w_i. That way, words can still be included in MWUs even if the model has had relatively little exposure to them, but only if the BTP with preceding words is comparatively large. As words are processed more often, this effect diminishes exponentially – in line with the increasing stability of BTPs.

We consider the MWUs discovered in this fashion as approximations of the types of MWU representations created by human learners – the underlying assumption being that internally coherent sequences of words are good candidates for cognitively plausible MWUs. This assumption derives its justification from the fact that the MWUs discovered by the CBL can be used to model results from Bannard and Matthews (2008) and Arnon and Clark (2011) – cf. McCauley and Christiansen (2014) – two key studies which motivated our hypothesis regarding the effect of MWUs on word learning. This track record notwithstanding, there is of course no guarantee that a particular MWU discovered by the CBL is really represented in the minds of language users, but it is our expectation that model-derived MWUs are more likely to be cognitively represented than randomly selected word sequences. In the following

analyses, we attempt to confirm this via comparison to a random baseline. The baseline model operates just like the CBL, except insofar as it randomly decides whether or not to group two successive words into an MWU. That is, the baseline model also incrementally processes a given input utterance, considering each word for inclusion into an MWU. But instead of using BTP to decide whether or not the current and the preceding word form part of an MWU, it relies on a random coin toss to make that decision. To avoid confusion, we refer to the units discovered by the baseline as *word sequences*, whereas we continue to use the term *MWUs* to refer to the units discovered by the CBL.

Corpora

McCauley and Christiansen (2011, 2014) used a corpus of child-directed speech (CDS) to discover MWUs with the CBL. Children learn primarily in the context of CDS, which differs quite markedly from the type of speech used by adults to address other adults (adult-directed speech, henceforth ADS). Among other things, CDS consists of shorter phrases, contains more pauses, shows a wider range of pitches, and is composed of a limited vocabulary (Saxton, 2010). These differences are, in turn, likely to affect the language acquisition process at various levels (Matychuk, 2005; Saxton, 2009). It makes sense that McCauley and Christiansen (2011, 2014) – modeling child-produced speech and child-elicited experimental results – chose an input corpus that reflects the unique linguistic environment of English-speaking children.

In the current study, however, we are interested in adult processing *in addition to* language acquisition. If we were to use a corpus of CDS, we would implicitly claim that adult lexical processing and child word learning are equally strongly affected by MWUs found in CDS – even though adults' primary linguistic input differs substantially from CDS. We address this challenge by using two different input corpora: one that is similar to the collection of corpora used by McCauley and Christiansen (2011, 2014), and an additional size-matched corpus of ADS. When carrying out correlational analyses, we then assume that MWUs in CDS are a more direct determinant of word learning in children, whereas MWUs in ADS are a more direct determinant of adult lexical processing. Consequently, when measuring the effect of MWUs on child word learning, we base the analyses on MWUs found in

CDS; and when examining the effect of MWUs on adult lexical processing, we focus on MWUs found in ADS.

The CDS corpus is based on eight British English corpora from the CHILDES database (MacWhinney, 2000a) (cf. appendix A for an enumeration). An aggregated CDS corpus is created by first ordering the transcripts from all included corpora by the age of the child addressed in each transcript. We then extract all utterances made by any adult whose utterances were transcribed (usually the mother or father of the child or children in question, sometimes another relative or an experimental confederate). The full CDS corpus contains 4,869,472 tokens of child-directed speech, produced by 201 adults in interactions with 133 different children.

By aggregating different corpora, we are conflating the language directed at children from different backgrounds. However, limiting ourselves to particular CHILDES corpora severely restricts the amount of available data, while working with data from several corpora is likely to increase the detectability of robust, corpus-independent patterns. At the same time, we include only British English corpora and exclude American English corpora, which increases comparability with the size-matched ADS corpus.

The ADS corpus is based on the informal spoken component of the 100-million-word British National Corpus (henceforth BNC) (Burnard, 2007), a resource designed to represent a wide cross-section of spoken and written British English. Due to the methodological challenges inherent in collecting representative spoken samples, the BNC mostly consists of written material. The spoken component comprises 10.58 million tokens, 6.28 million of which cover rather formal spoken English. The remaining 4.30 million tokens consist of transcribed conversations among adults, collected from 124 adult respondents who were given a recording device, together with instructions to record their everyday conversations. Except for the respondents, interlocutors were not aware of being recorded. Transcribed material was then included in the corpus only if all interlocutors had given consent upon being informed of the recordings. This informal spoken component of the BNC is a suitable source of ADS to compare against the CDS corpus.

The CDS and ADS corpora are taken from the same variety of English (British English) and are similar with respect to the number of tokens and interlocutors. Important differences have to do with the number of word types and the mean

utterance length (cf. Table 2.1). Despite containing a similar number of tokens each, the CDS corpus contains 30 % fewer word types than the ADS corpus, with utterances in the ADS corpus being on average two tokens longer than utterances in the CDS corpus. These differences are expected and likely reflect general differences between ADS and CDS (Saxton, 2010).

Table 2.1: Relevant statistics for the ADS and CDS corpora.

measure	adult-directed speech	child-directed speech
nr. adult speakers	124	201
nr. tokens	4,233,645	4,869,472
nr. types	34,267	25,109
median utterance length	5 (IQR: 7)	4 (IQR: 4)

2.2.2 Results and discussion

Running the CBL on the ADS and CDS corpora results in two different sets of MWUs – Table 2.2 summarizes relevant statistics about their distribution (upper section). There are relatively fewer MWU tokens (first row) and relatively more MWU types (second row) in ADS, while the median number of tokens per MWU (third row) is a bit smaller in CDS. And even though the overall statistics are roughly similar, the baseline model extracts comparatively more unique word sequences from CDS, with the median length of sequences from both ADS and CDS being larger than the corresponding lengths of CBL-extracted MWUs. This indicates that the MWUs discovered by the CBL deviate from randomly selected word sequences.

The MWUs discovered by the CBL have a tendency to span comparatively more tokens as they decrease in frequency. For example, the five most frequent MWUs in the CDS corpus are *that's right* (frequency count: 5,705), *oh dear* (4,566), *is it* (4,538), *isn't it* (4,445) and *come on* (4,410). The five most frequent MWUs in the ADS corpus are *you know* (2,644), *oh yeah* (2,028), *is it* (1,797), *it is* (1,754), and *isn't it* (1,650). Among the lower-frequency MWUs, we find constructions such as *knife and fork* (CDS, with a frequency of 7), *glass of wine* (CDS, 4), *come across* (CDS, 3), *point of view* (ADS, 12), *I apologize* (ADS, 2), or *beg pardon* (ADS, 2).

Table 2.2: Relevant statistics about the distribution of MWUs in ADS and CDS.

model	*measure*	ADS	CDS
Chunk-Based Learner	nr. MWU tokens	834,205	1,117,465
	nr. MWU types	495,610	467,849
	median MWU length	5 (IQR: 4)	4 (IQR: 3)
random baseline	nr. word sequence tokens	663,953	955,698
	nr. word sequence types	520,482	592,735
	median word sequence length	6 (IQR: 4)	5 (IQR: 3)

Upper section: MWUs extracted by the Chunk-Based Learner. Lower section: MWUs extracted by the random baseline model

The most frequent word sequences extracted by the baseline model often overlap with the most frequent MWUs discovered by the CBL: The baseline is bound to extract many of the short and frequent MWUs which the CBL discovers, simply because they are so frequent that even a random method will discover them by chance. As we consider less and less frequent MWUs, however, the degree of overlap weakens. For example, the overlap between the top 5,000 CBL-derived MWUs and the top 5,000 baseline-derived word sequences is 70 % for ADS and 66 % for CDS, but this shrinks to 36 % and 49 % if we consider the top 100,000 units. There is thus a principled difference in the types of MWUs discovered by the CBL and the word sequences extracted by the random baseline, in spite of the considerable overlap between the most frequent items. In the subsequent analyses, this should be reflected in a difference between results obtained with the CBL-extracted MWUs and results obtained with the baseline word sequences.

To derive the key independent variable for the remaining analyses, we count the number of different MWUs within which each target word appears. For example, suppose our target words are *boy*, *sit*, and *nice*. We would then consult the two sets of MWUs and count, for each of the three words, all MWUs which contain the word. We find that *boy* appears within 1,725 different CDS MWUs and 510 different ADS MWUs, *sit* within 3,046 CDS MWUs and 1,122 ADS MWUs, and *nice* within 3,838 CDS MWUs and 2,527 ADS MWUs. To illustrate the types of MWUs we have counted, Table 3.3 lists two high- and two low-frequency MWUs, in CDS and

Table 2.3: The two most frequent and two of the least frequent MWUs containing the target words *boy*, *sit*, and *nice*.

target word	ADS MWU	CDS MWU	ADS freq.	CDS freq.
boy	good boy	good boy	116	736
	old boy	clever boy	35	301
	there is a clever boy	poor little boy	3	3
	good old boy	oh you naughty boy	2	2
sit	sit down	sit down	106	324
	sit there	sit up	19	107
	sit in the back	sit on your chair	3	3
	you sit there	can I sit down	2	2
nice	very nice	that's nice	132	354
	that is nice	is that nice	88	219
	isn't it nice	do I look nice	3	3
	you look nice	looks quite nice	2	2

ADS, for each of the three example words.

In the immediately following analysis, we (1) evaluate the impact of this variable on child word learning and adult word recognition and (2) verify the assumption that ADS is the relevant linguistic input for adults, while CDS is the relevant input for children. Following this, in analysis III, we compare the results from analysis II to results obtained with the baseline model. Lastly, in analysis IV, we compare the effect of MWUs on word learning to their effect on word recognition.

2.3 Analysis II: The effect of multi-word units on word learning and word recognition

We now turn to the first of three correlational analyses. Our primary objective is to evaluate the potential impact of MWUs on word learning in children as well as on word recognition in adults. In line with the *MWU acquisition hypothesis*, we expect a beneficial effect of MWUs on the former; and given the *MWU processing hypothesis*, we also expect a facilitatory effect on the latter.

We use corpus-derived age of first production (AoFP) estimates to track word learning and reaction times (RTs) from a lexical decision task to track word recognition. Given a set of words with associated RT and AoFP values (henceforth *target words*), it is easy to count the number of MWUs within which each target word appears (a measure denoted by *#MWUs*). In addition, we also count how often individual target words occur within each corpus (denoted by *#Freq*). These predictors are then correlated with RTs and AoFP. We perform both full as well as partial correlations – correlating #Freq with the dependent variables while controlling for #MWUs, and correlating #MWUs with the dependent variables while controlling for #Freq. In each case, we expect the correlation coefficient to be negative: More frequent words should be recognized more quickly and learned earlier than less frequent words, and similarly for words appearing within a large number of different MWUs.

Recall that we use MWUs estimated from ADS and CDS to account for differences in the input received by adults and children. We assume that the linguistic input received by adults is best approximated by the ADS corpus, whereas the linguistic input received by children is best approximated by the CDS corpus. Thus, the two independent variables (#MWUs and #Freq) are estimated from CDS and ADS, resulting in two variants each: ADS-#Freq and CDS-#Freq, as well as ADS-#MWUs, and CDS-#MWUs. Since RTs are elicited from adult subjects, we consider the ADS variants relevant for their analysis; and since AoFP is based on child productions, we consider the CDS variants relevant for the correlations with AoFP. Given these assumptions, it would be methodologically dubious to correlate the two CDS predictors with RTs, or the two ADS predictors with AoFP. Nevertheless, instead of ignoring these possible correlations, we compare the CDS predictors to their ADS counterparts. If our reasoning is correct, we should expect RTs to correlate more strongly with the ADS predictors, while AoFP should correlate more strongly with the CDS predictors.

2.3.1 Method

Target words

The set of target words consists of all word forms which occur in both the CDS and the ADS corpus and for which both AoFP and RT estimates are available

(7,481 words). Target words are based on raw word forms, without any kind of pre-processing (e.g. stemming, lemmatization, or part-of-speech tagging).

Age of first production

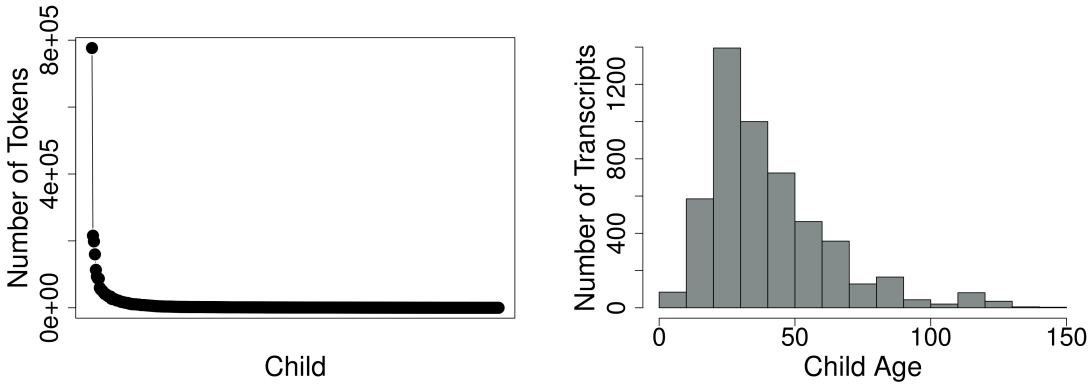
The first of two dependent variables, AoFP, measures word learning in children. Our reasoning is that words which are first produced early in development are easier to learn than words which are produced later. Ease of learning is likely determined by various factors, such as frequency in the child's input. Thus, a negative correlation between e.g. CDS-#Freq and AoFP would indicate that early-learned words are frequent in CDS; and a plausible interpretation would be that frequency of exposure leads to early word learning by exerting a facilitatory effect on one or more of the various processes involved in word learning.

We estimate AoFP from the transcribed speech of children addressed in a second collection of CHILDES corpora, without overlap with the CDS corpus. The rationale for using a second collection of corpora has to do with a possible confound. In the current study, we evaluate the effect of MWUs in ADS and CDS on two dependent variables – AoFP and RTs. If we were to use the speech produced by the children addressed in the CDS corpus to estimate AoFP, the difference in effect on AoFP between MWUs in CDS and MWUs in ADS might simply be due to the fact that both the dependent (AoFP) one of the independent variables (MWUs in CDS) have been estimated from related corpora. To avoid this issue, we estimate AoFP from an unrelated corpus.

The AoFP corpus is based on 44 American English corpora from the CHILDES database, which together contain 3,188,944 tokens produced by 463 children. The number of tokens contributed by individual children varies, with large longitudinal studies contributing a few thousand tokens for a single child each and some cross-sectional studies contributing only hundreds of words per child (cf. Figure 2.1a). The children in most transcripts are between ten and 70 months old, with relatively fewer transcripts for children between one and ten or 70 and 150 months (cf. Figure 2.1b).

In a balanced data set, utterances for each child would span the same age ranges, with the same number of words for each child – and consequently, the distributions in Figures 2.1a and 2.1b would be completely flat. We could then identify first usages

Figure 2.1: Rank distribution of the number of tokens uttered by each child (left) and distribution of transcripts by child age in months (right).



(a) nr. of tokens by child

(b) nr. of transcripts by age

in each child's data and take the mean, obtaining an average AoFP value for every target word. But because of the current uneven distributions, such a procedure would introduce noise into the AoFP estimates. Suppose, for example, that we have data from one child for ages 2 – 5, and data from ten additional children for ages 5 – 6. Suppose, furthermore, that the younger child uses a particular word for the first time at age 3, while all the older children use it from the earliest recorded time on (age 5). In cases like these, it is plausible to assume that most of the older children would have been using the word in question since well before their data were collected. Thus, by including their first usages in the average AoFP, we would artificially inflate the estimate.

To avoid this issue, we treat a word as having been learned at the earliest developmental stage at which any child within the corpus produces it. In doing so, it is possible that we still include first usages which are also artificially inflated (because the child may have been using the word prior to the commencement of data collection), but at least we do not exacerbate the problem by averaging across AoFP values. In spite of these precautions, it is still possible that our AoFP estimates do not, after all, correspond very closely to the ages at which children learn words. To ensure methodological validity, we thus correlate AoFP with age of acquisition estimates collected via an elicited production task. The correlation is strongly positive (see section 3.1.4 below), strengthening our confidence in the AoFP estimates.

Developmental stage is defined in terms of mean length of utterance (MLU) – the average child utterance length, in tokens, within a transcript. We induce MLU rather than age estimates because children who are close in age may nevertheless be far apart in terms of language development. Being a more robust estimator, MLU controls for such developmental differences (Parker and Brorson, 2005). Since transcripts contain varying numbers of utterances, the average utterance length per transcript is biased with respect to transcript length. We rectify this issue by estimating MLU for each transcript via statistical bootstrapping, wherein the sampling distribution of the population is approximated by drawing random samples from the data (Davison and Hinkley, 1997). Each bootstrap is based on 1,000 random samples with replacement, with the sample size equal to the number of child utterances per transcript. We thus induce MLU rather than AoFP estimates but will, for simplicity, refer to a word's MLU value as its AoFP. To induce a value for a given word, we calculate the set of MLUs γ for all transcripts within which the word appears and assign it the smallest value in γ. We perform this procedure for all 29,055 word types identified via this method.

Adult reaction times

The second dependent variable – RTs from a lexical decision task – measures word recognition in adults. Following the word recognition literature, we assume that words with fast RTs are easier to recognize than words with slow RTs. A negative correlation between e.g. ADS-#Freq and RTs would then indicate that words which are frequent in ADS tend to be quickly recognized; and a possible interpretation would be that frequency of exposure leads to fast word recognition in adults by strengthening the word's representation in long-term memory.

RTs are taken from the English Lexicon Project (Balota et al., 2007), which contains RTs from a lexical decision task for 40,481 mono- and multi-syllabic English words. Data were collected from participants recruited at six different U.S. universities (mean age $\approx$ 23 years) – meaning that just as AoFP, RTs were collected from native speakers of American English.

In the lexical decision task, subjects were presented with a string of letters corresponding to either an English word or a non-word, following which they were required to press a button if they thought the string was a word and another button

if they thought the string did not correspond to a word. The time taken between stimulus presentation and button press was averaged across participants, resulting in a mean RT estimate for each word.

Validity of AoFP and relationships among dependent variables

With the dependent variables in place, it is important to ensure methodological validity of the AoFP estimates. The advantage of using a collection of CHILDES corpora to estimate AoFP lies in the large number of words we can cover, but it is nevertheless desirable to compare AoFP to estimates elicited in controlled experiments. In addition, we ought to verify that AoFP and RTs are not too strongly correlated – to avoid potential difficulties in interpreting correlations with the independent variables, as well as to ensure that AoFP and RTs measure different underlying processes.

Our approach is methodologically related to work concerned with *age of acquisition*. Beginning with Carroll and White (1973), a large number of researchers have used adult estimates of when they learned to understand or use specific words to predict adult performance on various tasks (Barry et al., 2001; Bonin et al., 2004; Brysbaert and Cortese, 2010). However, this way of estimating age of acquisition may raise methodological concerns, as adult memory for childhood learning is very inaccurate (Baayen et al., 2016). To address this issue, Morrison et al. (1997) had children of varying ages perform a picture naming task. If a child is able to produce the correct noun (the picture name), he or she can be said to have learned the word. Presumably because of time constraints, Morrison et al. (1997) provide age of acquisition for a restricted set of 297 picturable nouns.

While the restricted focus makes their data less suitable for our analyses, Morrison et al. (1997)'s data are the only age of acquisition estimates for English that are directly derived from children. If our AoFP estimates are methodologically valid, we should expect their ordering to be strongly positively correlated with the order of Morrison et al. (1997)'s age of acquisition data. And indeed, for the 277 words shared between the two data sets, Spearman's rho is 0.61 ($p \leq 10^{-20}$), strengthening our confidence in the validity of AoFP. RTs from the English Lexicon Project correlate less strongly with both age of acquisition ($\rho = 0.35$, with $p \leq 10^{-8}$, for

284 shared words) as well as AoFP ($\rho = 0.31$, with $p \leq 10^{-20}$, for 10,883 shared words), suggesting that age of acquisition / AoFP and adult RTs measure different underlying processes.

Statistical analysis

For the choice of correlation coefficient, we use a particular formulation of Kendall's coefficient, Kendall's τ-b, as it addresses potential pitfalls with the data. Consider Table 3.4 as a snapshot of the available data, where each row represents a target word. From left to right, each column contains: the target word, its frequency of occurrence in CDS (CDS-#Freq), the number of unique MWUs within CDS that contain it (CDS-#MWUs), and the word's age of first production (AoFP). Two of the correlations we wish to examine are (1) CDS-#Freq vs. AoFP and (2) CDS-#MWUs vs. AoFP. Being frequency counts, both CDS-#MWUs and CDS-#Freq are non-normally distributed. In addition, data points are often tied on these two variables – i.e. they have the same value for either one or both. For example, the last two rows in Table 3.4 are tied on CDS-#Freq, and the last three rows are tied on CDS-#MWUs.

Table 2.4: Example data points. Statistics are estimated from the CDS corpus.

word	CDS-#Freq	CDS-#MWUs	AoFP
mummy	11,265	5,298	0.804
said	4,894	2,357	1.111
body	180	69	1.209
learn	162	69	2.405
covered	162	69	1.951

Kendall's τ-b addresses both issues. Unlike Pearson's r, which requires normality and is sensitive to outliers, τ-b makes no assumptions about the distribution of variables. And unlike Spearman's ρ, τ-b explicitly addresses tied data points (Agresti, 2010). Intuitively, given two different orderings of a set of data points, τ-b is a function of the number of data pairs which appear in the same order within both orderings, minus the number of pairs that appear in different orders. τ-b thus compares rankings of data points rather than real values. The approach taken, moreover, is maximally

general, ensuring resistance to discrepancies in the data. This generality comes with a decrease in statistical power; but this is compensated for by the amount of data, as we work with close to 7,500 target words.

Moving towards a more concrete description of τ-b, let X, Y be the rankings of target words according to two different variables (e.g. CDS-#MWUs and AoFP). A pair of target words t_i, t_j are then assigned ranks $x_i, x_j \in X$ and ranks $y_i, y_j \in Y$. The two pairs of ranks are *concordant* if they appear in the same order, i.e. if either $x_i < x_j \wedge y_i < y_j$ or $x_i > x_j \wedge y_i > y_j$. If the two pairs are ordered differently, they are *discordant*. Given the number of concordant pairs P and the number of discordant pairs Q, the correlation coefficient of X and Y is calculated as follows:

$$\tau\text{-}b_{XY} = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}$$

where X_0 is the number of pairs tied only in X and Y_0 is the number of pairs tied only in Y (pairs tied in both rankings are not considered).

In addition to such pairwise correlations, we calculate partial correlations – for example, we may want to correlate CDS-#MWUs and AoFP, controlling for CDS-#Freq. A partial correlation would then remove the variance shared between CDS-#Freq, CDS-#MWUs, and AoFP. Controlling for the ranking by a third variable (F), partial τ-b of the rankings X and Y is given by:

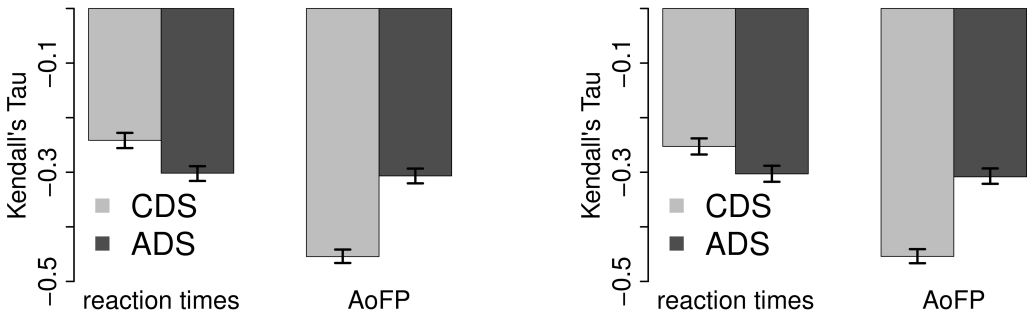
$$\tau\text{-}b_{XY,F} = \frac{\tau\text{-}b_{XY} - \tau\text{-}b_{FX} \times \tau\text{-}b_{FY}}{\sqrt{(1 - \tau\text{-}b_{FX}^2) \times (1 - \tau\text{-}b_{FY}^2)}}$$

95 percent confidence intervals for correlation coefficients are calculated via statistical bootstrapping (Davison and Hinkley, 1997), with each bootstrap based on 1,000 random samples with replacement, and a sample size equal to the number of data points. When comparing two correlation coefficients, we bootstrap 95 % confidence intervals for the difference between coefficients (again based on 1000 random samples with replacement). If zero is not contained within this interval, we can claim with 95 % certainty that the two correlation coefficients differ from one another.

2.3.2 Results and discussion

Correlations between response variables and predictors are summarized in Figures 2.2 and 2.3 (see appendix B for exact values). Figure 2.2 shows full pairwise correlations for #Freq (Figure 2.2a) and #MWUs (Figure 2.2b). #Freq and #MWUs are negatively correlated with RTs and AoFP: The more frequent a target word is and the more MWUs contain it, the earlier the target is produced by children, and the faster it is identified in a lexical decision task by adults. Furthermore, it does not matter whether we use #Freq or #MWUs: The overall picture is very similar, with AoFP being more strongly negatively correlated with CDS-derived predictors, while RTs are more strongly negatively correlated with ADS-derived predictors (95 % confidence interval for the absolute difference between the full correlations of RTs with ADS-#Freq and CDS-#Freq: 0.05 – 0.07; RTs with ADS-#MWUs and CDS-#MWUs: 0.04 – 0.06; AoFP with CDS-#Freq and ADS-#Freq: 0.14 – 0.16; and AoFP with CDS-#MWUs and ADS-#MWUs: 0.13 – 0.16).

Figure 2.2: Full pairwise correlations with ADS and CDS predictor variants.



(a) Correlations with #Freq

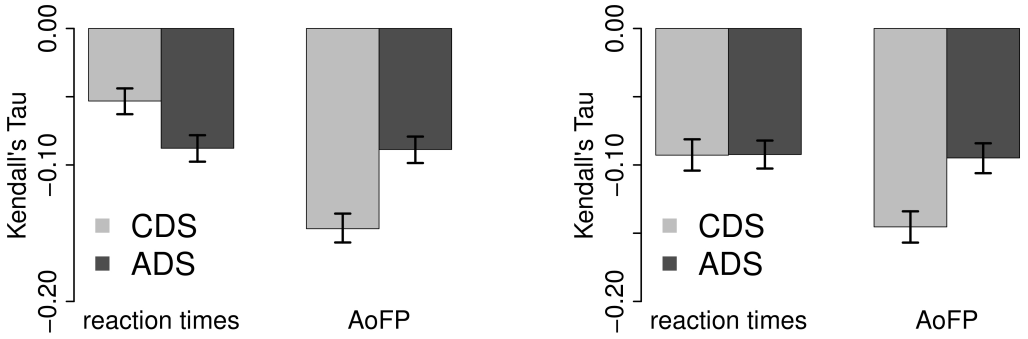
(b) Correlations with #MWUs

Figure 2.3 shows the corresponding partial correlations. Controlling for #MWUs (Figure 2.3a), RTs are still more strongly negatively correlated with ADS-#Freq, and AoFP is still more strongly negatively correlated with CDS-#Freq (95 % CI for the absolute difference between the partial correlations with RTs: 0.02 – 0.05; and with AoFP: 0.05 – 0.07). Controlling for #Freq, (Figure 2.3b), CDS-#MWUs is still more strongly negatively correlated with AoFP (95 % CI for the absolute difference:

0.04 – 0.06), while there is no significant difference between the correlations of RTs with CDS-#MWUs and with ADS-#MWUs (95 % CI: 0.00 – 0.01).

We thus have reason to suspect that the frequency of words in ADS affects RTs more strongly than the frequency of words in CDS. Similarly, the frequency of words in CDS appears to have a stronger effect on AoFP than frequency in ADS. The results furthermore suggest that MWUs in CDS affect AoFP more strongly. We cannot, however, detect a difference between the independent effects of MWUs in ADS and MWUs in CDS on RTs. The general trend is, nevertheless, quite clear: The ADS predictor variants are more strongly correlated with RTs, while the CDS variants are more strongly correlated with AoFP.

Figure 2.3: Partial pairwise correlations with ADS and CDS predictor variants.



(a) Correlations with #Freq

(b) Correlations with #MWUs

In summary, both predictors are negatively correlated with RTs and AoFP – suggesting that frequency and MWUs facilitate both word learning in children and word recognition in adults. Furthermore, the ADS variants are generally more strongly correlated with RTs, and the CDS variants are generally more strongly correlated with AoFP – validating the assumption that ADS is the relevant linguistic input for adults, while CDS is the relevant source of input for children. For the remaining comparison, we choose to correlate ADS-#Freq as well as ADS-#MWUs with RTs, and CDS-#Freq as well as CDS-#MWUs with AoFP. That is, we consider (1) the effect of both predictors from ADS on RTs and (2) the effect of both predictors from CDS on AoFP.

2.4 Analysis III: Comparison with a random baseline

We have established that #MWUs has a facilitatory effect on AoFP and RTs, and that this effect cannot be reduced to the frequency of target words. However, words which appear in a large number of MWUs could be likely to also appear in a large number of randomly selected word sequences. As a consequence, the effect of #MWUs on the two response variables could be due to collinearity with the number of random sequences within which each target word occurs. We use a random baseline to control for this possibility.

2.4.1 Method

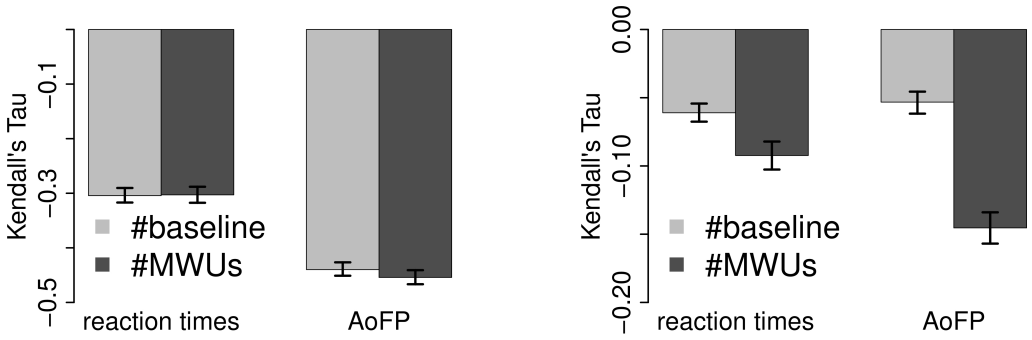
Recall that the baseline model is a mirror version of the CBL, except insofar as it uses a random decision to group successive words into sequences, instead of the backward transitional probabilities used by the CBL. As a result, the MWUs discovered by the CBL are coherent sequences of words, whereas the sequences extracted by the baseline lack this internal coherence. In analogy to the #MWUs measure, we count the number of baseline-extracted word sequences within which each target word appears, and we denote this measure *#baseline*.

The target words and statistical analysis remain unchanged from the previous analysis. And as in analysis II, we compare correlation coefficients – namely, we compare the correlations of AoFP and RTs with #MWUs to the corresponding correlations with #baseline. If there is a unique facilitatory effect of #MWUs, the correlations with #MWUs should be stronger than the correlations with #baseline.

2.4.2 Results and discussion

As explained in the foregoing analysis, we correlate ADS-#MWUs with RTs and CDS-#MWUs with AoFP. For the current analysis, this means that we compare (1) the correlation of ADS-#MWUs with RTs to the correlation of ADS-#baseline with RTs and (2) the correlation of CDS-#MWUs with AoFP to the correlation of CDS-#baseline with AoFP. These comparisons are summarized in Figure 2.4.

Figure 2.4: Comparison of correlations with MWUs from the Chunk-Based Learner and word sequences from a model which randomly groups words into MWUs.



(a) Full correlations

(b) Partial correlations

We cannot detect a statistically significant difference between the full correlations of #MWUs and #baseline with the two response variables (Table 2.4a) (95 % confidence interval for the absolute difference between the full correlations with RTs: 0.00 – 0.00; and with AoFP: 0.00 – 0.01). However, a significant difference emerges once we control for #Freq (Table 2.4b): while both #MWUs and #baseline are negatively correlated with RTs and AoFP, the partial correlations with #MWUs are stronger (95 % confidence interval for the absolute difference between the partial correlations with RTs: 0.02 – 0.04; and with AoFP: 0.08 – 0.10). Given that a difference emerges once frequency is controlled for, the absence of a difference between the full correlation coefficients is likely an artifact of frequency. In other words: compared to #baseline, #MWUs is in fact the stronger predictor. MWUs defined on BTPs between successive pairs of words are, therefore, likely to uniquely facilitate both child word learning and adult word recognition – above and beyond what could be explained by either frequency of exposure or a random baseline.

2.5 Analysis IV: Comparing the effect on word learning to the effect on word recognition

The aim of this last analysis is to compare the effect of MWUs across word learning in children and word recognition in adults. That is, we attempt to ascertain which of the two areas is more strongly affected by MWUs. We have no reason to expect a stronger effect on either area. Given that children's first lexical representation are likely fossilized MWUs, it is possible that MWUs have a particularly strong effect on word learning and a comparatively weaker effect on adult lexical processing; but it is also possible that MWU representations become more entrenched over the course of development, resulting in an even stronger effect on adult word recognition.

2.5.1 Method

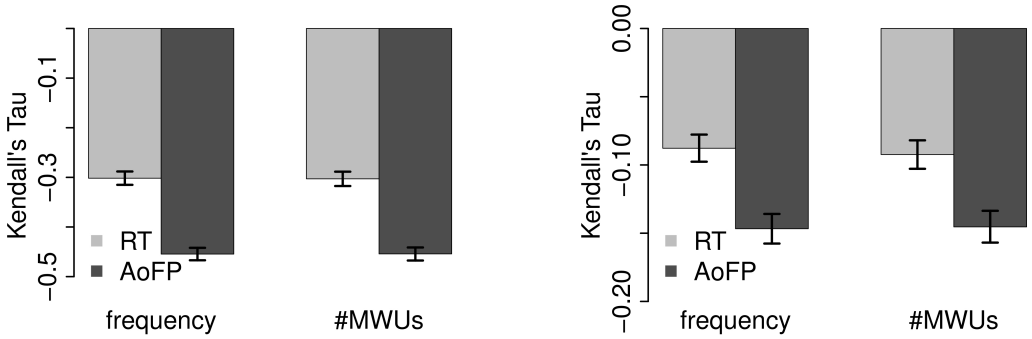
The target words and statistical method remain unchanged from the previous two analyses. Here, we compare the correlation with each predictor across the dependent variables. That is, we ask which of the two predictors has a stronger potential impact on AoFP, and which has a stronger potential impact on RTs.

2.5.2 Results and discussion

Recall that we use the ADS predictor variants for correlations with RTs and the CDS variants for correlations with AoFP, i.e. we correlate ADS-#Freq and ADS-#MWUs with RTs, and CDS-#Freq and CDS-#MWUs with AoFP. This means that we compare (1) the correlation of ADS-#Freq with RTs to the correlation of CDS-#Freq with AoFP, and (2) the correlation of ADS-#MWUs with RTs to the correlation of CDS-#MWUs with AoFP. Figure 2.5 visualizes these comparisons.

The full correlations (Figure 2.5a) of CDS-#Freq and CDS-#MWUs with AoFP are more strongly negative than the correlations of the corresponding ADS predictor variants with RTs (95 % CI for the absolute difference between ADS-#Freq vs. RTs and CDS-#Freq vs. AoFP: 0.13 – 0.17; ADS-#MWUs vs. RTs and CDS-#MWUs

Figure 2.5: Comparison of correlations with predictors across dependent variable.



(a) Full correlations

(b) Partial correlations

vs. AoFP: 0.13 – 0.17). This state of affairs remains unchanged when we control for the other predictor (Figure 2.5b) (95 % CI for the absolute difference between ADS-#Freq vs. RTs and CDS-#Freq vs. AoFP: 0.04 – 0.07; ADS-#MWUs vs. RTs and CDS-#MWUs vs. AoFP: 0.04 – 0.07). Thus, even when factoring out MWUs, the effect of CDS-#Freq on AoFP is stronger than the effect of ADS-#Freq on RTs; and even when factoring out frequency, the effects of CDS-#MWUs on AoFP is stronger than the effect of ADS-#MWUs on RTs. This pattern suggests that frequency and MWUs have a stronger effect on child word learning and a relatively weaker effect on adult word recognition.

2.6 General discussion

2.6.1 The effect of multi-word units

The analyses reported above revealed a negative correlation between the two response variables and the number of MWUs within which words appear. That is, words which appear in relatively many of the MWUs discovered by the CBL tend to be first produced at comparatively early stages in development and tend to be identified relatively quickly by adult subjects in a lexical decision task. Importantly, the correlations surpass the effect of a random baseline and persist even when the

frequency of target words is controlled for.

We also found a negative correlation between the two response variables and the frequency of individual words. This is not surprising: Word frequency has been established as a predictor of word recognition (Balota et al., 2004), with more frequent words being recognized more quickly. In language acquisition, frequency effects are likewise well-attested – including a positive effect of frequency on the age at which children learn words (Ambridge et al., 2015). The effect of #MWUs, on the other hand, constitutes novel evidence for a beneficial impact of MWUs on both word learning and word recognition.

We began this paper by proposing the *MWU acquisition hypothesis*, according to which the formation of MWU representations precedes and then facilitates the formation of single-word representations. On the basis of this hypothesis, we expected that words contained in relatively many of the MWUs discovered by our model will be learned comparatively early in development. This prediction is borne out by the negative correlation of AoFP and the number of MWUs per target word. We also hypothesized that MWU representations facilitate adult word recognition (*MWU processing hypothesis*), leading us to expect that words contained in many model-derived MWUs will be quickly recognized by adults in a lexical decision task. The negative correlation of RTs and the number of MWUs per target word substantiates this prediction.

With the evidence in place, what is lacking is a compelling account of *how* MWUs affect learning and processing. While possible explanations are necessarily going to be exploratory, we nevertheless attempt to synthesize and build on insights from the literature.

2.6.2 Multi-word units in word recognition

While we referred to the literature on word recognition and contextual diversity as the foundation for the *MWU processing hypothesis*, it is not clear exactly how MWUs facilitate word recognition. One possibility is hinted at in the preceding section and has to do with the retrieval of word forms from memory.

Just like word learning, word recognition relies on a memory component: Recognizing a string of letters as corresponding to a particular word form should certainly

involve accessing a lexical representation. In fact, while prominent models of word recognition (Seidenberg and McClelland, 1989; Coltheart et al., 2001) differ in terms of implementation details, they agree on this core process. The beneficial impact of word frequency on RTs from lexical decision tasks can then be interpreted as as facilitatory effect of exposure on retrieval of words from memory. This basic mechanism is implemented in Seidenberg and McClelland (1989)'s connectionist model, which strengthens neuronal connections involved in processing specific words with every exposure; or in Coltheart et al. (2001)'s lexical model, which imposes frequency-based accessibility thresholds on word representations.

Revised or additional mechanisms are needed to accommodate findings pertaining to contextual diversity. Referring to the rational analysis of memory (Anderson and Milson, 1989; Anderson and Schooler, 1991), Adelman et al. (2006) suggest that word accessibility could be governed by likely need, arguing that the more contexts a word appears in, the higher the likelihood that the word will be needed in new contexts. Alternatively, a recent line of work suggests that language processing involves an expectation generation mechanism, which facilitates processing of highly predictable words and reduces memory load for such words (Altmann and Mirković, 2009; Elman, 2009). Based on this idea, Johns et al. (2014) suggest that highly contextually diverse words, being difficult to predict, are more reliant on strong memory representations. Thus, although the specifics are uncertain, these possible explanations have in common that memory takes center stage.

Note that these proposals carry over to an MWU-based view. Contextual diversity is generally quantified as the number of documents or paragraphs within which words appear. We have proposed that contextual diversity causes the internal linking-together of words into MWUs, and that this is what ultimately causes measurable effects such as a correlation with RTs. This explanatory shift is based on a number of studies showing that both children (Bannard and Matthews, 2008) and adults (Arnon and Snider, 2010; Arnon and Priva, 2014; Arnon et al., 2017) are likely to form MWU representations. In other words, if people link words to one another to form MWU representations, then MWUs – not paragraphs or documents – are the relevant cognitive units. Consequently, it should be the interaction between MWU and single word representations which facilitates word recognition, not the interaction between words and documents / paragraphs. The idea that MWU and single word representations interact, in various ways, during the processing of both MWUs

and individual words is supported by previous work: Sprenger et al. (2006) showed that idiomatic phrases both prime and are primed by their constituent words, while Jacobs et al. (2016) found that equally frequent adjective-noun phrases are more easily recognized if they contain an easily recognizable noun.

Thus, if we accept the idea of a facilitatory interaction between word and MWU representations, Adelman et al. (2006)'s and Johns et al. (2014)'s suggested explanations for the beneficial effect of contextual diversity on word recognition carry over to our results with MWUs. Adapting Adelman et al. (2006)'s proposal to an MWU-based framework, it is also possible that the more MWUs a word appears in, the greater the likelihood that it will be needed in new contexts, and this is why such words have particularly accessible memory representation. And adapting Johns et al. (2014)'s suggestion, it could be that people utilize the mental links which form the basis for MWUs in order to predict upcoming words. This prediction process would then be less accurate for words contained in relatively many MWUs – hence necessitating stronger memory representations for such words.

To sum up, we have zeroed in on retrieval of words from memory as a core component of word recognition, and we have reviewed possible ways in which MWUs could interact with the retrieval process. Here, our core argument is that the more MWUs a given word is contained in, the faster it should be retrieved from memory, and the faster it should consequently be recognized.

2.6.3 Multi-word units in word learning

While we proposed the *MWU acquisition hypothesis* based on existing evidence, it is unclear exactly how MWUs exert their facilitatory effect on word learning. Learning to use words is a complex task – subsuming, among other things, the segmentation of phonological forms (Saffran et al., 1996), understanding the intentions of others (Baldwin, 1991; Carpenter et al., 1998), and integrating information across sensorimotor modalities (Lakoff, 1987; Barsalou, 1999). MWUs could potentially interact with several of these processes.

Word segmentation is, perhaps, the most probable candidate process. Consider Peters (1983)'s proposal that early-acquired MWUs, being stored in long-term memory, are gradually segmented into smaller units – units which are themselves stored in

memory, where they are again subject to segmentation. In this fashion, children could bootstrap small-grained linguistic units from an initial inventory of larger chunks. Later work concerned with children's early productions supports this view, showing that in spite of between-child differences in the degree of reliance on initial storage of unanalyzed patterns, all children may in fact rely on this strategy to some extent (Pine and Lieven, 1993). Evidence from perception studies, meanwhile, suggests that infants segment and store both actual and possible words – phonological forms which are word-like but do not correspond to words (Marchetto and Bonatti, 2013; Ngon et al., 2013). In light of the evidence from production, it is plausible that some of these early-segmented units contain several words – i.e., that children sometimes segment multi-word chunks before they begin to segment individual words from within those chunks. Thus, some early fossilized MWUs are likely to be (partially) undersegmented chunks (this could, for example, apply to the MWU *find-it*, used by Tomasello (1992)'s daughter to express desire for an absent object). In this scenario, the more initially undersegmented MWUs contain a given word, the earlier it is going to be segmented. We would then expect this early segmentation to translate into early induction of meaning, as children would have more time to establish the word's meaning, compared to late-segmented words.

The CBL, however, does not start from unanalyzed sequences. Instead, it operates on fully segmented words and builds those up into MWUs. In spite of this, the units it discovers still overlap with the chunks discovered by a simple segmentation algorithm. McCauley et al. (2015) compared the CBL to a segmentation method which initially stores whole utterances (represented as continuous streams of phonemes) as potential words and splits future utterances based on stored exemplars. Comparing the units discovered by the two methods, at identical points in time, shows that some MWUs are discovered by both algorithms. Thus, the MWUs discovered by the CBL correspond, to some extent, to chunks which could result from the initial storage of unanalyzed input.

Note also that CBL-derived MWUs would likely also overlap with chunks discovered via by tracking transitional probabilities between syllables, a segmentation strategy that infants have been shown to employ in artificial words segmentation tasks (Saffran et al., 1996; Aslin et al., 1998). The CBL already tracks transitional probabilities between words, and most English words are monosyllabic. It follows that similar results would be obtained if we ran the CBL on (unsegmented) syllabified

English utterances.

Of course, under-segmentation need not be the only way in which children form MWU representations. It is possible that the combination of smaller units proceeds side-by-side with segmentation, and that the two methods constitute complementary ways of discovering MWUs (McCauley et al., 2015). There may thus be additional processes involved in word learning, beyond segmentation, that interact with MWUs. Consider the process of establishing a word's meaning: Following successful segmentation, to establish the meaning of an object name, children need to create a link between word form and object referent. Estes et al. (2007) have shown that children are capable of executing these two steps in sequence. In the first of two experimental phases, infants listened to a sequence of syllables containing an easily segmented phonological form. In a subsequent object-label-learning task, this phonological form was presented together with a set of novel forms. Infant subjects were able to map the phonological form from the previous phase to an object, but failed to do so with the novel forms. In principle, children are thus able to first segment a meaningless sequence of sound from an incoming speech stream; later, they are able to recognize the sequence within a new context and can map it to a referent.

Crucially, this series of steps involves the recognition of a stored word form before its meaning is established. And it is here that MWUs could again facilitate word learning. To see how, suppose that the subjects in Estes et al. (2007)'s experiment have segmented a phonological form, stored it in memory, and are about to retrieve it in order to map it to a referent. Suppose that this word form is part of one or more MWU representations. Children would then have access to fully-fledged MWU representations without having access to the meaning of each individual word. When encountered, some word forms could thus be more quickly retrieved from memory because they are part of one or more MWUs: If a word form, through MWUs, is linked to many other words, it should be more easily primed for retrieval from memory. And if it is easier to retrieve a word form from memory, we would expect fewer necessary exposures to word forms and their referents to establish a link between the two, compared to word forms which are part of relatively few MWUs.

We have thus identified two possible mechanisms, word segmentation and retrieval of word forms from memory, that are likely to be part of word learning; and we

have spelled out ways in which MWUs could interact with these mechanisms. In both cases, MWUs are expected to have a beneficial impact: The more MWUs a given word is contained in, the easier it should be to segment the word from an unsegmented stream of speech, and the easier it should be to retrieve it from memory prior to establishing its referent.

2.6.4 Word learning vs. word recognition

We have attempted to sketch ways in which MWUs could benefit both word learning and word recognition, providing a possible explanation for (1) the negative correlation between #MWUs and AoFP and (2) the negative correlation between #MWUs and RTs. However, we have not yet considered the results of analysis IV, in which we compared the effects of #MWUs and #Freq across the two dependent variables. The results revealed (1) that CDS-#Freq was more strongly correlated with AoFP than ADS-#Freq with RTs and (2) that CDS-#MWUs was more strongly negatively correlated with AoFP than ADS-#MWUs with RTs. This suggests that frequency of occurrence and involvement in MWUs both affect word learning more strongly than word recognition and indicates a relatively stronger impact of language input during language acquisition, compared to the impact of input on adult processing. Generally speaking, this could be due to greater plasticity during language development and, as a consequence thereof, an increased sensitivity of children to various properties of the linguistic input.

The stronger potential impact of MWUs on word learning could also have to do with the range of cognitive processes which rely on MWUs. We have argued that both word learning and word recognition rely on retrieval of stored word forms from memory: During word learning, segmented word forms need to be retrieved in order to establish their meaning (e.g. to map them to a referent), while the retrieval of lexical representations from memory is so central to word recognition that the two are near-synonymous. In fact, we would argue that word learning subsumes word recognition, alongside many other sub-processes. If true, it is perhaps not unexpected that MWUs should have a stronger impact on word learning: If MWUs benefit the word recognition process, and if word learning subsumes word recognition in addition to other mechanisms which could likewise benefit from MWUs, then

MWUs should have an overall stronger impact on word learning.

2.6.5 Limitations and next steps

Being correlational in nature, the results reported in this study are compatible with three interpretations: (1) direct causation, (2) spurious correlation, and (3) reversed causation. We have argued for possibility (1) – i.e., we have argued that involvement of words in MWUs directly causes them to be produced earlier and recognized more quickly than words which are not involved in (as many) MWUs. While we have attempted to ground our interpretation of the observed correlations in previous results and observations, causality can only be established through future experiments with human subjects.

In the following chapters, we work our way to deriving a testable prediction for such experiments, focusing on the learning mechanism which seems to us the most promising explanation for the observed correlation between #MWUs and AoFP: the wholesale storage of MWUs, during early speech segmentation, as undersegmented chunks. This strategy could be used to further segment stored chunks into smaller units (including words) by comparing them to one another or to novel speech input (Peters, 1983), and the negative correlation between #MWUs and AoFP could be explained if we assume that words contained in many chunks (with a high #MWUs count) are more easily identified as independent units – and thus learned – than words contained in fewer chunks (with a low #MWUs count).

In the next chapter, we follow up on this and address concerns having to do with the possibly spurious nature of our result. Once this is done, we turn towards further analyses that will allow us to formulate a falsifiable prediction for experiments with human subjects.

Chapter 3

Isolating the effect of multi-word units on child word learning

Previous studies have suggested that children possess cognitive representations of multi-word units (MWUs) and that MWUs can facilitate the acquisition of smaller units contained within them. We propose that the formation of MWU representations precedes and facilitates the formation of single-word representations in children. Using different computational methods, we extract MWUs from two large corpora of English child-directed speech. In subsequent regression analyses, we use age of first production of individual words as the dependent and the number of MWUs within which each word appears as an independent variable. We find that early-learned words appear within many MWUs – an effect which is neither reducible to frequency or other common co-variates, nor to the number of context words contained in the MWUs. Our findings support accounts wherein children acquire linguistic patterns of varying sizes, moving gradually from the discovery of MWUs to the acquisition of small-grained linguistic representations.

3.1 Introduction

Frequently co-occurring word combinations have been investigated in studies examining both child (Bannard and Matthews, 2008; Arnon and Clark, 2011; McCauley and Christiansen, 2014) and adult processing (Arnon and Snider, 2010), with mount-

ing evidence that children and adults represent such sequences separately from their constituent words. Indeed, given that many English word sequences have idiosyncratic meanings which cannot be derived from the meaning of their constituent words (e.g. *pay attention to*, *leave of absence*, *you're welcome*), it is reasonable to expect language users to store such semantically opaque sequences in memory. Findings from the literature, however, extend beyond this: In addition to non-compositional constructions, people are likely to also lexicalize frequent but semantically transparent formulaic sequences (Wray, 2008). Here, we use the term *multi-word unit* (MWU) to refer to any sequence of words – compositional or not – which is likely to be lexicalized, and we investigate the role of MWUs in the acquisition of individual words.

More concretely, we expect a facilitatory interaction between the acquisition of MWUs and the acquisition of their constituent words. Provisional evidence for a beneficial impact of MWUs on the acquisition of smaller linguistic units was collected by Arnon and Clark (2011), who showed that children make fewer inflectional errors on known words if the words are contained within frequent MWUs. Usage-based approaches to language acquisition, meanwhile, suggest that children acquire a repertoire of both lexically specific and more abstract multi-word constructions (Tomasello, 2009; Behrens, 2009). Based on this, we propose that children sometimes possess MWU representations before they form representations of the words contained within them, and that these MWU representations then facilitate the acquisition of single-word representations. We dub this the *MWU acquisition hypothesis*.

With the availability of a growing number of corpora of child-caregiver interactions on the one hand (MacWhinney, 2000b) and the development of methods for the extraction of MWUs from corpora on the other hand (McCauley and Christiansen, 2014; Brooke et al., 2014), we are in a position to investigate the kinds of MWUs children are likely to acquire. Concretely, we extract MWUs from two large corpora of transcribed child-directed speech, using (a) a computational model employed by McCauley and Christiansen (2014) to account for findings from the language acquisition literature and (b) an algorithm, developed by Brooke et al. (2014), intended to build a comprehensive lexicon of psychologically plausible MWUs. We view extracted MWUs as an approximation of the types of MWUs children might discover and use the number of MWUs within which a given word is contained as an inde-

pendent variable.

Throughout, we use the age at which children first produce words (age of first production / AoFP) as an index of word learning: If a word is first produced relatively early in development, we assume that this is in part because it is easy to learn when and how to use it. Given the *MWU acquisition hypothesis*, we expect a facilitatory effect of the number of MWUs in which a word appears on its AoFP. This effect, moreover, should be uniquely attributable to MWUs – and not to individual word frequency, semantic co-variates, or the number of context words contained in MWUs. Number of co-occurring context words has previously been shown to predict age of acquisition for words (Hills et al., 2010). But if our proposal is correct, such an effect should disappear once MWUs are taken into consideration.

3.2 Related work

3.2.1 Language acquisition

MWUs have emerged as an important theoretical concept in usage-based approaches to Language Acquisition (Tomasello, 2009). Within this broad theoretical framework, learners' linguistic representations are conceived of as continually complexifying entities, with the developed cognitive system containing both lexically specific and more abstract patterns. At early stages in development, most representations are lexically specific, and child language is "(partially) formulaic and item-based" (Behrens, 2009, p. 393). That is, child language development is thought to involve representations which are lexically specific and span multiple words.

Experimental evidence for the existence of children's MWU representations comes from Bannard and Matthews (2008), who presented 2 and 3 year-olds with frequent MWUs like *a drink of tea* and matched infrequent MWUs like *a drink of milk* that differed in the last word. 2 and 3 year-olds were faster to repeat frequent MWUs, and 3 year-olds were also faster to repeat the first three words if they formed a frequent MWU with the fourth word. Since the final word and the final bigram (e.g. *of tea* and *of milk*) were matched for frequency, the processing advantage for frequent MWUs can only be attributed to the frequency of the entire MWU, rather

than to the frequencies of its component words, suggesting that children have access to cognitive representations of MWUs. Bannard and Matthews (2008) argue, furthermore, that their subjects were likely familiar with the words comprising the MWUs, which implies the existence of (partially) independent MWU and single-word representations.[1]

In addition, Arnon and Clark (2011) found that MWUs interact with the acquisition of morphemes: 4;6 year-olds produced more correct irregular plurals after familiar lexically specific frames than after general questions. Subjects were presented with depictions of several objects. The object name was elicited either with a labeling question or with a lexically specific frame. For example, on one particular trial the objects were sheep, the lexically specific frame was *Count some –*, and the labeling question was *What are all these called?* 4;6 year-olds were more likely to complete the lexically specific frame with *sheep* and would provide relatively more incorrect plural forms – like the over-regularized *sheeps* – in response to the labeling question. This suggests that MWUs like *count some sheep* affect the way in which some of the smaller units contained within them are learned.

3.2.2 Computational modeling

The above-cited results by Arnon and Clark (2011) and Bannard and Matthews (2008) have been modeled by McCauley and Christiansen (2014). In a comprehension phase, their model segments a corpus of child-directed speech into MWUs. In a production phase, it generates child-produced utterances based on stored MWUs. Given a corpus, MWUs are extracted by comparing the conditional probability of the current word given the preceding word to a running average of all such probabilities, for all words so far encountered one position to the left of the current word. If this backward transitional probability (BTP) is larger than the running average, the current and preceding word are part of an MWU. The process continues until the BTP falls below the average, at which point the current MWU is stored in memory.

Extracted MWUs can then be used to re-construct child-produced utterances. McCauley and Christiansen (2011) compared model-derived to child-produced utter-

[1] The same argument can be made for adults, who are faster to recognize and produce frequent four-word MWUs in similar experiments (Arnon and Snider, 2010).

ances across 13 corpora from the CHILDES database (MacWhinney, 2000b). On average, about 60 % of utterances were successfully re-produced – illustrating that a purely MWU-based system can account for a majority of child-produced utterances. Importantly, MWUs discovered by the model can also be used to model results from Bannard and Matthews (2008) and Arnon and Clark (2011). In both cases, stimuli were sequences of words – constructions like *a drink of tea* in the former and *count some sheep* in the latter study. McCauley and Christiansen (2014) assigned a *chunkedness* score to each stimulus by calculating the product of BTPs between the MWUs used by the model to re-produce each stimulus. In each study, differences in scores reflected differences in subjects' performance: Stimuli with lower reaction times in Bannard and Matthews (2008)'s study were assigned a larger chunkedness score, as were stimuli which elicited a larger proportion of correctly inflected nouns in Arnon and Clark (2011)'s study.

3.2.3 Natural language processing

McCauley and Christiansen's (2011, 2014) model can be situated in a tradition that measures association strength between pairs of words. Words are then grouped together if their association strength exceeds a particular threshold. McCauley and Christiansen (2014, 2011) use BTP as the measure of association. Other options include pointwise mutual information or log likelihood (cf. Pecina, 2010, for an overview). All association-based methods require an arbitrary threshold for inclusion of words in MWUs. In addition, there is no consensus on which association measure is best. An alternative approach is to identify frequent n-grams – called *lexical bundles* –, but this requires very high frequency thresholds (Biber et al., 2004). There is, then, no generally accepted way of extracting MWUs from corpora, nor is it common practice to evaluate whether extracted MWUs correspond to psychologically real entities.

Work by Brooke et al. (2014) has recently begun to address these issues. Their method operates at the token level, identifies MWUs of varying sizes, and relies on two parameters: a frequency threshold and a maximum MWU size. Broadly speaking, the algorithm considers all possible segmentations of a given sentence into n-grams that meet a pre-specified frequency threshold. Then, that segmentation is selected which maximizes the predictability of each word within its n-gram. The

stated goal of this work is to develop a method for the extraction of an MWU lexicon that would correspond to knowledge of MWUs possessed by native speakers. The system has since been refined by Brooke et al. (2015), who also introduced first steps towards evaluating MWU lexicons.

3.3 Hypothesis

According to the *MWU acquisition hypothesis*, children sometimes acquire MWU representations before they acquire representations of the individual words contained in MWUs, and access to MWU representations then facilitates acquisition of the words contained in them.[2] While this hypothesis is grounded in the literature, it is not clear via which mechanisms MWUs might aid the word learning process. Consequently, our goal is to provide evidence *that* MWUs uniquely facilitate word learning, and not *how* this process unfolds. Below, we nevertheless sketch two possible scenarios.

One possibility is that children initially acquire MWUs as unanalyzed units. This could result from an initial undersegmentation of the input: Words, before their meaning is established, need to be identified from a continuous stream of sound. Early in development, children might sometimes segment multi-word chunks before they begin to segment individual words from within those chunks. Thus, some early fossilized MWUs are likely to be (partially) undersegmented chunks. In this scenario, the more initially undersegmented MWUs contain it, the earlier a given word is going to be segmented. We would then expect this early segmentation to translate into early induction of meaning.

A second possibility is that children discover some words before establishing their meaning. They would then go on to discover MWUs containing those words, at which point they have access to fully-fledged MWU representations without having access to the meaning of each individual word. The more MWUs contain a given word, the more words it is going to be linked to – and the more words will prime its retrieval, making it more salient for the learner. On average, a word with many links

[2]Note that we do not claim that the acquisition of MWUs *always* precedes the acquisition of single words, but merely that this happens often enough to have a measurable impact on word learning.

will be more easily retrieved than a word with few links. Because of this, we would expect fewer necessary exposures to establish the meaning of a word which forms part of relatively many MWUs, compared to words contained in fewer MWUs.

As mentioned, we do not distinguish between these two and other such possibilities. Instead, we aim to broadly corroborate the *MWU acquisition hypothesis* by showing that MWUs uniquely facilitate word learning: If, all else being equal, words contained in many MWUs are learned earlier than other words, this would be indicative of a developmental pattern which begins with the formation of MWU representations and then proceeds to the acquisition of individual words.

3.4 Method

Our method is the following: First, we extract MWUs from two corpora of English child-directed speech (CDS) and estimate age of first production (AoFP) for the words produced by the children addressed in the CDS corpora. We then use the number of MWUs within which each target word appears (*#MWUs*) as an independent variable – next to several co-variates – in a linear regression analysis, with AoFP as the dependent variable. If the *MWU acquisition hypothesis* is true, we expect a unique facilitatory effect of *#MWUs* on AoFP.

3.4.1 Child-directed speech

We use two corpora of CDS, which both consist of the adult-produced utterances from several corpora on the CHILDES database (MacWhinney, 2000b). Some corpora are based on cross-sectional studies, while others are longitudinal. In addition, subjects vary in age. Regardless, each corpus consists of standardized transcripts, based on recordings of child-caregiver interactions. In order to maximize the amount of data, we ignore possible fine-grained differences between age cohorts and compile a North-American corpus (NA-CDS) from 45 American English corpora[3] and a

[3]Corpora names (see http://childes.talkbank.org/access/ for references): Bates, Bernstein, Bliss, Bloom70, Bloom73, Bohannon, Braunwald, Brent, Brown, Carterette, Clark, Cornell, Demetras1, Demetras2, ErvinTripp, Evans, Feldman, Garvey, Gathercole, Gleason, HSLLD, Hall, Higginson, Kuczaj,MacWhinney, McCune, McMillan, Morisset, Nelson, NewEngland, Peters,

British English corpus (BE-CDS) from eight British corpora[4]. Table 3.1 summarizes statistics.

Table 3.1: Relevant corpus statistics.

measure	CDS-BE	CDS-NA
nr. tokens	4,681,925	6,389,963
nr. types	24,929	37,128
median length of utt.	4 (IQR: 4)	4 (IQR: 4)
nr. adult speakers	201	774
nr. children addressed	134	441
mean child age (months)	33 (SD: 9)	41 (SD: 23)

3.4.2 Extraction of multi-word units

To extract MWUs from the CDS corpora, we use McCauley and Christiansen's (2014) model as well as Brooke et al.'s (2014) method. McCauley and Christiansen's (2014) model – called *Chunk-Based Learner* (CBL) – processes a given corpus utterance by utterance and word by word. Processing an utterance u is initiated by incrementing the frequency count of the first word $w_1 \in u$ by 1 and creating a new MWU with w_1 as its only member. For each subsequent word w_i at utterance position $1 < i \leq length(u)$, the model keeps track of the number of times w_i has been encountered so far, as well as how often the immediately preceding word w_{i-1} has occurred one position to the left of w. The model then calculates the backward transitional probability (BTP) of w_i and w_{i-1}: $p(w_{i-1}|w_i)$. If this probability is larger than the average BTP across all words which have occurred one position to the left of w in all utterances so far considered, w_i is added to the current MWU. Else, the current MWU is added to a set M, and a new MWU is created – again with w_i as its only member. In this way, the model discovers MWUs of size 2 or larger, as well as single-word units, collected in M. In our analyses, we use all MWUs which occur at least twice in the input corpus.

Post,Providence, Rollins, Sachs, Snow, Soderstrom, Sprott, Suppes, Tardif, Valian, VanHouten, VanKleeck, Warren, Weist

[4]Belfast, Fletcher, Manchester, Thomas, Tommerdahl, Wells, Forrester, Lara

As a second model, we use the method from Brooke et al. (2014)[5]. We refer to it as *Prediction Based Segmenter* (PBS), as it splits utterances into n-grams whose component words are maximally predictable. The basic idea is that given an n-gram $w_1...w_n$, the conditional probability of any word w_i given the remaining subsequence $w_1...w_{i-1}, w_{i+1}...w_n$ should be maximal. In essence, the algorithm splits utterances into n-grams such that each word's predictability is maximized, capturing the intuition that words within MWUs are more predictive of one another than words outside of MWUs – but see Brooke et al. (2014) for a more in-depth explanation. Specifying a maximum n-gram length of ten – longer than most utterances in the corpus – , we use the PBS to segment utterances into either single-word units or MWUs with a minimum size of two and a maximum size of ten. As with the CBL, we retain all MWUs which occur at least twice.

Running the models on the two CDS corpora results in four different sets of MWUs, whose distributions are summarized in Table 3.2. The CBL results in a larger number of shorter MWUs, while the PBS identifies MWUs that are a bit longer. There are generally more MWU types than word types (compare Table 3.1).

Table 3.2: Relevant statistics about the distribution of MWUs.

corpus	*measure*	CBL	PBS
CDS-BE	MWU tokens	1,073,037	978,804
	MWU types	465,447	387,391
	median length	4 (IQR: 3)	5 (IQR: 4)
CDS-NA	MWU tokens	1,40,8614	1,338,173
	MWU types	628,252	492,863
	median length	4 (IQR: 3)	5 (IQR: 4)

3.4.3 Age of first production

To induce AoFP, we start from a corpus of child-produced utterances, treating a word as having been learned at the earliest developmental stage at which any child within the corpus can produce it. *Developmental stage* is defined in terms of mean

[5]available online: `http://www.cs.toronto.edu/~jbrooke`

length of utterance (MLU) – the average child utterance length, in tokens, within a transcript. Since transcripts have varying lengths, we estimate MLU for each transcript via statistical bootstrapping, wherein the sampling distribution of the population is approximated by drawing random samples from the data (Davison and Hinkley, 1997). Each bootstrap is based on 1,000 random samples with replacement, with the sample size equal to the number of child utterances per transcript. We thus induce MLU rather than AoFP estimates, since MLU is a more robust estimator of development (Parker and Brorson, 2005): Children who are close in age may nevertheless be far apart in terms of language development. For simplicity, we still refer to a word's MLU value as its AoFP. To induce a value for any word, we calculate the set of MLUs γ for all transcripts within which the word appears and assign it the smallest value in γ.

We perform this procedure for each word produced by the children addressed in the two CDS corpora – once for the NA data and once for the BE data, meaning that we end up with two AoFP data sets: 441 children are addressed in the CDS-NA corpus and together produce 29,188 different words, each of which is assigned an AoFP value; and 134 children are addressed in the CDS-BE corpus, producing 14,747 different words, again each with its own AoFP value.

3.4.4 Regression analyses

In the regression models, we use AoFP as the dependent variable. The first key independent variable is the number of different MWUs within which a given target word appears (**#MWUs**). For example, assuming our corpus is CDS-NA and our target words are *girl* and *sit*, we count the unique MWUs which contain these two words. To illustrate this, Table 3.3 shows the five most frequent MWUs, in CDS-NA, containing the two words. Counting all such MWUs, we end up with 113 (PBS) and 230 MWUs (CBL) for *girl*, and 253 (PBS) and 488 (CBL) MWUs for *sit*. The second key independent variable is the number of unique context words appearing in all MWUs within which a given target word is contained (**#ctxt**). If MWUs aid word learning, we should see a facilitatory effect of *#MWUs* on AoFP, and this effect should not be reducible to *#ctxt*. If a target word appears within a large number of MWUs, it will also tend to co-occur with a large number of context words. We posit, however, that MWUs – not individual words – are the cognitively relevant

units; and we predict, therefore, that it is the number of MWUs – not the number of co-occurring context words – which affects learning.

Table 3.3: The five most frequent MWUs, found in CDS-NA, for the target words *girl* and *sit*.

word	CBL	PBS
girl	good girl (410)	good girl (440)
	little girl (110)	little girl (175)
	that's a girl (101)	a girl (98)
	a girl (68)	that's a good girl (59)
	that's a good girl (57)	the little girl (51)
sit	sit down (627)	sit down (846)
	sit up (88)	sit up (141)
	sit here (46)	you sit (117)
	sit over here (46)	you wanna sit (87)
	sit down please (41)	come sit (85)

Frequency counts for the MWUs are give in parentheses.

Further, we include the following co-variates: the corpus-frequency of each target word (**freq**), number of syllables (**syl**), phonological neighborhood density (**phon**), and concreteness ratings (**con**). Given a target word, *phon* is defined as the number of homophones, plus the number of words that can be derived from the target by either adding, deleting, or substituting a single phoneme. *phon*, together with *nsyl*, is derived from a syllabified version of the Carnegie Mellon University (CMU) pronouncing dictionary (Bartlett et al., 2009).[6] Concreteness ratings for 40,000 lemmas are taken from Brysbaert et al. (2014)[7], who collected them from over four thousand participants via Mechanical Turk. Since ratings were collected for lemmas, whereas we work with word forms, we assigned the lemma rating to all word forms which correspond to the lemma. Regression analyses are based on all words for which *phon*, *syl* and *con* estimates are available: 7,265 words in CDS-BE and 5,724 words in CDS-NA. Table 3.4 shows three example data points.

[6] `http://webdocs.cs.ualberta.ca/~kondrak/cmudict.html`

[7] `http://crr.ugent.be/archives/1330`

To increase the generality of this study's implications, we use AoFP from children who were not addressed in the corpus used to estimate *#MWUs*, *#ctxt*, and frequency. In other words, we use AoFP from the children addressed in the CDS-NA corpus for regression models which include *#MWUs*, *#ctxt* and frequency counts from CDS-BE; and we use AoFP from the children addressed in CDS-BE for regression models which include independent variables from CDS-NA.

Table 3.4: Example data points from the CDS-BE corpus, with *#MWUs* and *#ctxt* estimated via the PBS.

word	freq	con	phon	syl	#ctxt	#MWUs	AoFP
goes	3,183	2.19	16,661	1	430	156	0.51
lunch	1,175	4.31	1,175	1	168	57	1.29
running	853	4.27	853	2	86	46	1.16

3.5 Results

Table 3.5 presents results of four linear regression analyses (2 methods for MWU extraction × 2 CDS corpora). All variables are log-transformed, and *#ctxt* as well as *#MWUs* are increased by 1, in order to avoid problems from zero counts. The baseline models with all co-variates (second column) explain between 38 and 44 percent of the variance in AoFP. *Freq* and *con* have facilitatory effects, while there are no statistically significant effects for *phon* and *nsyl*. Given that increased frequency of exposure is associated with early word learning (Ambridge et al., 2015), the effect of *freq* is not surprising, while the effect of *con* implies that words associated with concrete concepts tend to be early-acquired.

Adding *#ctxt* to the baseline models (third column) leads to a significant increase in R^2, with a facilitatory effect of *#ctxt*. Adding *#MWUs* to the baseline models (fourth column) also improves the fit, with a facilitatory effect of *#MWUs*. Interestingly, the effect of *#MWUs* is stronger than the effect of *#ctxt*. Neither effect is reducible to the frequency of target words, their concreteness, their phonological complexity, or the density of their phonological neighborhoods. In models which include the covariates plus *#ctxt* and *#MWUs* (fifth and sixth columns), *#MWUs*

Table 3.5: Effects of log-transformed *#ctxt* and log-transformed *#MWUs*.

	Effect (ΔR^2 in %)				
Data + corpus	Covariates baseline	Log-#ctxt	Log-#MWUs	Log-#ctxt unique	Log-#MWUs unique
CBL					
CDS-BE	44.85 * * *	1.23 * * *	1.73 * * *	0.34 (I) * * *	0.85 * * *
CDS-NA	38.33 * * *	0.87 * * *	1.35 * * *	0.13 (I) * * *	0.61 * * *
PBS					
CDS-BE	44.85 * * *	0.78 * * *	1.52 * * *	0.55 (I) * * *	1.29 * * *
CDS-NA	38.33 * * *	0.47 * * *	1.09 * * *	0.18 (I) * * *	0.79 * * *

The effects of *#ctxt* and *#MWUs* were calculated after those of the co-variates had been included. Unique effects are those with the indicated variable entered last (i.e. *#ctxt* after covariates + *#MWUs*, or *#MWUs* after *#ctxt* + covariates). I = inhibitory effect of indicated variable.

continues to exert a facilitatory effect; but importantly, *#ctxt* now has an inhibitory effect on AoFP. This pattern suggests that the initial facilitatory effect of *#ctxt* is due to collinearity with *#MWUs*.

Our results imply that it is involvement in a large number of MWUs – not co-occurrence with a large number of context words – which drives word learning. Furthermore, the effect of MWUs may be limited to MWUs consisting of relatively few words. Hence, when factoring out *#MWUs*, co-occurrence with a large number of context words inhibits acquisition of the target words; and when factoring out the effect of context words, the positive effect of *#MWUs* persists.

3.6 Conclusions and next steps

We began this chapter with a review of studies which suggest that children acquire representations of MWUs and that MWUs could facilitate the acquisition of smaller linguistic units contained within them. Based on this, we proposed the *MWU acquisition hypothesis*, according to which the formation of MWU representa-

tions precedes and facilitates the formation of individual word representations. The facilitatory effect of *#MWUs* on AoFP supports this hypothesis. More broadly, it supports accounts of language development wherein children acquire linguistic units at various levels of granularity, transitioning gradually from MWUs to more small-grained units.

Our results also have implications for a previous finding: Hills et al. (2010) found that the sum of unique context words occurring within a window of five words to the left and right of each target word predicts age of acquisition of the targets. We also observed a facilitatory effect of *#ctxt*. However, an inhibitory effect of *#ctxt* emerged once *#MWUs* was controlled for. Thus, given that their measure is similar to *#ctxt*, it is possible that Hills et al. (2010)'s result is due to collinearity with the number of MWUs within which target words appear.

In formulating the hypothesis, we purposefully remained agnostic with respect to the specific mechanisms involved in the facilitatory interaction between the acquisition of MWU and single word representations. Accordingly, our results support a general class of theories wherein MWUs are acquired before single words. These could be usage based approaches to language acquisition (Tomasello, 2009), but also proposals such as Peters (1983)'s, according to which early-acquired MWUs are undersegmented chunks which are gradually segmented into smaller units – units which are themselves stored in memory, where they are again subject to segmentation.

In the following chapter, we take this idea as a starting point, operationalizing the notion of *chunk* at the syllable rather than the word level. We then extract different types of multi-syllable chunks from corpora transcribed child-directed speech, in order to examine which types of chunks have the strongest potential effect on word learning. This, in turn, allows us to more closely specify the mechanisms whereby chunks facilitate word learning.

Chapter 4

Children probably store short rather than frequent or predictable chunks: Quantitative evidence from a corpus study

One of the tasks faced by young children is the segmentation of a continuous stream of speech into discrete linguistic units. Early in development, syllables emerge as perceptual primitives, and the wholesale storage of syllable chunks constitutes a possible strategy for bootstrapping the segmentation process. Here, we investigate what types of chunks children store. Our method involves selecting syllabified utterances from corpora of child-directed speech, which we vary according to (a) their length in syllables, (b) the mutual predictability of their syllables, and (c) their frequency. We then use the number of utterances within which words are contained to predict the time course of word learning, arguing that utterances which perform well at this task are also more likely to be stored, by young children, as undersegmented chunks. Our results show that short utterances are best-suited for predicting when children acquire the words contained within them. In addition, we also find that short utterances are the most likely to correspond to words. Together, these results suggest that children extract and store short, word-like chunks. We discuss implications for models of speech segmentation and for work on formulaic multi-word sequences.

4.1 Introduction

The present study investigates undersegmented chunks in child language development. Previous work suggests that young children sometimes store speech sequences such as *Oh dear* or *Where's it gone* as internally unanalyzed chunks, without having discovered smaller constituents such as words or phonemes (Lieven et al., 1992; Pine and Lieven, 1993). We argue that this is a result of word segmentation: When children do not yet know what the meaningful units in their language are, they could initially rely on wholesale storage of chunks, which are then further segmented by comparing chunks to one another and to incoming speech (Peters, 1987). In this chapter, we are interested in the nature of such undersegmented chunks.

As one possibility, children could extract and store frequently recurring speech sequences. Frequency effects are pervasive in language development, with children acquiring frequent words, morphemes, and even syntactic constructions before less frequent exemplars (Ambridge et al., 2015). Consequently, it would be sensible to expect preferential storage of particularly frequent chunks. Alternatively, perhaps frequency is less important than input properties that indicate whether a given sequence corresponds to a discrete linguistic unit, such as a word or a morpheme. Next to frequent chunks, we thus consider two other, potentially more word-like chunk types: especially internally predictable and particularly short chunks.

A landmark study by Saffran et al. (1996) first demonstrated that children can exploit conditional probabilities between adjacent syllables in order to extract nonsense words from a continuous stream of speech. This raises the more general possibility that, during the word segmentation process, children use syllabic predictability to extract multi-syllable chunks from speech. Chunks with high syllabic predictability might be more likely to correspond to discrete linguistic units than frequent syllable sequences, and perhaps this inclines children to store interally predictable rather than frequent undersegmented chunks.

Yet another possibility is that stored chunks are neither frequent nor predictable, but simply short. Compared to long speech sequences, short sequences are unlikely to contain smaller linguistic units and might thus appear more word- or morpheme-like to the language-acquiring child. As a consequence, perhaps children are more likely to store short rather than frequent sequences as undersegmented chunks. In

this study, we investigate all three chunk properties – (1) whole-chunk frequency, (2) syllabic predictability, (3) chunk length –, and we ask to what extent children rely on these properties during the extraction of an initial chunk vocabulary.

The remainder of the chapter is structured as follows. First, we survey evidence for the existence of unsegmented chunks in young children, arguing that they emerge as a by-product of word segmentation. Following this, we provide a brief sketch of our method, which involves selecting multi-syllable utterances as sequences that could potentially be stored (by young children) as undersegmented chunks. Varying the syllabic predictability, frequency, and length of selected utterances, we evluate which multi-syllable utterances (henceforth *MSUs*) perform better at predicting the time course of word learning.

Our method extends previous work by Grimm et al. (2017), who found that words contained in a large number of multi-word phrases tend to be learned early in development. Referring to Peters (1987), Grimm et al. (2017) suggest that children store some phrases as undersegmented chunks. Chunks are then compared to one another in order to identify shared sub-units. And the more chunks contain a particular unit (such as a word), the easier it should be to discover that unit. We expand on this by evaluating whether short, frequent, or internally predictable MSUs perform better at predicting when their constituent words are learned – arguing that well-performing MSUs are more likely to be stored within children's early proto-lexica.

4.1.1 Evidence for children's unanalyzed chunk vocabularies

Young children sometimes produce utterances in ways which suggest that they are treated as (partially) unanalyzed wholes. Peters (1987) surveyed various examples, including e.g. the child utterance *I don't know where's Emma one*, which appears to consist of the previously heard utterances *I don't know* and *Where's Emma one*; or *I all very mucky too*, given in response to the statement *We're all very mucky*.[1] Observations like these led Peters (1987) to suggest that children could extract and store in memory uninterrupted sequences of speech. Then, once a larger inventory of chunks has been collected, children might compare stored chunks to one another and to incoming speech, bootstrapping a vocabulary of smaller units.

[1]Both examples were originally reported by Clark (1974).

This proposal receives support from a systematic investigation conducted by Lieven et al. (1992), who analyzed the productive vocabularies of twelve English-speaking children through parental reports and analyses of child-caregiver interactions. Child-produced multi-word utterances were coded as *frozen phrases* if they contained at least two words which had not previously occurred in isolation within the vocabulary of the child – or if they contained only one such word, so long as the word had not occurred in the same position within a previous utterance. Lieven et al. (1992) found that their subjects' productive vocabularies, at 50 and 100 produced units (phrases or words), contained around 20 % frozen phrases. This reliance on frozen chunks, although practiced to different degrees by different subjects, seems to be a strategy shared by all children (Pine and Lieven, 1993).

Once an initial vocabulary of chunks has been stored, Peters (1987) proposed, children could gradually discover structural patterns by identifying chunk-internal positions of variability. Lieven et al. (2009a) implemented a similar idea in a computational method that reconstructs child utterances on the basis of earlier productions. The method first attempts to match a given utterance with earlier child productions and, if this is not possible, inserts abstract slots. For example, upon observing the utterances ***I go*** *bathroom* and ***I go*** *home*, it could create an ***I go*** *LOCATION* construction. Lieven et al. (2009a) report that between 20 and 40 % of their two-year-old subjects' utterances could be exactly matched to previous productions, while the majority of non-exact matches required the insertion of just a single slot. These results are echoed by Bannard et al. (2009) and Borensztajn et al. (2009), who also worked with child-produced speech and applied methods for grammar induction that can discover both lexicalized and abstract constituents.

The early building blocks of child language, then, appear to include unanalyzed chunks. Such findings can be situated within a usage-based approach to language acquisition (Behrens, 2009; Tomasello, 2009), a framework which conceives of early linguistic representations as lexically specific units that often span multiple words. Representations are refined and become more abstract over time, and the developed cognitive system operates with both lexically specific and more abstract patterns.[2] Unanalyzed chunks, that is, should only exist for a short developmental window, when children are faced with the task of segmenting continuous speech into discrete

[2]This idea re-surfaces in accounts of adult linguistic competence which include constructions as constituents (Goldberg, 2006; O'Donnell, 2015).

units. But once that process is complete, smaller linguistic units should replace the initial chunk vocabulary. We next review converging evidence from empirical studies and computational models of word segmentation in support of this notion.

4.1.2 Undersegmented chunks during word segmentation

One of the first challenges faced by children during language development is what Peters (1987) called the *initial extraction problem*: Without knowledge of the units in their target language(s), which speech sequences should children pick out as hypothesized linguistic units? Early perception studies showed that 2-month-olds demonstrate improved discrimination of syllable-like sequences (Bertoncini and Mehler, 1981) and are proficient at storing information pertaining to the syllabic – but not the phonemic – structure of speech (Jusczyk and Derrah, 1987). Follow-up work suggests that even 4-day-old neonates perceive speech in terms of syllables (Bijeljac-Babic et al., 1993). And on the computational modeling side, it is possible to segment speech into units that closely resemble syllables by tracking changes in sonority (Räsänen et al., 2015, 2018) – i.e., by attending only to changes in audibility, without reliance on prior linguistic knowledge. The syllable thus presents a good candidate for an early perceptual primitive in speech.

As one possible segmentation strategy, children could focus on sequences characterized by high transitional probabilities (TPs) between syllables.[3] In a seminal study, Saffran et al. (1996) exposed eight-month-olds to synthesized streams of nonsense words, with no cues to word boundaries other than the co-occurrence patterns of syllables. Within-word TPs of four different three-syllable nonsense words (e.g. *padoti* or *golabu*) were 1.0 (e.g. *go* was always followed by *la*), while TPs between syllables spanning word boundaries were 0.33 (e.g. *bu* could be followed by the first syllable of three other words). In the testing phase, subjects listened longer to sequences which spanned word boundaries than to the more internally predictable nonsense words. Infants typically pay more attention to novel stimuli, and less to familiar ones. Saffran et al. (1996)'s results thus imply that subjects were familiar with the internally predictable nonsense words. Infants, that is, appear capable of exploit-

[3]In psycholinguistics, the term *transitional probability* has come to denote conditional probabilities between units. Conditional probability is a measure of association strength between two elements that is normalized by the frequency of the non-conditional element.

ing statistical regularities between syllables to segment words from fluent speech. Aslin et al. (1998) replicated these results while keeping the frequencies of nonsense words constant, demonstrating that TPs provide a useful cue even when they are not correlated with frequency.[4]

There are, of course, other potential segmentation cues, such as stress or co-articulation (Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003). Sensitivity to certain cues seems to be present at an early age, while other cues are only used at later stages. For example, seven-month-olds exhibit sensitivity to TPs but not to stress, while nine-month-olds can exploit stress patterns in an artificial segmentation task (Thiessen and Saffran, 2003). Thiessen and Saffran (2003) hypothesize that this indicates an early exploitation of statistical structure in order to extract a first set of words. These are then used to discover language-specific stress patterns, which can help to further segment the input. Extracted units *could* correspond to actual words, but this need not always be the case. Some units, extracted via reliance on statistical structure, could be stored as undersegmented chunks; and by comparing chunks to one another, children could discover language-specific segmentation cues, bootstrapping further segmentation. This bootstrapping approach to segmentation has the potential to explain other patterns in language development, such as the emergence of phonemic categories before the presence of a large receptive lexicon: If children approach segmentation by constructing a proto-lexicon of chunks, early phonemic contrasts could emerge as a result of identifying minimally different chunks (Martin et al., 2013).

Under such a proposal, undersegmented chunks are a side-effect of the segmentation process, and they would become fully analyzed once that process is complete. Evidence from computational models of word segmentation supports this view. The models described by Goldwater et al. (2009), for example, start from phoneme sequences, which are then segmented on the basis of statistical regularities between phonemes.[5] Discovered units include words, but also many undersegmented chunks. Another segmentation strategy, not mutually exclusive with reliance on statistical

[4]Various studies have since shown that the underlying mechanism can operate on non-linguistic auditory as well as visual stimuli, and that it is not restricted to humans. See Aslin (2017) for a review.

[5]See Phillips and Pearl (2015) for similar models which operate on syllables rather than on phonemes.

structure, is the wholesale storage and gradual breaking-down of full utterances. In this possible scenario – going back at least to Peters (1987) – children initially store full utterances as holistic units, and novel input sequences are only split if another unit (stored in memory) is contained within them, leading to the discovery of more and more fine-grained units. Computational models which implement this strategy (Monaghan and Christiansen, 2010; Lignos and Yang, 2010; Lignos, 2012) achieve excellent performance[6] and thus demonstrate how a large number of undersegmented chunks could accumulate as by-products of the segmentation process. Crucially, chunk-based segmentation algorithms are not just an engineering idea. Two-month-olds show improved memory for speech when it is contained in clause-like units, compared to being presented in list form or spanning clause boundaries (Mandel et al., 1994). At the same time, it has been demonstrated that six-month-olds can use their own name or the word *mommy* to segment unfamiliar words from novel sequences (Bortfeld et al., 2005).

Undersegmented chunks, in summary, are a plausible by-product of the segmentation process. In the current study, we ask which types of chunks children initially extract from speech. In language development, frequent items are generally learned before less frequent items (Ambridge et al., 2015), and one could thus expect children to preferentially extract and store frequent chunks. Support for the role of frequency during segmentation comes from findings that eight-month-olds can detect words within fluent speech on the basis of their frequency (Jusczyk and Aslin, 1995), and eleven-month-olds appear to store highly frequent syllable sequences that span word boundaries as well as highly frequent disyllabic nonsense words (Ngon et al., 2013).

Perhaps, however, frequency is less important than the perceived unity of a given syllable sequence. That is, perhaps children store syllable sequences which appear to form a discrete unit and cannot, for all intents and purposes, be segmented into smaller units, such as words or morphemes. For example, we can reasonably expect that short sequences are more likely to correspond to words or morphemes than longer sequences. Thus, if children store chunks as hypothesized words or morphemes, perhaps they simply store uninterrupted speech sequences that happen to be particularly short.

[6]Cf. Phillips and Pearl (2015), who compared several state-of-the-art Bayseian segmenters to Lignos (2012)'s model. As long as the input is represented in terms of syllables, and not in phonemes, the chunk-based segmenter performs similarly to Bayseian approaches.

Alternatively, given that TPs are an early segmentation cue (Saffran et al., 1996; Aslin et al., 1998; Thiessen and Saffran, 2003), children might extract and store sequences whose syllables are especially mutually predictive. Syllable sequences presumably exist along a spectrum of predictability, with some consisting of syllables that always and only occur with one another, while others have a more variable internal structure. If the goal of segmentation is the discovery of discrete units, then sequences with stronger internal predictability might be more likely to be considered as hypothesized words or morphemes – and therefore to be stored as chunks.

4.2 Goal and method

We consider the following research question: When extracting undersegmented chunks from speech during first language acquisition, are children more likely to extract (a) frequent, (b) internally predictable, or (c) short syllable sequences? We investigate (a) because frequent items, being acquired before less frequent exemplars (Ambridge et al., 2015), may simply be associated with a general learning advantage. We examine (b) and (c), on the other hand, because children might be biased to extract discrete linguistic units from unsegmented input; and short or predictable sequences, in contrast to frequent items, should have a higher chance of corresponding to such units.

Before answering the core research question, we first attempt to verify the assumption that short and predictable MSUs are more word-like than frequent MSUs. This is done by selecting various sets of multi-syllable utterances (MSUs) from the input English-speaking children typically receive. We refer to these as *chunk sets* – selections of uninterrupted syllable sequences which children could potentially store as chunks. If we are correct in assuming that short and internally predictable MSUs are more word-like, we should find that chunk sets with short and predictable MSUs are better-suited for selecting single-word utterances than sets with frequent MSUs.

After examining which types of chunk sets contain more words, we evaluate the likelihood that children store the MSUs in a given chunk set as unanalyzed units. One difficulty with devising such a method is that chunks might only be stored for brief periods and might only rarely be produced, if children use them at all. Because of this, methods for tapping into the chunk vocabulary of children should not rely on

child productions. Instead, we evaluate MSUs according to how well they perform at predicting when their constituent words are learned.

This method has previously been introduced by Grimm et al. (2017), who used an existing computational model (McCauley and Christiansen, 2014) to extract multi-word phrases from corpus data. Extracted phrases were used to predict the developmental stage at which children learn to produce the words contained within them. For this purpose, the incidence of the phrases containing each word was determined and correlated with the developmental stages at which children first produce the words. The correlation is negative, even when controlling for the frequency of words – i.e., words contained in many different phrases tend to be learned earlier than words contained in fewer phrases.

By way of explanation, Grimm et al. (2017) refer to segmentation: If phrases are stored as chunks, it should be easier to identify words contained in a large number of phrases, relative to words contained in fewer phrases. This would follow from an approach to word segmentation wherein the comparison of stored chunks leads to the detection of common sub-sequences – a strategy which Peters (1987) refers to as *phonological matching*. Assuming *phonological matching*, encountering a particular sub-sequence within many chunks could be advantageous in at least two ways: (1) Finding a particular sub-sequence within many different chunks might strengthen its hypothesized status as an independent unit; and (2) the more chunks contain a given word, the greater the chance that units which are encountered in the future can be split from one of those chunks – a strategy infants could, in principle, use during segmentation (Bortfeld et al., 2005).

Expanding on this, we evaluate chunk sets according to how well included MSUs perform at predicting the age at which children first produce the words contained within them. MSUs which are stored as chunks should perform well, whereas those that are never stored should perform poorly. Thus, whole-sequence frequency will be implied as a determinant of chunkhood to the extent that chunk sets containing frequent MSUs can predict when their component words are learned; syllabic predictability will be implied to the extent that internally predictable MSUs predict word learning; and sequence length will be implied to the extent that chunk sets with short MSUs predict word learning.

4.3 Analysis I: Chunk selection

In this analysis, we describe the method used to select chunk sets, which we define as subsets of the MSUs found in English child-directed speech. Our method involves ranking MSUs by (1) syllable length, (2) syllable predictability, and (3) frequency – followed by selecting the top N MSUs from each ranking.

4.3.1 Method

To select chunk sets from English child-directed speech (CDS), we rely on three properties: (1) the overall frequency of MSUs in CDS, (2) their length in syllables, and (3) the average predictability of adjacent syllables. Given a set of MSUs from a corpus of CDS, we rank MSUs by (1) – (3), and we select the top N items from each ranking. MSUs are ranked from most to least frequent, from shortest to longest, and from most to least predictable. We thus obtain three chunk sets – corresponding to the N most frequent, N shortest, and N most internally predictable MSUs.

Corpora

We extract MSUs from transcribed CDS, which differs markedly from the speech used by adults to address other adults. Among other things, CDS consists of shorter phrases, contains more pauses, and is composed of a more limited vocabulary (Saxton, 2010). Its properties appear to facilitate word segmentation and word learning (Thiessen et al., 2005; Yurovsky et al., 2012), making it the obvious corpus choice. We obtain CDS samples from various corpora of transcribed speech exchanged between caretakers and young children, taken from the CHILDES database (MacWhinney, 2000a). A typical corpus consists of various transcripts based on interactions (e.g. reading a book, playing a game) involving a child or group of children and their caretakers. Given that individual corpora contain at most a few hundred thousand words, we collapse various English CHILDES sources into a North American corpus (NA corpus) and a British English corpus (BE corpus).[7] Since most corpora in the CHILDES database are transcribed at the word level, whereas we are interested

[7]See appendix B for a list of included corpora.

in processes which precede the segmentation of speech into words, we syllabify all corpora – motivated by the observation that neonates and infants perceive speech in terms of syllables (Bertoncini and Mehler, 1981; Jusczyk and Derrah, 1987; Bijeljac-Babic et al., 1993). We convert each word to a syllable representation by relying on a syllabified version of the Carnegie Mellon University (CMU) pronouncing dictionary (Bartlett et al., 2009).[8] We keep only those utterances whose words have an entry in the CMU dictionary. About 80 % of utterances survive this syllabification process. Table 4.1 summarizes other relevant statistics.

measure	BE	NA
# adult speakers	280	737
# children addressed	247	743
mean child age (months)	32.66 (SD = 9.25)	41.39 (SD = 23.45)
# utterances	1,467,445	1,319,102
mean utterance length (words)	4.55 (SD = 3.69)	4.46 (SD = 3.46)
# tokens	6,690,453	5,890,443
# types	49,206	35,699
# syllabified utterances	1,190,858	1,083,618
mean utterance length (words)	4.42 (SD = 3.45)	4.08 (SD = 3.09)
# syllabified tokens	5,266,479	4,428,993
# syllabified types	19,931	14,156

Table 4.1: Child-directed speech statistics.

Possible chunks

We consider full utterances from CDS as possible chunks, i.e. as syllable sequences from which to select chunk sets. Sampling smaller sequences would require mechanisms for decomposing utterances and could confound the results. For example, a decomposition based on TPs would pre-suppose that children prioritize syllable predictability when extracting chunks from speech. Working with full utterances avoids this problem. Moreover, the storage of utterances presents an easy solution to Peters

[8] `http://webdocs.cs.ualberta.ca/~kondrak/cmudict.html`

(1987)'s *initial extraction problem*: If children have no knowledge about linguistic units, the most straightforward hypothesis is to consider uninterrupted stretches of speech as potential units. We thus consider MSUs from CDS as candidates for inclusion in chunk sets.

However, to lessen the probability that included utterances are not idiosyncratic to particular child-caretaker dyads, we require that MSUs are produced by adults from at least two different CHILDES corpora. This reduces the number of available utterances, in the BE and NA corpus, from more than 1,000,000 to about 50,000 each. The reason for this fairly drastic step lies in the nature of our corpus material: Because we collapse data from a large number of different CHILDES corpora (10 for the BE and 41 for the NA corpus), with hundreds of child and adults speakers, most MSUs will not form part of the input received by the children addressed in the different corpora. For example, the BE corpus contains the adult-produced MSU, *On Wednesday he ate through three plums*. Unsurprisingly, this MSU is only used once, to address a particular child, in a situation that is unlikely to occur with any of the other children whose input we are considering. Because of this, it would not make sense to include it as an utterance that children could, in general, store as an undersegmented chunk. Thus, to reduce the likelihood that such idiosyncratic MSUs are included in the aggregated BE and NA corpora, we filter MSUs by the number of individual CHILDES corpora within which they occur – requiring them to be used, on independent occasions, by caretakers from at least two of the (41 + 10 = 51) CHILDES corpora.

Furthermore, given that we consider the syllable as a primitive unit, single-syllable utterances are already fully segmented and cannot be considered as undersegmented chunks. For this reason, we require that the utterances included in chunk sets contain at least two syllables (i.e., we consider *multi-syllable* utterances / MSUs). Finally, to control for repetition, we exclude MSUs that consist of repeated occurrences of a single word. The three criteria (more than one syllable, no repetitions, used in at least two CHILES corpora) are met by 50,199 MSUs in the BE corpus and by 57,151 MSUs in the NA corpus.

Selection of chunk sets

From the available MSUs, we wish to select the N most frequent, N shortest, and N most internally predictable items as chunk sets. We thus need to fix the size of each chunk set to some N, where N must be smaller than the number of all MSUs. Otherwise, there would only be one chunk set, and it would contain all MSUs. Given some N, we then select MSUs according to their frequency, their length in syllables, and the mutual predictability of their syllables. We determine frequency by counting how often MSUs appear in CDS, length by counting the number of syllables in each MSU, and predictability by averaging over the conditional probabilities between adjacent syllables.

More formally, each syllable u_i within the MSU $u_1, u_2...u_n$ can be associated with a set P_i of conditional probabilities:

$$P_i = \begin{cases} \{p(u_i|u_{i-1}), p(u|u_{i+1})\} & \text{if } i > 1 \wedge i < n \\ \{p(u_i|u_{i-1})\} & \text{if } i > 1 \wedge i = n \\ \{p(u_i|u_{i+1})\} & \text{if } i = 1 \wedge i < n \end{cases}$$

The predictability score of a given MSU is then defined as the average of the conditional probabilities associated with the syllables in a given MSU. This definition is inspired by the oft-replicated finding that infants are sensitive to the TPs (conditional probabilities) between syllables (Saffran et al., 1996; Aslin et al., 1998; Thiessen and Saffran, 2003), suggesting that the local predictability of syllables within sequences is an early segmentation cue.

At this point, N is an obvious tweakable parameter. As mentioned, N must be smaller than the number of available MSUs. Otherwise, the only chunk set would contain all MSUs, and we would not be able to distinguish between especially frequent, short, or internally predictable MSUs. At the same time, N should not be extremely small either. For example, it would not make sense to set $N = 1$. But even values in the tens or hundreds might not be sufficiently large. Since we wish to predict the age at which words are learned from the number of MSUs within which these words are contained, it would be good to operate with fairly large chunk sets, to ensure that a majority of target words will in fact appear within some MSU.

For the current illustrative purpose, we set $N = 10,000$. In subsequent analyses, however, we report results for many possible choices of N.

4.3.2 Results and discussion

To illustrate the chunk set selection procedure, we focus on example sets from the BE corpus – consisting of the $N = 10,000$ shortest MSUs, the N most frequent MSUs, and the N most predictable MSUs. Table 4.2 summarizes statistics pertaining to the three sets. As expected, the average syllable count of the N shortest MSUs is lowest; the average frequency count of the N most frequent MSUs is highest; and the average predictability score the of N most internally predictable MSUs is largest. Overlap between the three sets is limited to below 30 %, indicating that the chunk sets contain fundamentally different types of MSUs.

chunk set	short	frequent	predictable
mean frequency	18 (SD: 212)	**31** (SD: 220)	13 (SD: 179)
mean predictability	0.10 (SD: 0.14)	0.15 (SD: 0.13)	**0.30** (SD: 0.09)
mean length (syllables)	**2.36** (SD: 0.48)	3.73 (SD: 1.45)	4.42 (SD: 1.61)
overlap with shortest	—	28.6 %	15.49 %
overlap with most frequent	—	–	25.5 %
overlap with most predictable	—	—	—

Table 4.2: Statistics for chunk sets with the $N = 10,000$ shortest, most frequent, and most predictable MSUs.

Table 4.3 contains example MSUs from each chunk set.[9] The most predictable MSUs (e.g. *brilliant*, *breakfast*) correspond to syllable sequences whose component syllables, if they do occur, have a high chance of occurring within the given MSUs. For example, the syllable corresponding to *brill-* occurs only to the left of the syllable corresponding to *-iant*, and the syllable corresponding to *-iant* occurs only to the right of the syllable corresponding to *brill-*. This means that the conditional probabilities associated with the two syllables are both 1.0, leading to a 1.0 average

[9]For readability, each MSU is presented in its orthographic transcription. But note that in our experiments, MSUs are represented as unsegmented syllable sequences. For example, the orthographic transcription *that's wonderful* is underlyingly represented as *thats1-won1-der0-ful0* (1 = primary stress, 2 = secondary stress, 0 = no stress).

N shortest			N most frequent			N most predictable		
MSU	freq	pred	MSU	freq	pred	MSU	freq	pred
more bricks	9	0.03	okay	12,101	0.57	vampire	3	1.00
push out	1	0.00	uhu	7,613	0.98	brilliant	317	1.00
nice tea	2	0.00	that's right	7,474	0.20	breakfast	30	1.00
quiet	23	0.86	pardon	5,033	0.75	trowel	4	0.99
stop there	3	0.00	that's it	4,823	0.08	uhu	7,613	0.98
a leg	2	0.02	come on	4,734	0.31	grandad	35	0.97
bread yeah	1	0.00	oh dear	4,697	0.46	children	13	0.96
train what	2	0.00	what's that	3,747	0.19	Fraser	1,627	0.96
left eye	2	0.00	thank you	3,002	0.51	nonsense	4	0.96
right back	3	0.00	oh no	2,945	0.04	hello	1,680	0.95
London	15	0.49	good girl	2,293	0.36	jigsaw	15	0.95
red bear	2	0.00	there you go	2,262	0.10	hungry	6	0.94
what room	1	0.00	I don't know	2,248	0.19	costume	4	0.94
the farm	2	0.21	what is it	2,225	0.13	husband	1	0.94
window	20	0.81	is it	2,151	0.20	croissant	10	0.94

Table 4.3: Top 15 MSUs from chunk sets containing the (1) N shortest, (2) N most frequent, and (3) N most internally predictable MSUs.

predictability score for *brilliant*. Strikingly, the 15 most internally predictable MSUs all correspond to individual words – with both very high and very low frequency counts.

The 15 most frequent MSUs, on the other hand, include single words (e.g. *okay*) and idiomatic sounding multi-word utterances (e.g. *oh dear*, *I don't know*). The much lower predictability scores associated with these MSUs indicate that their syllables are less strongly tied to one another: Even though MSUs such as *I don't know* are frequently used, the syllables corresponding to *I*, *don't* and *know* are also frequently used in MSUs other than *I don't know*. The shortest MSUs, finally, correspond to both disyllabic words (e.g. *quiet, window*) as well as disyllabic multi-word utterances (e.g. *stop there*). Since these MSUs are only selected according to length in syllables, their frequency counts and predictability scores are quite variable.

4.4 Analysis II: Which multi-syllable utterances correspond to single words?

In considering frequent, predictable, and short MSUs, we have been assuming that the latter two MSU types are more word-like than the former. The previous analysis certainly suggests that the most internally predictable MSUs are more word-like than the most frequent MSUs – with the top 15 predictable items all corresponding to single-word utterances (cf. table 4.3). It is possible, however, that the top 15 MSUs are special cases, with fewer single-word utterances among the MSUs further down the rank distribution. In the current analysis, we use a more rigorous method to determine which of the three metrics (syllable length, frequency, or syllabic predictability) is best-suited for selecting MSUs that correspond to single words. If our initial assumption is correct, MSU length and syllabic predictability should be better-suited for selecting single-word MSUs than whole-sequence frequency.

4.4.1 Method

Any method used to establish which of the three metrics is most useful for selecting single-word MSUs should address two key issues: (1) the need for an appropriate performance metric, and (2) the potentially confounding effect of the tweakable parameter N (chunk set size). We address both issues below.

Classification metrics

We would like to quantify whether different types of chunk sets – i.e., subsets of the available MSUs – are well-suited for selecting single-word MSUs. In the best case, a given chunk set will contain all and only single-word MSUs; and in the worst case, it will not contain any single-word MSUs. We can thus frame the selection of chunk sets as a classification task, where MSUs included in a particular chunk set are classified as *words*, and excluded MSUs are classified as *non-words*.

To quantify classification performance, we use a *precision* and a *recall* metric: the proportion of words contained within a given chunk set, and the proportion of words correctly selected out of all available words. More formally, let N be the chunk set

size, W_C the number of words within the chunk set, and W the number of words outside of the chunk set. Precision and recall are then defined as follows:

$$Precision = \frac{W_C}{N} \tag{4.1}$$

$$Recall = \frac{W_C}{W_C + W} \tag{4.2}$$

Precision equals 1.0 if and only if the chunk set contains only single-word MSUs, and recall equals 1.0 if and only if the chunk set contains all available single-word MSUs. A chunk set that contains all and only single-word MSUs will thus lead to maximum precision and recall. To quantify the notion that well-performing chunk sets should maximize both precision and recall, we track overall classification performance via the *F-score*, defined as the harmonic mean of precision and recall (a measure of classification performance commonly used in computational linguistics and studies investigating speech segmentation in children – see e.g. Goldwater et al. (2009) and the references therein). While we do not expect to achieve maximum scores with our chunk sets, we nevertheless expect to obtain informative differences between classification outcomes.

Effect of chunk set size

The parameter N (chunk set size) could in principle take any value between 1 and the total number of MSUs (50,199 for the BE corpus and 57,151 for the NA corpus). Crucially, robust results should emerge across all choices of N – excluding only large and small values. Large values close to the number of all available MSUs should lead to similar result for the three chunk sets, since each set will contain the same selection of MSUs. But N should not be too small either: The BE and NA corpus contain 1,856 and 2,159 single-word MSUs, respectively, and chunk sets containing fewer MSUs cannot maximize recall. However, as long as N is neither too small nor close to the number of all MSUs, we should see similar results. We examine this by calculating classification performance for various N.

Statistical analysis

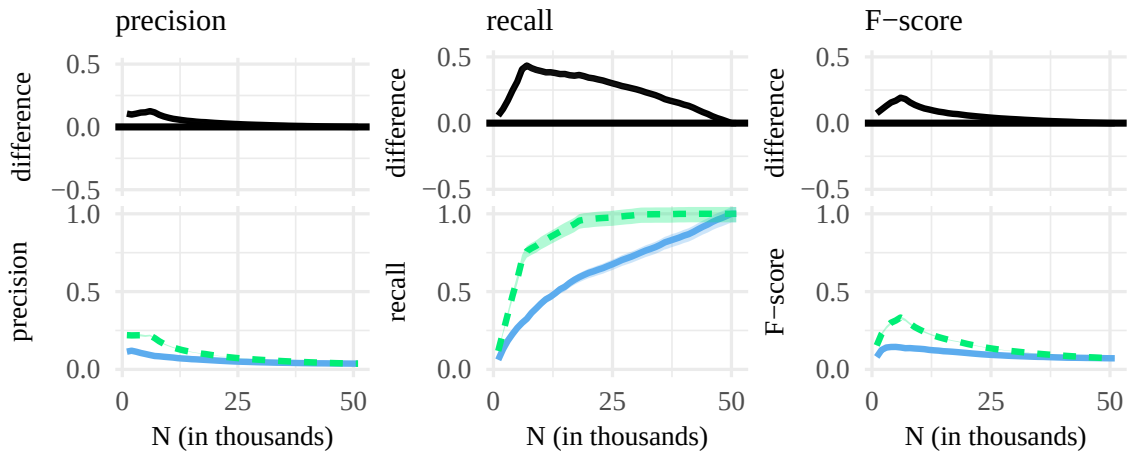
We calculate 95 % percent confidence intervals for precision, recall, and F-score via statistical bootstrapping (Davison and Hinkley, 1997), with each bootstrap based on 100 random samples with replacement, and a sample size equal to the number of data points.

For example, consider a chunk set of size $N = 10,000$, selected from the 50,199 MSUs in the BE corpus. In this case, each of the 10,000 MSUs included in the chunk set is assigned a *word* label, and the remaining 40,199 MSUs are labeled as *non-words*. To bootstrap confidence intervals for the three classification metrics, we first take a random sample (with replacement) of 50,199 MSUs (all available data points). Next, we calculate precision and recall for this sample, based on the labels assigned during the classification step. By repeating this procedure 100 times, we obtain a normal distribution of classification metrics – and their 95 % confidence intervals correspond to the range between the 2.5th and the 97.5th percentiles. When comparing metrics derived from two different chunk sets, we bootstrap 95 % confidence intervals for the difference between them. If zero is not contained within this interval, we can claim with 95 % certainty that the difference is not due to chance.

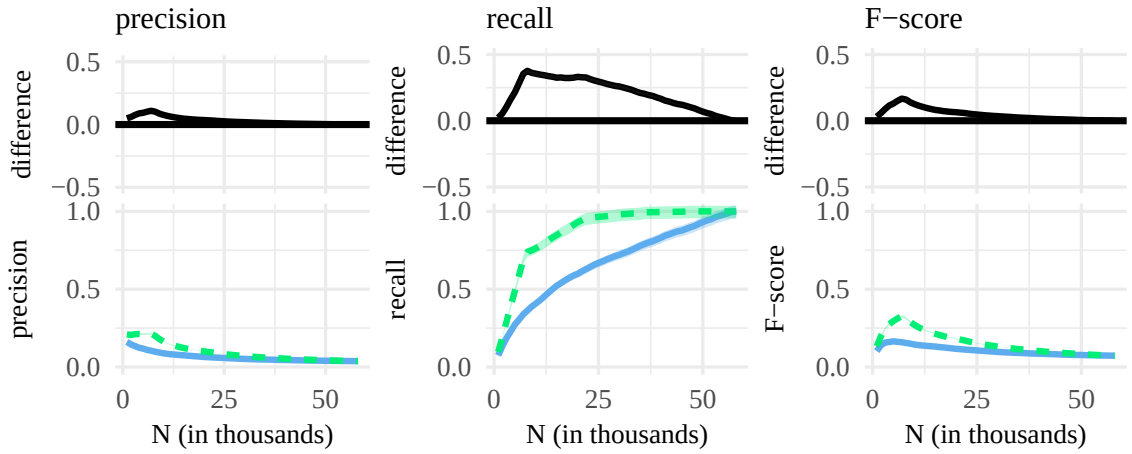
4.4.2 Results and discussion

We compare classification metrics associated with three different chunk sets – containing either the shortest, the most frequent, or the most internally predictable MSUs. This design yields three pairwise comparisons of chunk sets: (1) shortest vs. most frequent, (2) shortest vs. most predictable, and (3) most predictable vs. most frequent – each conducted for three metrics of classification performance (precision, recall, F-score), using chunk sets taken from two corpora of English CDS (the BE and the NA corpus). The comparisons are summarized, in turn, by figures 4.1, 4.2, and 4.3 below. Each figure plots, as a function of N, classification performance for two different chunk sets, as well as the difference between performance scores. On the x-axis, we increment N in steps of 1,000 – beginning at $N = 1,000$ and ending at the maximum possible chunk set size.

Figure 4.1 shows classification performance for chunk sets containing the N shortest and the N most frequent MSUs. Across both corpora, precision is highest at $N =$



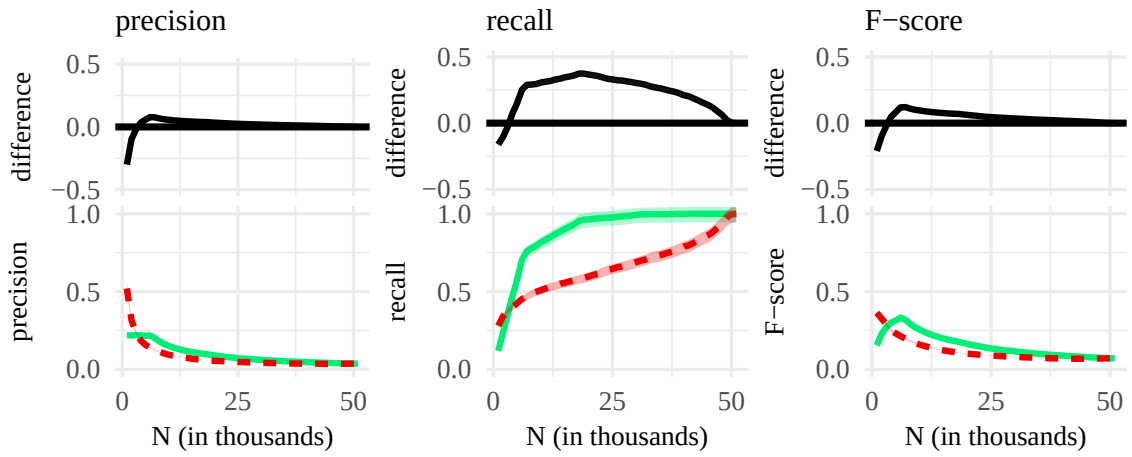
(a) Chunk sets taken from the BE corpus.



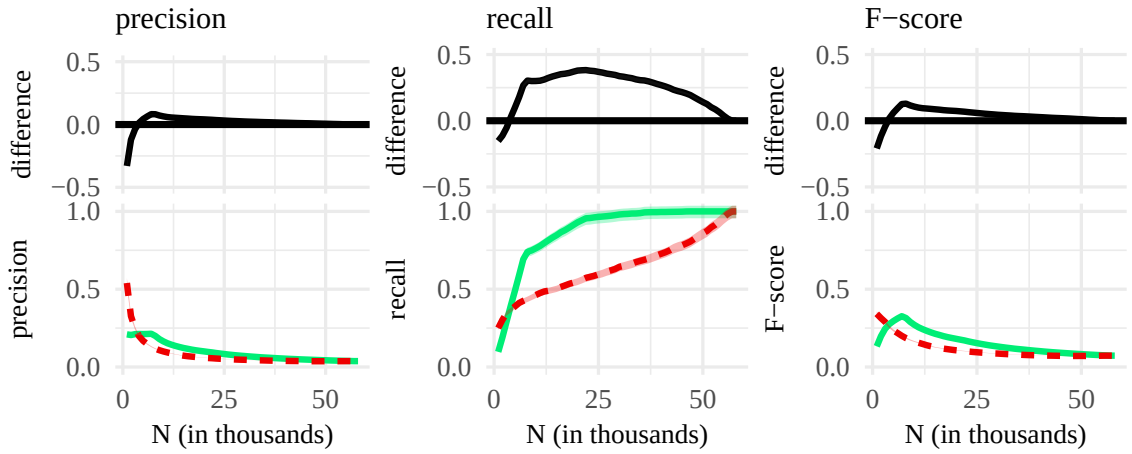
(b) Chunk sets taken from the NA corpus.

Figure 4.1: Bottom of each subplot: classification performance for the N shortest (green line) and N most frequent MSUs (blue line), with 95 % confidence intervals. Top: difference between green and blue line, with 95 % confidence intervals.

1,000, where it is just above 0.2 for the shortest and between 0.1 – 0.15 for the most frequent MSUs. That is, ca. 20 % of the shortest 1,000 MSUs correspond to single words, while the same is true for only 10 – 15 % of the most frequent MSUs. Precision then decreases with an increasing chunk set size – to about 15 % and 8 % at $N = 10,000$, and to ca. 7 % and 4 % at $N = 25,000$. At $N = 50,000$, the two chunk sets each contain almost all available MSUs, so precision scores derived from either set are very close to one another. However, until the chunk sets contain approximately half of the available MSUs, precision is clearly higher for the N shortest MSUs, with the scores approaching each other as N is further increased.



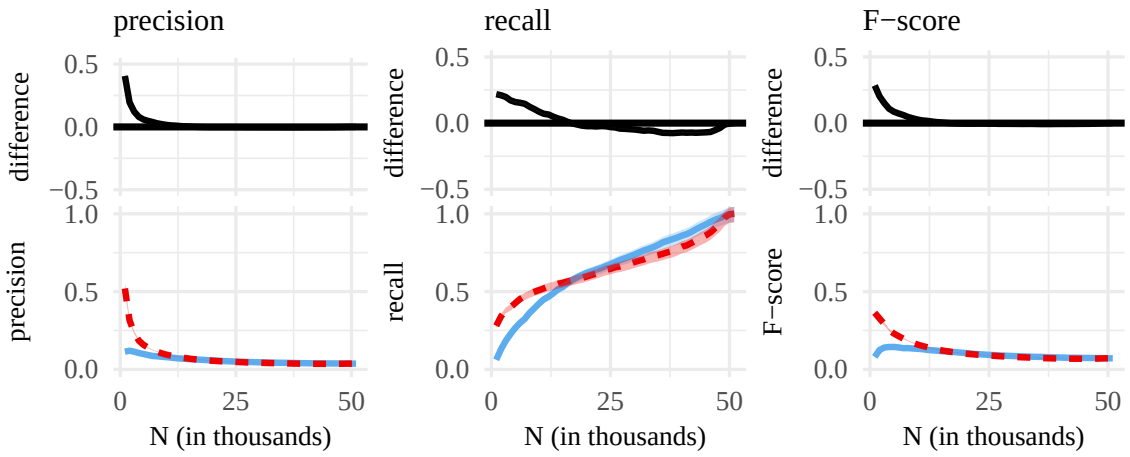
(a) Chunk sets taken from the BE corpus.



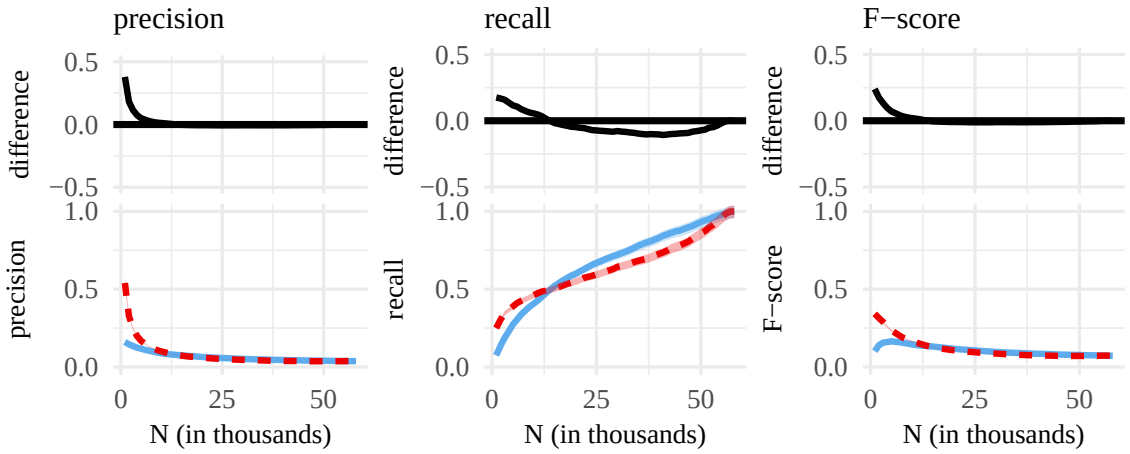
(b) Chunk sets taken from the NA corpus.

Figure 4.2: Bottom of each subplot: classification performance for the N shortest (green line) and N most internally predictable MSUs (red line), with 95 % confidence intervals. Top: difference between green and red line, with 95 % confidence intervals.

Recall increases rather than decreases over successive chunk set sizes. This is because recall can be maximized, at the cost of low precision, by assigning the *word* label to every MSU. Thus, at $N = 50,000$, recall is close to 1.0 for both chunk sets (i.e., they each contain close to 100 % of single-word MSUs) simply because they contain almost all available MSUs, while precision is close to zero (i.e., the proportion of selected single-word MSUs is very low). Conversely, at $N = 1,000$, recall is minimized, while precision is maximized. Thus, smaller chunk sets contain a large proportion of words, but the majority of single-word MSUs remains undetected. Crucially, with the exception of chunk sets close to the maximum possible size, recall is generally



(a) Chunk sets taken from the BE corpus.



(b) Chunk sets taken from the NA corpus.

Figure 4.3: Bottom of each subplot: classification performance for the N most internally predictable (red line) and N most frequent MSUs (blue line), with 95 % confidence intervals. Top: difference between red and blue line, with 95 % confidence intervals.

higher for short rather than frequent MSUs.

The harmonic mean of precision and recall (F-score) is maximized at $N \approx 10,000$ (short MSUs ≈ 0.25; frequent MSUs ≈ 0.15). Generally, chunk sets containing short rather than frequent MSUs translate into significantly higher F-scores. The difference begins to disappear at around $25,000$, reflecting the fact that as we increase the size of chunk sets, the MSUs contained within them tend to overlap more. As long as we focus on small N, however, chunk sets containing short MSUs are clearly better-suited for selecting words.

When comparing short to internally predictable MSUs (figure 4.2), we find that very small chunk sets with predictable MSUs ($N = 1,000$ and $N = 2000$) contain a noticeably larger amount of single-word MSUs than equally sized chunk sets with frequent MSUs (50 % vs. 10 – 15 % at $N = 1,000$; 20 % vs. 12 – 15 % at $N = 2,000$). At $N = 1,000$, predictable MSUs also yield slightly better recall; but almost all subsequent chunk sets capture a much larger proportion of the available single-word MSUs if they are selected according to syllable length rather than predictability. This is reflected in the F-score, which is significantly higher for short MSUs, as long as N is not too small or too large. Generally, then, chunk sets containing short rather than predictable MSUs tend to be better-suited for selecting utterances corresponding to individual words.

The only exception to this comes in the form of the 1,000 – 2,000 most predictable MSUs, which tend to be words more often than their counterparts in equally sized chunk sets with short MSUs. One possible explanation for this pattern is that our predictability metric picks up on low-frequency words, with syllables that occur in only a handful of syllabic contexts, leading to relatively high conditional probabilities for MSUs containing such syllables. But the most predictable MSUs include both low-frequency words (*e.g. vampire, husband, costume*), as well as more common words such as *hello* or *brilliant* (cf. table 4.3). Moreover, the average frequency of the 1,000 most predictable MSUs (62 in the BE corpus, 67 in the NA corpus) is actually *higher* than the average frequency of the 1,000 shortest MSUs (26 in the BE corpus, 15 in the NA corpus) – demonstrating that highly predictable MSUs are not all low-frequency items.

In the last remaining comparison (predictability vs. frequency, figure 4.3), the high classification performance of small sets containing predictable MSUs exceeds the performance associated with (small) sets of frequent MSUs. Unlike MSU length, whole-sequence frequency is not associated with particularly high recall scores – and larger chunk sets containing frequent MSUs perform, at best, only slightly better than (larger) chunks sets of predictable MSUs. On the whole, predictability thus wins out over frequency.

Syllable length – chiefly due to high recall – in turn won out over predictability (figure 4.2) and clearly lead to better performance than frequency (figure 4.1). Ordered from worst to best classification performance, that is, we obtain the following

ranking: (1) frequency, (2) syllabic predictability, (3) length in syllables. Of course, a small number of constituent syllables does not guarantee that a given MSU will in fact correspond to a single word. But by and large, selections of short MSUs are better-suited for picking out single-word utterances than either frequent or internally predictable MSUs.

This verifies our initial assumption that whole-sequence frequency is a poorer indicator of wordhood than either sequence length or syllabic predictability. In the following analysis, we investigate which of the three MSU types are most likely to be stored, during early speech segmentation, as undersegmented chunks.

4.5 Analysis III: Which multi-syllable utterances best predict the age of first production of words?

In the previous analysis, we examined whether short, frequent, or predictable MSUs are more likely to correspond to single words. Now, we evaluate how well the three different types of MSUs predict the age at which their component words are acquired, arguing that MSUs which are well-suited for predicting word learning are also more likely to be stored as undersegmented chunks. Since frequency of occurrence seems to confer a general learning advantage (Ambridge et al., 2015), children might preferentially store frequent MSUs as chunks. It is also possible, however, that children are biased to extract and store more discrete, word-like MSUs. If true, we should expect children to store short and possibly internally predictable MSUs, to the exclusion of more frequent items.

4.5.1 Method

Following Grimm et al. (2017), we use the MSUs in a particular chunk set to predict the age at which children first produce the words contained within the MSUs. Grimm et al. (2017) found that words which are contained in a large number of multi-word phrases are produced at earlier stages than words contained in fewer phrases. As a possible explanation, they argued that children commit phrases to long-term

memory as holistic chunks – i.e., before they have discovered that the phrases are composed of smaller linguistic units. As a result, the more chunks containing a particular word X are stored in long-term memory, the higher the likelihood that children discover X as a separate linguistic unit – and the earlier they subsequently produce X. We thus evaluate how well the MSUs from different chunk sets perform at predicting the age of first production (henceforth *AoFP*) of their component words. If children store frequent MSUs as chunks – prior to having detected the words contained within those chunks –, then frequent MSUs should perform best. Conversely, if they store short or internally predictable MSUs, short or predictable MSUs should perform best.

We implement this idea by using AoFP as a dependent variable in multiple linear regressions. Given a chunk set and a set of words with associated AoFP values (henceforth *target words*), we count – for each target word – how many MSUs within the chunk set contain it. The resulting value, the number of MSUs per target word (henceforth $\#MSU$), is then used as an independent variable. We denote this measure $\#MSU\text{-}F$ when calculated based on the N most frequent MSUs, $\#MSU\text{-}S$ when calculated based on the N shortest MSUs, and $\#MSU\text{-}P$ when calculated based on the N most predictable MSUs. Thus, by using $\#MSU\text{-}F$, $\#MSU\text{-}S$, and $\#MSU\text{-}P$ to predict AoFP, we evaluate how well the shortest, most frequent, and most predictable MSUs perform at predicting the time course of word learning. If children store short MSUs as chunks, then $\#MSU\text{-}S$ should perform best at predicting AoFP; if they store frequent MSUs, $\#MSU\text{-}F$ should perform best; and if they store predictable MSUs, $\#MSU\text{-}P$ should emerge as the best-performing predictor.

To evaluate performance, we track two statistics: (1) the regression coefficient (β), measuring how strongly the AoFP of targets decreases as we increase $\#MSU$; and (2) the amount of variance within AoFP that can be accounted for by including $\#MSU$ in the regression models (R^2). We expect that a robust result should lead to comparable effects across the two statistics. For example, if words contained within predictable MSUs are learned earlier than words contained within frequent or short MSUs, words with high $\#MSU\text{-}P$ counts should be learned earlier than words with high $\#MSU\text{-}F$ or $\#MSU\text{-}S$ counts – and this should be reflected in stronger effects, across the two statistics, for $\#MSU\text{-}P$.

Age of first production

Selecting suitable AoFP data is critical, as the procedure used to obtain AoFP estimates could confound the results. Specifically, children might produce chunks without having learned about the words within them. We should make sure, in other words, that AoFP estimates are based on word productions which are not performed in the context of the MSUs used to predict AoFP. We control for this in the first of two AoFP data sets, which we estimate from the children addressed in the two CDS corpora. And to ensure the robustness of these corpus-derived AoFP estimates, we replicate our results on an existing data set derived from parent-report questionnaires.

Corpus-derived AoFP The first AoFP data set is estimated from the transcribed speech of the children addressed by the caregivers in the two aggregated CHILDES corpora.[10] Here, we treat a word as having been acquired at the earliest developmental stage at which any child within a corpus produces it. In doing so, we only consider word productions from outside of the adult-produced MSUs. For example, if a child produces the word *day* as part of the MSU *what a great day*, and this MSU is also used to predict AoFP, we do not consider the child production. Further, we do not consider word productions if the words are produced within sub-sequences of adult-produced MSUs. We would not, that is, consider child productions like *it's a great day*, since *a great day* is a sub-sequence of *what a great day*. Corpus-derived AoFP thus is a conservative estimate, where a given word is considered as learned at the earliest developmental stage at which any child first produces it – in a context without overlap with the adult-produced MSUs.

Developmental stage is defined in terms of mean length of utterance (MLU) – the average child utterance length, in tokens, within a transcript (CHILDES corpora consist of transcripts, recorded at different points during the target child's development). We induce MLU rather than AoFP estimates because children who are close in age may nevertheless be far apart in language development. Being a more robust estimator, MLU controls for developmental differences (Parker and Brorson, 2005). Since transcripts contain varying numbers of utterances, the average utterance length per transcript is biased with respect to transcript

[10] See appendix B, table B.1.

length. We rectify this issue by estimating MLU for each transcript via statistical bootstrapping (Davison and Hinkley, 1997). Each bootstrap is based on 10,000 random samples with replacement, with the sample size equal to the number of child utterances per transcript. We thus induce MLU rather than AoFP estimates but will, for simplicity, refer to a word's MLU as its AoFP. To calculate an estimate for a given word, we bootstrap the set of MLUs γ for all transcripts within which a child uses the word outside of an adult-produced MSU, and we choose the smallest value in γ as the word's AoFP. Performing this procedure for all words produced by children in at least two of the considered CHILDES corpora, we obtain AoFP estimates for the aggregated BE and NA corpus – covering 7,565 and 9,482 different child-produced words.

CDI-derived AoFP The corpus-derived AoFP estimates are sensitive to high-frequency words, making it desirable to replicate results on data that do not rely on language sampling. We obtain such AoFP estimates from the wordbank database (Frank et al., 2017)[11], a repository with results from parent-report questionnaires (MacArthur–Bates Communicative Development Inventories / CDI). Wordbank archives data from various administrations of the CDI. The largest English data set pools responses from parents of 6,945 (American) English-speaking children between the ages of 16 and 30 months and covers 680 words and phrases.[12] At each of the 15 months covered by the questionnaires, parents had to indicate whether their child produces a list of words. Word-level data are then represented as the percentage of parents who reported, for a given month, that their child can successfully produce the word in question. Excluding compounds, phrases, and words that are specific to particular children (baby sitter's name, child's own name, pet's name), we derive AoFP estimates for 647 words by counting words as having been learned if at least 50 % of the children were reported to produce it. Due to the design of the CDI, we cannot rule out that parents reported on child productions of chunks instead of individual word productions. Corpus-derived AoFP, which controls for chunk productions, is thus of primary importance. And to increase confidence in the robustness of results, CDI-derived AoFP is used to replicate results achieved with the former.

[11] Available online: `http://wordbank.stanford.edu/`

[12] Data were downloaded on 01/08/2018.

Since the children whose parents filled in the CDI forms were no older than 30 months, we restrict the MSUs included in chunk sets for the analyses with CDI-derived AoFP – considering only MSUs which were produced in the presence of children aged 30 months or less.

Validity of AoFP estimates

It would raise methodological concerns if we simply assumed the validity of corpus-derived AoFP. The CDI-derived estimates, on other hand, have been validated on different measures of children's expressive vocabularies (Dale, 1991; Fenson et al., 2007). This is why we include CDI-derived estimates, and why it is important that similar results are obtained with both data sets. To further increase our confidence in both types of estimates, we compare them to the only publicly available English age of acquisition estimates that come directly from children: Morrison et al. (1997) had children of varying ages perform a picture naming task; and if a child was able to produce the correct picture name, he or she was considered to have acquired the word.

Presumably because of time constraints, Morrison et al. (1997) provide age of acquisition for a restricted set of 297 picturable nouns. While insufficient for our analyses, we can still use their data to verify our estimates: The correlation between their estimates and corpus-derived AoFP is strongly positive (Spearman's $rho = 0.65$ for the BE children and $rho = 0.59$ for the NA children, based on 274 and 272 shared words, respectively; $p < 10^{-8}$). The correlation with CDI-derived AoFP is also fairly strong ($rho = 0.50$, based on 117 shared words, $p < 10^{-8}$). This pattern strengthens our confidence in the validity of (corpus- and CDI-derived) AoFP estimates.

Co-variates

The independent variable is *#MSU*, which we use to predict AoFP. Grimm et al. (2017) found that a similar predictor is negatively correlated with AoFP, leading us to also expect a negative correlation between *#MSU* and AoFP (meaning that words contained in many MSUs would be learned earlier than words contained in

fewer MSUs). But such a correlation could be due to collinearity with several co-variates, the most obvious of which is word frequency. Frequency of exposure is associated with a general learning advantage (Ambridge et al., 2015), and words with a high *#MSU* count tend to be frequent. Grimm et al. (2017) controlled for frequency, but there are other possible confounds.

We attempt to remedy this by including the following co-variates: (1) the corpus frequency, in CDS, of each target word (*Freq*), (2) concreteness ratings (*Con*), (3) length in syllables (*Nsyl*), and (4) phonological neighborhood density (*PhonN*).[13] *Freq* must be included to control for frequency of exposure, and *Con* is included to control for semantic properties of target words. *Nsyl* and *PhonN*, meanwhile, are meant to control for confounds having to do with the phonological properties of target words. Concreteness ratings for 40,000 lemmas are taken from Brysbaert et al. (2014)[14], who collected them from over 4,000 participants via Mechanical Turk. Since ratings were collected for lemmas, we assigned the lemma rating to all its word forms. Given a target word, *PhonN* is defined as the number of homophones, plus the number of words that can be derived from the target by either adding, deleting, or substituting a single phoneme. *PhonN*, together with *Nsyl*, is derived from the syllabified CMU pronouncing dictionary that was also used to convert our corpora to syllable representations. Braginsky et al. (2016) have recently shown that variables similar to *Freq*, *Con*, and *Nsyl* predict age of acquisition: Early-acquired words tend to be frequent, concrete, and (at least in English) short. We additionally include *PhonN*, as words in dense neighborhoods tend to be early-learned, possibly due to a memory advantage of highly connected lexical representations (Storkel, 2004, 2009).

Statistical analyses

When working with the corpus-induced AoFP data, we use AoFP estimates from children who were not addressed in the corpus used to calculate *#MSU*. In other words, we use AoFP from the children addressed in the NA corpus for regression models which include *#MSU* and frequency counts from the BE corpus; and we use AoFP from the children addressed in the BE corpus for regression models which include independent variables from the NA corpus. This design de-couples the inde-

[13] See appendix B, table B.2, for a collinearity analysis.

[14] `http://crr.ugent.be/archives/1330`

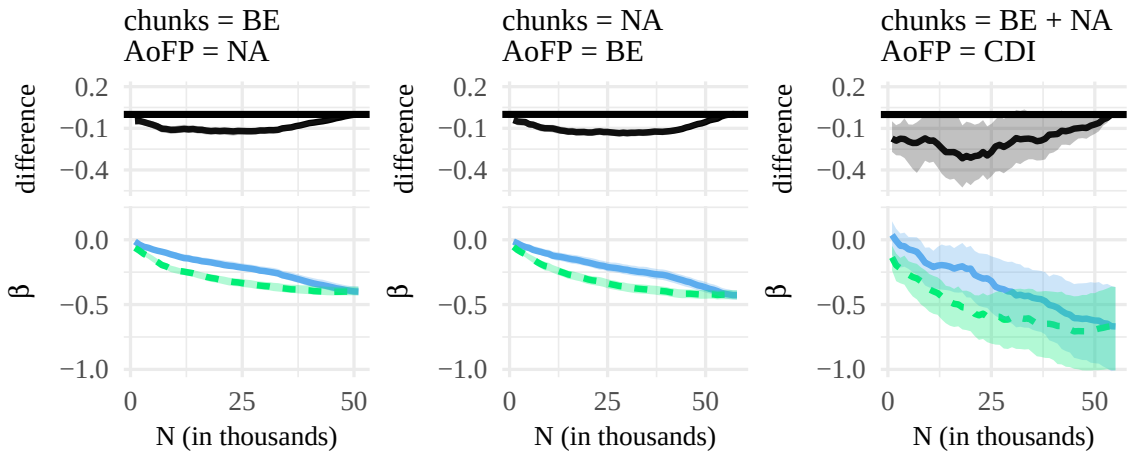
pendent variable from corpus-induced AoFP and is meant to increase the generality of our study's implications. Since the CDI-derived AoFP estimates come from an external source, we use MSUs from both the BE and NA corpus to predict the CDI data – although restricted, as mentioned, to MSUs produced in interactions with children aged 30 months or less.

This leaves us with three different corpus-AoFP pairings: (1) BE corpus with AoFP from NA children, (2) NA corpus with AoFP from BE children, and (3) age-restricted BE and NA corpus with CDI-derived AoFP. The corpus material used in each analysis contains around 50,000 MSUs. Regression analyses are based on all words for which *PhonN*, *Nsyl*, *Con*, and AoFP estimates are available: 6,208 and 5,577 words for analyses (1) and (2), and 615 words for analysis (3). In order to avoid problems from zero counts, $\#MSU$ was increased by 1. All variables were log-transformed and then standardized (via transformation to Z-scores). We compute 95 % percent confidence intervals for regression coefficients and R^2 values via statistical bootstrapping (Davison and Hinkley, 1997), with each bootstrap based on 100 random samples with replacement, and a sample size equal to the number of data points. When comparing two effects, we bootstrap 95 % confidence intervals for the difference between them. If zero is not contained within this interval, we can claim with 95 % certainty that the difference is not due to chance.

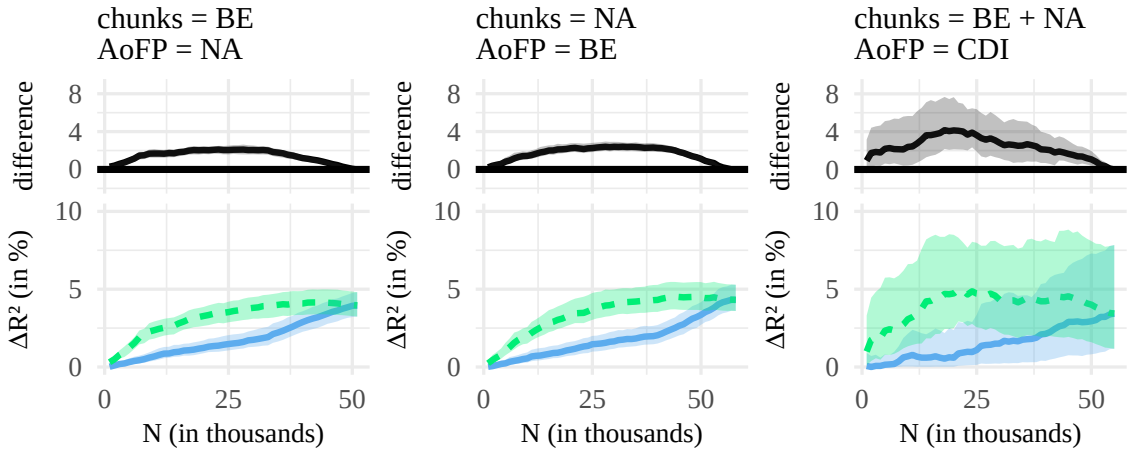
4.5.2 Results and discussion

We compare the effects associated with three independent variables ($\#MSU\text{-}S$, $\#MSU\text{-}F$, $\#MSU\text{-}P$), resulting in three pairwise comparisons: (1) $\#MSU\text{-}S$ vs. $\#MSU\text{-}F$, (2) $\#MSU\text{-}S$ vs. $\#MSU\text{-}P$, and (3) $\#MSU\text{-}P$ vs. $\#MSU\text{-}F$. Each of these comparisons is conducted for two statistics (β, R^2) and three corpus-AoFP pairings (calculate $\#MSU$ from BE corpus and AoFP from NA corpus; calculate $\#MSU$ from NA corpus and AoFP from BE corpus; calculate $\#MSU$ from age-restricted NA plus BE corpus and use CDI-derived AoFP). Fig 4.4 summarizes the first set of comparisons, for (1) $\#MSU\text{-}S$ vs. $\#MSU\text{-}F$. Figure 4.5 then summarizes (2) $\#MSU\text{-}S$ vs. $\#MSU\text{-}P$, and figure 4.6 summarizes (3) $\#MSU\text{-}P$ vs. $\#MSU\text{-}F$. We discuss each comparison in turn.

Figure 4.4a shows, as a function of N, the regression coefficients for $\#MSU\text{-}S$ and



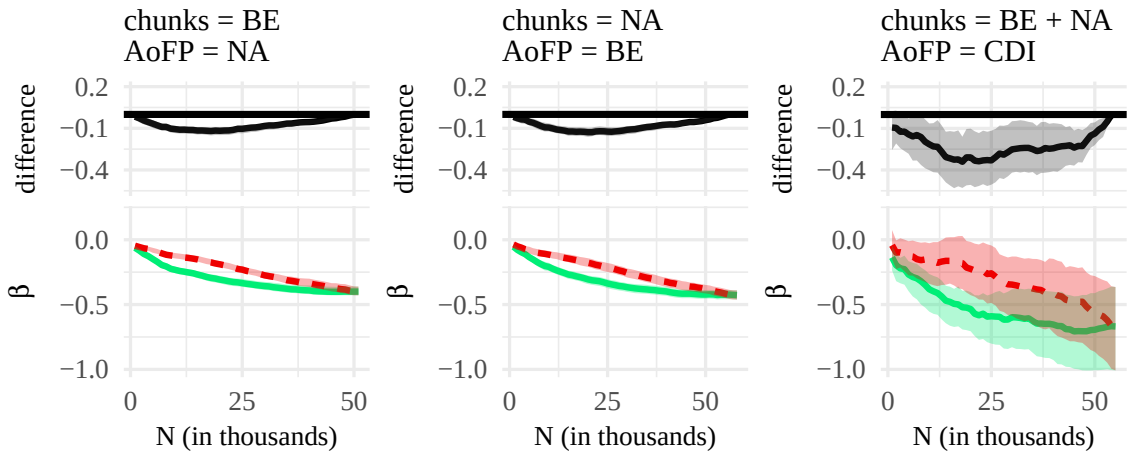
(a) Bottom: regression coefficients (β), with 95 % confidence intervals. Top: difference between green ($\#MSU\text{-}S$) and blue ($\#MSU\text{-}F$) line, with 95 % confidence intervals.



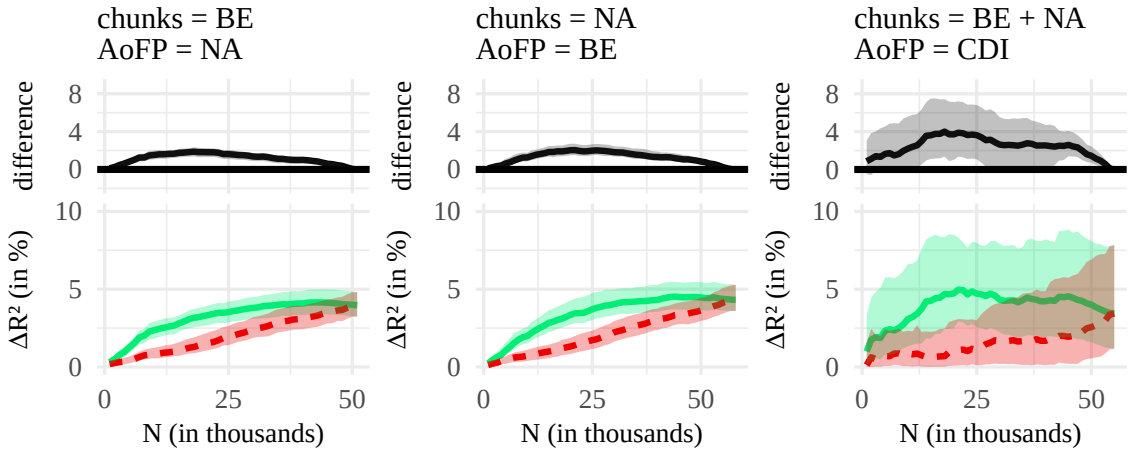
(b) Bottom: amount of variance in AoFP (ΔR^2 in %), with 95 % confidence intervals, that can be explained by $\#MSU$. Top: difference between green ($\#MSU\text{-}S$) and blue ($\#MSU\text{-}F$) line, with 95 % confidence intervals.

Figure 4.4: Comparison of $\#MSU\text{-}S$ (green line) and $\#MSU\text{-}F$ (blue line).

$\#MSU\text{-}F$, as well as the difference between both; and fig 4.4b does the same for R^2. Similar to the plots presented in the previous analysis, each figure begins with $N = 1,000$, which is then incremented in steps of 1,000 until N is equal to the number of all available MSUs. For most N, the coefficient for $\#MSU\text{-}S$ is more strongly negative than the coefficient for $\#MSU\text{-}F$. Thus, the more MSUs contain a given word, the earlier that word is first produced, and this predictive relationship is stronger for $\#MSU\text{-}S$ than for $\#MSU\text{-}F$. We find a similar pattern for R^2: Across most N, $\#MSU\text{-}S$ can explain a larger amount of variance in AoFP than



(a) Bottom: regression coefficients (β), with 95 % confidence intervals. Top: difference between green ($\#MSU\text{-}S$) and red ($\#MSU\text{-}P$) line, with 95 % confidence intervals.

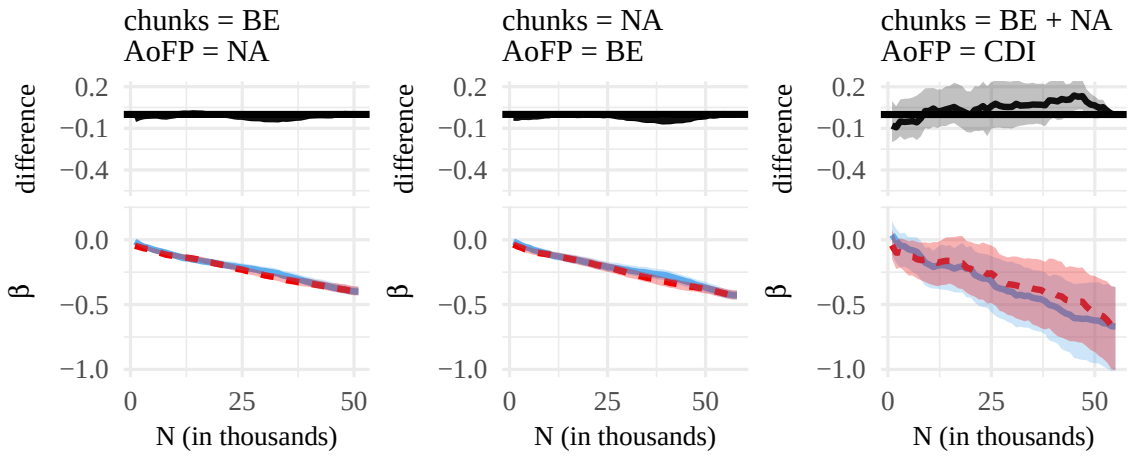


(b) Bottom: amount of variance in AoFP (ΔR^2 in %) that can be explained by $\#MSU$, with 95 % confidence intervals. Top: difference between green ($\#MSU\text{-}S$) and red ($\#MSU\text{-}P$) line, with 95 % confidence intervals.

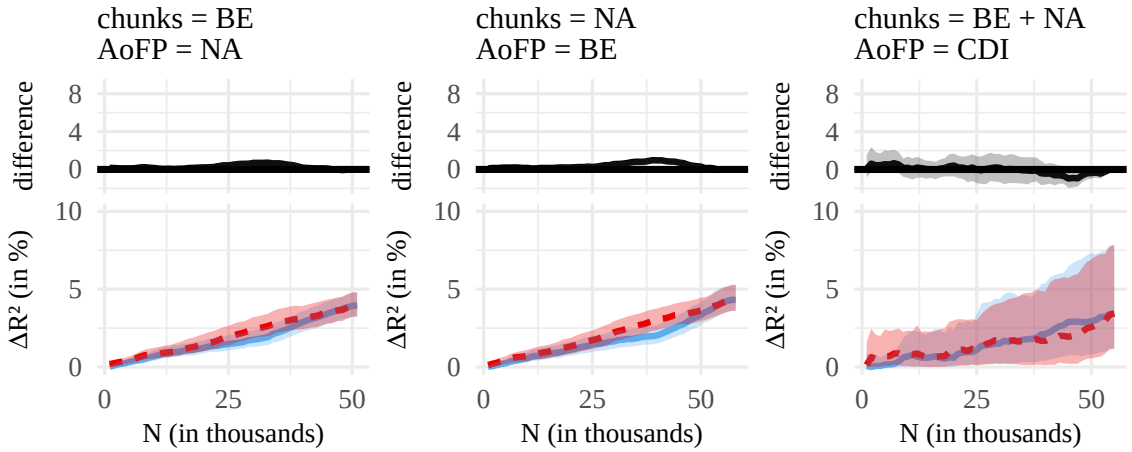
Figure 4.5: Comparison of $\#MSU\text{-}S$ (green line) and $\#MSU\text{-}P$ (red line).

$\#MSU\text{-}F$. We can state, then, that $\#MSU\text{-}S$ performs better at predicting AoFP.

This pattern is similar across all three pairings of corpus and AoFP data, although the confidence intervals are much larger when using CDI-derived AoFP. This is probably due to the smaller number of data points: The regression models with CDI estimates are based on 615 words, while the regressions with corpus-derived estimates include approximately ten times the number of words. As a result, we operate with less statistical power when conducting analyses with the CDI-derived estimates, and the differences between β / R^2 do not always reach statistical signif-



(a) Bottom: regression coefficients (β), with 95 % confidence intervals. Top: difference between red ($\#MSU\text{-}P$) and blue ($\#MSU\text{-}F$) line, with 95 % confidence intervals.



(b) Bottom: amount of variance in AoFP (ΔR^2 in %) that can be explained by $\#MSU$, with 95 % confidence intervals. Top: difference between red ($\#MSU\text{-}P$) and blue ($\#MSU\text{-}F$) line, with 95 % confidence intervals.

Figure 4.6: Comparison of $\#MSU\text{-}P$ (red line) and $\#MSU\text{-}F$ (blue line).

icance. The overall pattern, however, is similar across the different AoFP estimates – indicating that $\#MSU\text{-}S$ is indeed better-suited for predicting AoFP.

The only choices of N for which this is not true are (a) very small values and (b) values close to the largest possible value. Generally, β and R^2 take near-zero values at $N = 1,000$. This is because at 1,000 MSUs, we can only derive $\#MSU$ counts for a relatively restricted number of target words. But as we increase N, $\#MSU\text{-}S$ and $\#MSU\text{-}F$ begin to perform better. R^2 increases, and the coefficients associated with the two predictors now take negative values. Crucially, regression models with

$\#MSU\text{-}S$ outperform their counterparts with $\#MSU\text{-}F$.

At some point, the difference starts to decrease, until it disappears once N is equal to the number of all MSUs. This makes sense: If the two chunk sets contain all MSUs, $\#MSU\text{-}S$ and $\#MSU\text{-}F$ are calculated from the same selection of MSUs, and the two estimates will take the same value. A larger N means that the two chunk sets from which we calculate $\#MSU\text{-}S$ and $\#MSU\text{-}F$ overlap more and more, and the two estimates begin to converge. Thus, past a certain point, the differences in R^2 and β decrease.

We thus have good reason to claim that $\#MSU\text{-}S$ is better suited for predicting the time course of word learning than $\#MSU\text{-}F$. Figure 4.5 shows, moreover, that $\#MSU\text{-}S$ *also* outperforms $\#MSU\text{-}P$, with a pattern that is very similar to the one obtained in the previous comparison. At the same time, almost no significant difference emerges when comparing $\#MSU\text{-}P$ and $\#MSU\text{-}F$ (figure 4.6). Together, the three comparisons suggest that there is no (strong) difference in the effects obtained with $\#MSU\text{-}P$ and $\#MSU\text{-}F$, while $\#MSU\text{-}S$ performs consistently better at predicting AoFP than the other two $\#MSU$ counts. The implication for speech segmentation is that children are more likely to store short rather than frequent or internally predictable MSUs.

4.6 General discussion

In this chapter, we set out to determine whether children are more likely to extract and store (1) frequent, (2) short, or (3) internally predictable chunks during the segmentation process. In one of two analyses, we found that selections of short MSUs tend to contain more single-word utterances than selections of frequent or predictable MSUs, suggesting that sequence length is more useful cue to *wordhood* than the other two predictors. In a second analysis, we also found that short MSUs perform better at predicting the time course of word learning – suggesting that children store short rather than frequent or internally predictable syllable chunks. Together, the two findings imply that children tend to store short, word-like chunks, not frequent or internally predictable chunks.

Frequency plays an important role in first language acquisition, as it facilitates

learning at many different levels (Ambridge et al., 2015). But when it comes to the extraction of an initial repertoire of chunks from unsegmented input, children probably store short instead of frequent speech sequences. In this respect, sequence length also appears to be more important than syllabic predictability: When predicting the time course of word learning, we found no difference between predictable and frequent MSUs, which are both clearly outperformed by short MSUs. Since short MSUs are also the most word-like, this can be interpreted as evidence for a segmentation bias towards discrete or indivisible linguistic units – i.e., word-like sequences that cannot themselves be segmented into smaller units.

Note that we cannot rule out that children *also* store particularly frequent or internally predictable chunks. After all, there is some overlap between the shortest, most frequent, and most predictable MSUs (cf. analysis I); and involvement in frequent or predictable MSUs still predicts when words are learned, albeit to a lesser degree than involvement in short MSUs. Our results suggest, however, that short MSUs are comparatively more likely to be stored as chunks than either frequent or predictable MSUs. More generally, our findings imply that the length of syllable sequences is a more useful segmentation cue than syllabic predictability or whole-sequence frequency. Based on this, we predict a stronger causal role of sequence length during word segmentation. A possible direction for future work is to test this prediction experimentally, e.g. in an artificial word segmentation task.

Our results also have implications for research concerned with frequent multi-word sequences, which are sometimes referred to as *formulaic sequences*. Various studies have demonstrated that both adults (Arnon and Snider, 2010; Arnon and Priva, 2014) and children who have completed the segmentation process (Bannard and Matthews, 2008; Arnon and Clark, 2011) are faster to process formulaic multi-word phrases, and that this processing advantage cannot be reduced to the frequency of individual words. Such results suggest that language users represent some aspect(s) of frequent word sequences – above and beyond information about constituent words.

Since the subjects in these studies had completed the segmentation process, it is unlikely that they process multi-word phrases in a holistic fashion, without accessing component words. Indeed, other studies have collected evidence that access (in adult processing) to frequent trigrams (Arnon and Priva, 2014), to idioms (Sprenger et al., 2006), and to frequent adjective-noun and noun-noun phrases (Jacobs et al.,

2016) involves access to individual words. Post-segmentation, that is, language users appear to possess analyzed representations of multi-word phrases. This naturally leads to the question whether holistically stored chunks are retained past the segmentation stage as fully analyzed representations, or whether chunks are discarded once the segmentation process is completed. In the latter case, chunks and representations of frequent phrases would result from two different processes. One would be related to segmentation and involve the storage of larger units that are gradually analyzed, and the other would discover phrases through usage patterns within fully segmented input (Arnon and Christiansen, 2017).

The results presented in this study imply that children preferentially store word-like sequences as undersegmented chunks – which tend to be short, not frequent. As mentioned, we cannot rule out that children *also* store (some) frequent undersegmented chunks, in addition to short sequences. However, our results do suggest that children are comparatively more likely to store short, word-like sequences. This, in turn, supports accounts wherein representations for frequent multi-word sequences tend to emerge *after* the segmentation process has run its course. Arguing from the current results, in other words, we suggest that a sizable portion of cognitive representations for formulaic multi-word sequences cannot be traced back to undersegmented chunks in children.

This page is intentionally left blank.

Chapter 5

Using neural networks to model speech processing with cochlear implants

We introduce a novel machine learning approach for investigating speech processing with cochlear implants (CIs) – prostheses used to replace a damaged inner ear. Concretely, we use a simple perceptron and a deep convolutional network to classify speech spectrograms that are modified to approximate CI-delivered speech. Implant-delivered signals suffer from reduced spectral resolution, chiefly due to a small number of frequency channels and a phenomenon called channel interaction. The latter involves the spread of information from neighboring channels to similar populations of neurons and can be modeled by linearly combining adjacent channels. We find that early during training, this input modification degrades performance if the networks are first pre-trained on high-resolution speech – with a larger number of channels, and without added channel interaction. This suggests that the spectral degradation caused by channel interaction alters the signal to conflict with perceptual expectations acquired from high-resolution speech. We thus predict that a reduction of channel interaction will accelerate learning in CI users who are implanted after having adapted to high-resolution speech during normal hearing.

5.1 Introduction

Cochlear implants (CIs) are neural prostheses that can partially restore hearing to individuals with sensorineural hearing loss. This type of hearing loss results from damage to sensory cells within the cochlea, an inner-ear organ that converts pressure waves into nerve impulses. CIs rely on an external microphone to capture sound signals from the environment and filter them into frequency bands whose amplitude envelopes modulate digital pulse sequences. These are then used to excite neurons through appropriately placed electrodes within the cochlea, where designated locations are responsible for processing specific frequency ranges.

The CI-delivered signal suffers from reduced spectral resolution, caused by (1) a limited number of electrodes; and (2) channel interaction (Friesen et al., 2001; Jones et al., 2013), which results from a group of neurons being stimulated by more than one electrode. Partly because of this, certain speech perception tasks present difficulties for CI users. For example, CIs make it harder to detect vocal emotion (Chatterjee et al., 2015, 2018), perceive tones (Mao and Xu, 2017), or to recognize speech in the presence of competing talkers (Cullington and Zeng, 2008; Stickney et al., 2004).

Recent work attempts to explain such task-specific differences in terms of the acoustic cues attended to by CI users and normally hearing (NH) subjects. For example, Gaudrain and Baskent (2018) found that the former have difficulties in telling apart speech which differs in fundamental frequency or vocal tract length of the speakers – cues which NH listeners rely on to distinguish pitch (fundamental frequency) and speaker height (vocal tract length). In a similar vein, Moberly et al. (2014) measured sensitivity to cues involved in the detection of phonemic contrasts. They report that CI users rely on coarse-grained cues related to the overall amplitude of the signal, rather than fine-grained differences between formants. NH subjects, in contrast, attend to both types of cues.

The performance differences between NH listeners and CI users, then, follow from differences in how the two populations process speech. And it stands to reason that differences in processing are, to some extent, a result of the input received by the learners. In other words, the fact that CI-delivered speech is characterized by a coarse spectral resolution, rendering more fine-grained features inaccessible,

could conceivably push CI users to attend to coarse-grained acoustic cues; while NH individuals attend to coarse- *and* fine-grained features because the intact cochlea happens to transmits a greater level of spectral detail.

This, in turn, suggests that postlingually deaf CI users (PD-CI users), who receive CIs after a period of normal hearing, need to change the manner in which they process speech following implantation. This transition is likely to require neural re-wiring – which is presumably why the speech recognition performance of PD-CI users improves gradually, for well over a year, following implantation (Oh et al., 2003). In the current study, we train neural networks under conditions that mimic those of PD-CI users, with initial exposure to high-resolution speech (corresponding to a period of normal hearing), followed by exposure to low-resolution speech (corresponding to the period after implantation). In doing so, we examine the effect of channel interaction on the transition from high- to low-resolution speech.

By *channel interaction*, we mean the (partial) summation of electrical fields generated by neighboring electrodes, leading to a distortion of amplitude envelopes for frequency channels assigned to specific electrodes, prior to neural activation (Shannon, 1983). This is an important cause of variability in speech recognition outcomes among CI users, with higher levels of channel interaction being associated with poor performance (Stickney et al., 2006).

The mechanisms via which performance is affected clearly have to do with spectral resolution: In a task designed to measure the level of spectral detail utilized by subjects (spectral-ripple discrimination), higher levels of channel interaction were associated with poor outcomes (Jones et al., 2013). Moreover, increasing the number of available electrodes past eight does not increase performance (Friesen et al., 2001), suggesting that channel interaction makes adjacent electrodes less distinguishable. Thus, by decreasing spectral detail in the implant-delivered signal, channel interaction seems to inherently limit speech recognition performance.

Given this, one should naturally expect better performance if channel interaction was reduced. Here, we surmise that a reduction in channel interaction should *also* lead to faster learning in PD-CI users. Our basic argument is that post-implantation, PD-CI users may have to change the manner in which they process speech in order to accommodate the decreased spectral resolution in the implant-delivered signal. If channel interaction were eliminated, the signal would contain more detail, and

PD-CI users would presumably have to make fewer adjustments to transition to speech processing with the implant.

To explore this, we train deep neural networks under conditions modeled on those faced by PD-CI users and CI users who are born deaf – and thus learn to process CI-delivered signals without first having adapted to an intact cochlea. In the following section, we specify our objective and method more closely. In section 3, we provide more details about the speech recognition tasks, data, and pre-processing steps. Finally, we present the results in sections 4 – 6, ending with a general discussion in section 7.

5.2 Research question and general method

Our research question can be phrased as follows: Does channel interaction slow learning during the period after implantation in PD-CI users, compared to congenitally deaf CI users (CD-CI users)?

In contrast to PD subjects, CD-CI users are born deaf – and, if implanted early enough, develop good speech recognition abilities (Harrison et al., 2005). Crucially, CD-CI users are only exposed to the degraded implant-delivered signal, while PD-CI users first learn to process speech delivered through the intact cochlea (while normally hearing) and then adapt to the CI after implantation. During the adaptation process, PD-CI users presumably integrate novel aspects of the CI-delivered signal – including the mode of neural stimulation (electrical current) and a reduction in spectral resolution. In this chapter, we focus on spectral degradation caused by channel interaction, and we hypothesize that it slows adaptation to CIs in PD-CI users, with a comparatively smaller impact on learning in CD-CI users.

This effect could result from PD-CI users changing auditory processing strategies in order to transition from high-resolution input (delivered through the intact cochlea) to the spectrally degraded CI-delivered signal – for example, to emphasize coarse- rather than fine-grained spectral cues. Such a change in processing strategies is likely to require a certain amount of time. However, if the implant-delivered signal was less-coarse grained and thus more similar to high-resolution input, less time should be required, as that should reduce the amount of adaptation required on the part

of PD-CI users. Thus, if we increase spectral resolution by reducing or removing channel interaction from the implant, we should expect faster adaptation to CIs.

Importantly, we expect this effect to be absent in CD-CI users: Since this population never learns to process speech delivered through the intact cochlea, they should be able to adapt to a degraded spectral resolution without having to modify existing processing strategies. Thus, while channel interaction should equally degrade maximum performance in PD- and CD-CI users *after* both populations have adapted to the implant, its impact on the learning process should be stronger in PD-CI users.

To gather evidence for the hypothesized effect, we train neural networks on (1) high-resolution (high-res) spectrograms (X_h) with a large number of channels, intended as an approximation of what the intact cochlea delivers to the brain; (2) low-resolution (low-res) spectrograms, with a smaller number of channels suffering from channel interaction (X_l), to approximate CI-delivered input; and (3) medium-resolution (med-res) spectrograms – essentially low-res spectrograms *without* channel interaction (X_m), to approximate CI-delivered input if channel interaction was eliminated from the implants. We derive X_h by computing amplitude spectrograms with a large number of channels, X_m by constructing spectrograms with fewer channels, and X_l by linearly combining neighboring channels in X_m.

We conduct experiments in a postlingually deaf (PD) condition, where we train first on X_h, followed by further training on either X_l or X_m; and a congenitally deaf (CD) condition, where we train only on X_l or X_m. Networks trained on high-res speech should generalize poorly to X_l and perform better on X_m. Given sufficient training on X_l and X_m, CD and PD networks might eventually perform similarly on both input types. But at early epochs, before the PD networks have adapted to the decreased spectral resolution, we expect a larger performance difference in the PD condition.

This outcome would show that the spectral degradation introduced through channel interaction slows learning in deep neural networks as they adapt to low-res speech – *after they were pre-trained on high-res speech.* Assuming that deep learning is a reasonable model for pattern recognition in the brain, it would also suggest that decreased spectral resolution prevents PD-CI users from quickly adapting to CIs. Caution is surely required in making the connection to human processing; but since deep neural networks have proven useful for understanding brain-based sensory pat-

tern recognition (Yamins and DiCarlo, 2016), we consider them as an exploratory tool for investigating learning dynamics with CIs.

5.3 Materials and methods

Tasks and Data We train neural networks on gender and isolated word recognition, choosing the former because it presents difficulties for CI users (Gaudrain and Başkent, 2015; Gaudrain and Baskent, 2018) and the latter because CI users can perform it with high accuracy (Rouger et al., 2007). If similar patterns appear across both tasks, despite the different performance patterns with CIs, we can be relatively confident in their robustness.

We frame gender recognition as binary classification of utterances into *male* or *female*, based on data from the Texas Instruments Massachusetts Institute of Technology (TIMIT) corpus (Garofolo et al., 1993). TIMIT contains recordings of 630 speakers (70 % male, 30 % female) from 8 U.S. dialect regions. Each speaker read 10 different sentences, yielding 6,300 utterances spanning several seconds each (5.4 hours of speech).

The isolated word recognition task involves the classification of utterances into 30 word-classes, with data from the Google Speech Commands (GSC) corpus[1]. Collected via crowd-sourcing, the GSC contains 65,000 one-second utterances of 30 short words (18 hours of speech). 20 core words were pronounced five times by most speakers, and an additional 10 words (treated as *unknown words*) were pronounced once. For each corpus, we use a randomly selected 20 % of the data for validation, and another 20 % for testing. All WAV files have a sampling rate of 16,000.

Featurization We train neural networks on input that is approximately similar to what the brain uses for speech recognition. Generally, the auditory system processes nerve impulses distributed across time and frequency – with the decomposition into distinct frequency components implemented mechanically by the inner ear (NH subjects), or digitally by the implant (CI users). Modern CIs work with 12 – 22 electrodes, and signals are decomposed into as many frequency components.

[1] `https://research.googleblog.com/2017/08/launching-speech-commands-dataset.html` (data were downloaded on 08/08/2018)

The intact human cochlea, in contrast, contains thousands of sensory hair cells, where topographically coherent cell groups correspond to functional channels. The bandwidths of these cochlear channels have been investigated in various behavioral experiments, and the methods used continue to evolve (Shera et al., 2002). In auditory models, 30 channels are often used to cover the frequency range relevant for speech (Coath and Denham, 2007; McDermott and Simoncelli, 2011).

In this study, we approximate speech delivered through CIs and the intact cochlea via mel-scaled amplitude spectrograms (computed over windows of 50ms, strided by 10ms). Each spectrogram is a matrix with dimensionality $N \times T$, where N is the number of channels and T is the number of frames, with $x_{n,t}$ being the amplitude at time t and channel n. The channels, whose spacing and bandwidths conform to the perceptually motivated mel scale, cover the range between 200 and 7,000 Hz, similar to most CIs. A medium-resolution (med-res) condition with $N = 16$ channels approximates CI-delivered input in the hypothetical case that channel interaction was completely eliminated; and a high-resolution (high-res) condition with $N = 32$ channels approximates signals delivered through the intact cochlea.

We thus feed the networks with mel-scaled amplitude spectrograms, covering a frequency range similar to most CIs, with a number of channels similar to implants (med-res) or to the intact cochlea (high-res). In order to obtain low-resolution (low-res) input that approximates the signals transmitted through CIs, however, we still need to operationalize channel interaction – which can be approximated as a summation of potentials from individual electrodes Tang et al. (2011). Thus, given row $x_{n,*}$ in a med-res spectrogram x with N frequency bands, we obtain a low-res spectrogram with added channel interaction by linearly combining rows:

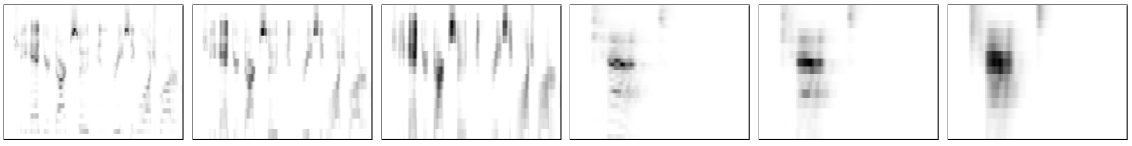
$$x_{n,*} = \begin{cases} x_{n,*} + x_{n+1,*}, & \text{if } n = 1 \\ x_{n,*} + x_{n-1,*}, & \text{if } n = N \\ x_{n,*} + x_{n+1,*} + x_{n-1,*}, & \text{otherwise} \end{cases}$$

To ensure that data points from all three input conditions are represented as a $32 \times T$ matrices, we duplicate each row in the low- and med-res spectrograms. Equally sized spectrograms are (1) necessary to train the same neural network on high- and either low- or med-res input and (2) roughly analogous to the contrast between the intact cochlea and CIs. With the latter, electrodes cover broad swaths of neurons; while

in the intact cochlea, groups of sensory hair cells cover fewer neurons. Similarly, in the low- and med-res spectrograms, broad areas contain information from a single channel; and in the high-res spectrograms, smaller areas contain information from more narrow channels.

Note that our spectrograms approximate rather than simulate CIs. For example, the implants use amplitude envelopes to modulate digital pulse trains, whereas we use raw amplitude spectrograms. Due to the exploratory nature of this study, however, we abstract from additional complexity. See figure 5.1 for example spectrograms.

Figure 5.1: Example spectrograms – in high-res (32 channels), med-res (16 channels), and low-res (16 linearly combined channels).



high-res med-res low-res high-res med-res low-res

(a) Utterance from the TIMIT corpus. (b) Word from the GS corpus.

Network Architectures One of the simplest neural networks available is the perceptron (PER). For this model, the probability that an input vector x is a member of class i (belonging to a stochastic variable Y, with two values for gender recognition and thirty values for word recognition) can be defined as:

$$p(Y = i|x, \theta) = \frac{e^{W_i \cdot x + b_i}}{\sum_j e^{W_j \cdot x + b_j}} \tag{5.1}$$

where the set of trainable parameters θ contains the weight matrix W as well as the bias b, and each input x is a spectrogram, converted from the original matrix into a vector format. This is done by concatenating the rows (= frequency bands) in a given spectrogram, with dimensionality $N \times T$, to create an input vector with length NT. Due to its simplicity, the PER is an appealing model choice, but it also comes with disadvantages: (a) It is a flat architecture, limiting performance; and (b) each column in the weight matrix W and bias b corresponds to exactly one value in x, so that the patterns detected by the model are bound to particular input coordinates.

To obtain better performance, and to model the invariance of auditory processing to spectro-temporal details (Bizley and Cohen, 2013), we also report results with a deep convolutional neural network (CNN), trained to ingest spectrograms in matrix format. The first three layers contain 2D convolutions – which facilitate the detection of local patterns at different positions (LeCun et al., 1995) by repeatedly applying sets of weights (filters) to $n \times m$ (filter size) sub-regions within the input, strided by $u \times v$ (stride). The convolutional layers are then followed by a fully connected layer, whose output is fed into a perceptron as in (1). For speedier convergence, we apply the Batch Normalization method (Ioffe and Szegedy, 2015) after each hidden layer.

Since we compare performance on different featurizations of the same data (high-res, med-res, low-res), a selection of hyperparameters meant to optimize performance on any of the three featurizations could confound the results. To address this, we chose hyperparameters so that the validation error improves steadily, without noticeable fluctuations – but we do not tweak them for maximum performance. Since the CNN has a fairly large number of tunable hyperparameters, it is still possible that we accidentally picked settings which favor a specific featurization. For the PER, however, the only hyperparameters are the learning rate (0.01) and mini batch size (32), dramatically reducing this risk.

Apart from the number of hidden layers and the application of batch normalization, hyperparameters for the CNN include the activation function (rectified linear function), regularization (0.1 dropout at each convolutional layer, 0.5 at the fully connected layer), number of filters in each convolutional layer (5), filter size (5×5), stride (2×2), number of hidden units in the connected layer (100), learning rate (0.1), and mini batch size (32).

The models are trained via mini batch gradient descent, on an Nvidia Titan X GPU, by minimizing the categorical cross-entropy of predicted and true class probabilities. The learning rate is adjusted via Adadelta (Zeiler, 2012), and training is terminated once the validation error has ceased to decrease for 10 epochs (early stopping with a patience of 10).

Statistical Testing The key statistic of the current study is the difference in performance between neural networks trained on different featurizations of the same corpus. For example, given the TIMIT corpus, let X_h be a high-res featurization, and let X_l be a low-res featurization. We might train a network A on X_h and a

network B on X_l, obtaining accuracy scores t_A, t_B by evaluating A and B on held-out portions of X_h and X_l, respectively. The question of interest then is whether we can reject the null hypothesis that $t_A = t_B$.

We can answer this via approximate randomization testing (ART), a simple approach that does not rely on assumptions about the data and is thus well-suited for application in machine learning (Noreen, 1989; Yeh, 2000). ART starts from the labels $C_A = \{c_A^1, ..., c_A^n\}$ and $C_B = \{c_B^1, ..., c_B^n\}$ assigned by the two networks to each data point in the held-out data. Each pair of labels c_A^i, c_B^i is then switched with probability $\frac{1}{2}$, and the difference in performance d' is re-calculated. The procedure is repeated R times, with r being the number of times that $d' \geq d$. For large R, $p = \frac{r+1}{R+1}$ approximates the significance level. We set $R = 10^5$.

Given N comparisons, we reject the null hypothesis if $p \leq \frac{0.05}{N}$. That is, we apply the Bonferroni method to correct for multiple comparisons, since these increase the chance of incorrectly rejecting the null hypothesis. In the analyses below, we conduct 18 model comparisons per corpus, obtaining a rectified significance threshold of $p \leq \frac{0.05}{18}$.

5.4 Analysis I: Preliminary comparisons

In this first analysis, we compare (1) the PER to the CNN and (2) performance on high-res spectrograms (32 channels) to performance on low-res spectrograms (16 channels with channel interaction). This serves as a sanity check: Due to the larger number of parameters and the location-independent, more generalizable features detected by the CNN, we expect worse performance with the PER; and due to the diminished level of spectral detail in the low-res featurization, we expect better performance on high-res input. Given our two network architectures (PER, CNN), the two featurizations (high-res, low-res), and the two tasks (gender, word recognition), we report results for $2 \times 2 \times 2 = 8$ models.

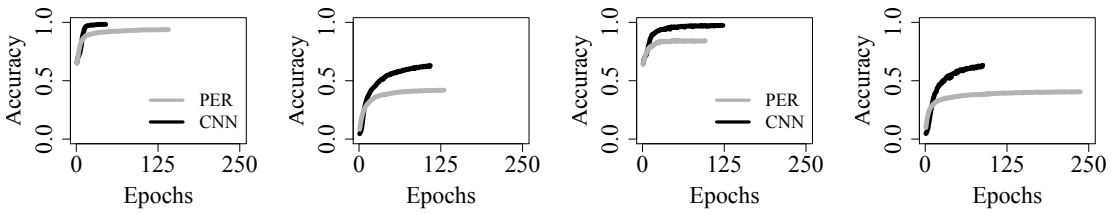
5.4.1 Results and discussion

Figure 5.2 shows validation accuracy over training epochs. It is immediately apparent that the CNN outperforms the PER. Indeed, test accuracy achieved with

the former is significantly higher: by ca, 10 – 15 % for gender recognition, and by about 25 % for word recognition ($p \leq 0.001$). The larger performance gap on word recognition is likely due to a higher degree of spectro-temporal variability in the data used for this task, so that the CNN gains a comparatively stronger advantage.

Turning to the distinction between low- and high-res input (table 5.1), we find that high-res spectrograms lead to better performance in three out of four model comparisons – as is expected, given the spectrally impoverished nature of the low-res spectrograms. On gender recognition, the difference obtained with the CNN is not statistically significant; and the difference obtained with the PER is less strong than the corresponding difference on the word recognition task. This indicates that loss of spectral detail is comparatively more detrimental for gender and less so for word recognition. These results make sense in light of performance patterns with CI users, who can solve isolated word recognition with high accuracy (Rouger et al., 2007) but struggle with gender recognition (Gaudrain and Baskent, 2018) – suggesting that the former is more easily solvable with the spectrally impoverished CI-delivered signal.

Figure 5.2: Validation accuracy over training epochs, for networks trained on high- or low-res spectrograms.



gender recognition word recognition gender recognition word recognition

(a) high-res (b) low-res

5.5 Analysis II: Effect of channel interaction

We next compare the performance of networks trained only on med-res spectrograms (16 channels *without* channel interaction) or low-res spectrograms (16 channels *with* channel interaction). These training regimes are idealized analogs of the conditions faced by CD-CI users – equipped with hypothesized CIs that completely eliminate

Table 5.1: Test accuracy, for networks trained on high- or low-res speech.

Task	Model	High-Res	Low-Res	Diff
gender	PER	93.1	84.3	**8.8** * * *
	CNN	98.9	96.9	**2.0** **
words	PER	43.1	41.1	**2.0** * * *
	CNN	63.1	63.5	-0.4

Diff = high-res accuracy minus low-res accuracy. Here and in the following table, *: $p \leq 0.05$; **: $p \leq 0.01$; * * *: $p \leq 0.001$ (significance thresholds are Bonferroni-corrected).

channel interaction (med-res spectrograms); or with CIs suffering from channel interaction, similar to the implants currently in use (low-res spectrograms).

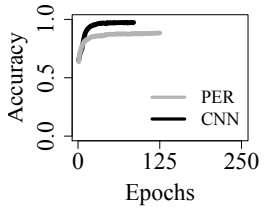
Given that channel interaction in CIs is associated with poor speech recognition performance (Jones et al., 2013), we should likewise expect our operationalization of channel interaction to limit speech recognition performance in neural networks.

5.5.1 Results and discussion

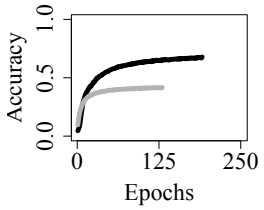
Figure 5.3 shows validation accuracy over epochs, for networks trained on med-res data. Although the learning trajectories appear broadly similar to those obtained on low-res speech (figure 5.2b), final test accuracy is significantly higher in the med-res condition, in three out of the four model comparisons: by 4.2 % (PER, $p \leq 0.001$) on gender recognition, as well as by 1.5 % (PER, $p \leq 0.001$) and 0.9 % (CNN, $p \leq 0.05$) on word recognition. For the CNN, no significant difference emerged when trained on gender recognition.

By reducing spectral detail, then, our operationalization of channel interaction limits accuracy. The observed performance degradations are generally weaker than the ones from the previous analysis, where we compared high- to low-res spectrograms. This is expected, since the difference (in spectral resolution) between the med- and low-res data is less strong than between high- and low-res input.

Figure 5.3: Validation accuracy over training epochs, for networks trained on med-res spectrograms.



gender recognition



word recognition

5.6 Analysis III: Effect of channel interaction with pre-training

In the foregoing analyses, we started training with randomly initialized parameters (weights and biases). Now, we use parameters that are pre-trained on high-res input (see figure 5.2a), and we ask how well they generalize to either (1) low-res or (2) med-res spectrograms. (1) mimics the learning conditions of PD-CI users, who transition to the CI-delivered signal after having adapted to high-res speech during a period of normal hearing; and (2) mimics a hypothetical case where PD-CI users, having adapted to high-res input, transition to CIs that do not suffer from channel interaction.

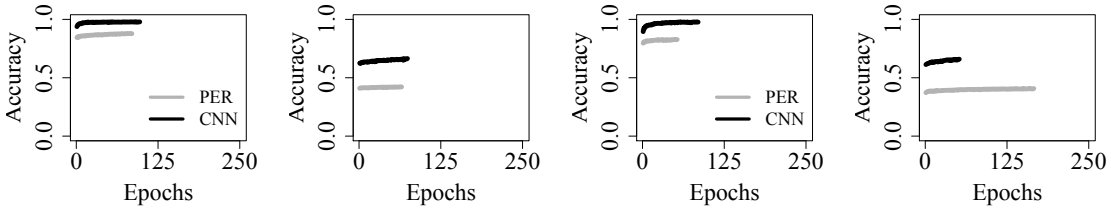
As before, the networks come in two variants, applied to two different tasks. They are then trained on low-res or med-res spectrograms, yielding eight pre-trained models.

5.6.1 Results and discussion

Figure 5.4 shows validation accuracy, on low- and med-res data, for models whose weights were pre-trained on high-res spectrograms. Notice how initial validation accuracy is already very high, compared to networks that were not pre-trained (low-res: figure 5.2b; med-res: figure 5.3).

Without pre-training, the weights are randomly initialized, and accuracy is at chance level before training commences (table 5.2, left side, bottom). The parameters acquired from high-res input, however, generalize to both the med- and low-res data,

Figure 5.4: Validation accuracy over training epochs, on med- and low-res spectrograms, for models that were pre-trained on high-res spectrograms.



gender recognition word recognition gender recognition word recognition

(a) med-res (b) low-res

affording high initial performance (left side, top). Table 5.2 shows, moreover, that the pre-trained parameters generalize better to the med-res data, with significantly lower initial accuracy on low-res input – both before training commences, as well as early during learning (after the first epoch). Once training has finished, the pre-trained networks still generally perform worse on low-res input, but the differences are not as strong as early during training.

The opposite pattern emerges for the randomly initialized models: Initially, these models make random guesses, regardless of input conditions. But after training has run its course, significant differences emerge. For the randomly initialized models, in other words, differences between med- and low-res input emerge and *increase* over time, whereas they are present from the beginning and *decrease* over time if the parameters are pre-trained on high-res spectrograms.

This can be explained by differences in spectral resolution between the input conditions. The pre-trained weights, being optimized for processing high-res spectrograms, generalize better to med- than to low-res spectrograms because the resolution of the former is closer to high-res input. After training, the strong initial differences between the med- and low-res conditions decrease, suggesting that the pre-trained networks compensate, to some extent, for the reduced resolution. But early during training, when they process low-res speech as if it was high-res speech, the stronger spectral degradation severely degrades performance.

The degraded spectral resolution in the low-res data is a direct result of channel interaction (which is absent in the med-res spectrograms). Thus, not only does our

Table 5.2: Test accuracy, for pre-trained (top) and randomly initialized (bottom) networks: before training on med- or low-res spectrograms (left), after the first epoch (middle), and after the final epoch (right).

Init.	Task	Model	med-r.	low-r.	Diff	med-r.	low-r.	Diff	med-r.	low-r.	Diff
pre-tr.	gender	PER	83.0	75.3	**7.7** * * *	85.1	79.4	**5.7** * * *	87.3	83.1	**4.2** * * *
		CNN	74.6	73.1	1.5	93.5	90.0	**3.5** * * *	97.6	97.3	0.3
	words	PER	41.0	36.5	**4.5** * * *	41.5	38.0	**3.5** * * *	42.6	41.1	**1.5** * * *
		CNN	62.7	58.4	**4.3** * * *	62.6	62.6	**1.0** * * *	66.7	65.8	**0.9** *
rand.	gender	PER	50.3	51.1	-0.8	63.4	63.0	0.4	87.1	84.3	**2.8** **
		CNN	59.2	59.2	0.0	65.8	63.4	2.4	96.2	96.9	-0.7
	words	PER	2.9	2.6	0.3	9.7	9.7	0.0	42.4	41.1	**1.3** *
		CNN	3.3	3.5	-0.2	5.1	5.0	0.1	67.0	63.5	**3.5** * * *

Diff = accuracy on med-res spectrograms minus accuracy on low-res spectrograms.

operationalization of channel interaction lead to reduced accuracy *after* training, but it also leads to slower initial performance gains for the pre-trained networks. These results imply that channel interaction in CIs should not only limit speech recognition performance after CI users have fully adapted to the implants, but that it should *also* slow learning in PD-CI users during the transition period after implantation.

5.7 Conclusions

In an effort to investigate the impact of channel interaction on learning in CI users, we trained neural networks on two types of spectrograms, intended to approximate CIs with and without channel interaction. We generally obtained poorer performance on the former – in spite of training the networks for as long as is necessary for performance to plateau. This suggests that channel interaction leads to the irrecoverable loss of crucial spectral detail, corroborating previous findings that CI users perform worse in the presence of channel interaction due to a decrease in spectral resolution (Friesen et al., 2001; Jones et al., 2013).

Apart from a performance degradation after training, we also found a negative impact of channel interaction on the early performance of networks pre-trained on spectrograms intended to approximate the speech delivered through the intact cochlea (high-res input). We observed that the effect was absent in models that were not pre-trained; and that the pre-trained networks recovered with additional training, until they performed similarly to their randomly initialized counterparts. This second effect, then, arises only if the models have adapted to high-res input, and it is reversible over time.

The implication for CIs is that spectral degradation caused by channel interaction should slow learning in PD- but not in CD-CI users. Prior to implantation, only the former adapt to normal hearing; and this might force them to unlearn certain processing strategies that may be applicable to normal hearing – but that do not generalize to CIs due to the impoverished nature of the implant-delivered signal. For example, PD-CI users might need to re-learn which acoustic features they attend to, with more emphasis on coarse- rather than fine-grained features after implantation. If spectral resolution is increased, there should be less need for such adaptation. Consequently, we predict that techniques for the reduction of channel interaction

(Landsberger and Srinivasan, 2009) will accelerate speech recognition improvement in PD-CI users.

More generally, our study demonstrates how machine learning can be used to shed light on questions in the field of CI research. Our approach allows us to quickly evaluate the impact of input modifications on auditory pattern recognition, without the difficulties involved in conducting behavioral studies (e.g. time constraints, ethical considerations). In similar machine learning experiments, other input properties could be examined. For example, one could model the electrical pulse trains generated by CIs and investigate how the pulsatile nature of the signal affects processing.

This page is intentionally left blank.

Chapter 6

Conclusion

In the preceding chapters, we examined bootstrapping operations in language development, defined as processes wherein learners use resources or information from one domain in order to (partially) solve a task from a different domain.

At the outset of this thesis, we noted that bootstrapping processes are prevalent in language development. For example, in order to group words into lexical categories, children pay attention to the phonological properties of those words (Durieux and Gillis, 2001; Fitneva et al., 2009) and rely on statistical regularities between co-occuring context words (Gerken et al., 2005; Mintz, 2003, 2006). Thus, children can use knowledge from different domains – word-level phonology and the distributional patterning of words – in order to break into lexical category acquisition. In the literature, this is often referred to as *phonological* and *distributional bootstrapping*, respectively. Other documented examples of bootstrapping processes include *syntactic bootstrapping*, where the manner in which verbs combine with their arguments is used to constrain possible action-verb mappings (Gleitman, 1990; Fisher et al., 2010); and *inflectional bootstrapping*, where word-referent mappings are constrained by inflectional morphology (Jolly and Plunkett, 2008).

Such synergistic interdependencies most likely arise from pervasive *neural reuse*: When new cognitive functions develop, the brain is unlikely to develop neural circuitry from scratch, but instead appears to recruit existing circuits whenever possible. This "fundamental organizational principle of the brain" (Anderson, 2010) implies that bootstrapping processes, resulting from the constant repurposing of

neural circuitry, are common in every area of cognitive development. Research on language development and work on neural reuse offer complementary perspectives on bootstrapping – situated at the the level of psychological mechanisms and the level of neural processing, respectively. In the current thesis, we took both perspectives.

In chapters 2 – 4, we applied regression models to statistics derived from various multi-word and multi-syllable chunks, extracted from large corpora of child-directed speech. There, our goal was to investigate how children use knowledge from the perceptual domain in order to break into the linguistic domain. In doing so, we stayed at the level of psychological mechanisms, considering the extraction and storage of syllable chunks as abstract computational operations that allow children to identify hypothesized linguistic units during early speech segmentation. In chapter 5, we then examined bootstrapping at the neural level, using deep neural networks to model the use of neural circuits, specialized for speech processing in the domain of normal hearing, to bootstrap category perception in cochlear implant users.

6.1 Summary of results

Speech segmentation

With respect to speech segmentation, we conducted a series of three studies – over the course of which we developed a method for gauging the facilitatory impact of multi-word chunks on child word learning (chapter 2); isolated this effect with respect to a number of potentially confounding variables (chapter 3); and compared effects across three different chunk types, defined at the syllable level (chapter 4).

Beginning in chapter 2, we extracted multi-word units (MWUs) from a corpus of child-directed speech, and we counted the number of MWUs containing a set of target words. The resulting statistic, the number of MWUs containing each target word, was then correlated with the age at which children first produce the targets (age of first production / AoFP). The correlation was negative, even when controlling for target word frequency, suggesting that early-learned words tend to be included in a large number of MWUs.

Following a proposal from Peters (1983), we suggested that children store MWUs as

undersegmented chunks before having segmented the words contained within them. If children compare stored chunks to one another and to incoming speech, words contained within a large number of stored chunks should be segmented (and discovered as independent linguistic units) before words contained within fewer chunks – explaining the negative correlation between AoFP and the number of MWUs per target word.

The first important result, presented in chapter 2, thus is the development of a new method for investigating early speech segmentation: By correlating the number of chunks per word with AoFP, we can evaluate how likely it is that children store a particular selection of chunks as undersegmented wholes. Chunks that result in strong negative correlations are more plausible candidates for wholesale storage during speech segmentation, whereas those that lead to weak correlations are less plausible candidates.

In chapter 3, we isolated the effect of multi-word chunks from a number of psycholinguistically relevant co-variates, beyond simple word frequency – using multiple linear regression instead of pairwise correlation. In addition to the co-variates, we also calculated an effect for the number of context words contained within the MWUs used to compute the key independent variable (the number of MWUs per target word). The effect of this variable was such that the more context words a target word co-occurs with, the *later* it is learned – yet the more MWUs contain a word, the *earlier* that word is learned. This result implies that there is a unique facilitatory effect of multi-word chunks, different from the effect of co-occuring words, underscoring the plausibility of our earlier interpretation that children segment larger chunks before discovering smaller linguistic units.

In chapter 4, finally, we compared correlations between AoFP and three different types of chunks, focusing on (a) particularly frequent, (b) particularly internally predictable, and (c) particularly short chunks. Overall, short chunks were more strongly associated with early word learning – and, apart from that, were more likely to correspond to words than the other two chunk types. Out of the three types considered, that is, short chunks are both most word-like and most likely to be stored, by young children, as undersegmented chunks.

The implication for early speech segmentation is that children store undersegmented chunks that are similar to discrete linguistic units such as words – which are short,

first and foremost, rather than frequent or internally predictable. Additional implications arise for cognitive representations of formulaic multi-word sequences in older language users. These are often conceptualized as corresponding to particularly frequent sequences. Yet our results suggest that children extract short, word-like chunks, not frequent multi-word chunks. It follows that formulaic multi-word sequences are unlikely to originate as undersegmented chunks during the speech segmentation process.

Category perception with cochlear implants

In chapter 5, we used deep neural networks to model speech perception with cochlear implants (CIs). The experimental design involved training neural networks on spectrograms intended to approximate the spectrally degraded speech signals delivered through CIs. A crucial input modification pertained to the presence of channel interaction: the spread of information, prior to neural activation, to neighboring channels. Across the board, the networks performed worse if channel interaction was present in the spectrograms.

Beyond that, an additional difference emerged for networks that were pre-trained on high-resolution spectrograms, before exposure to CI-like input. For both types of networks – pre-trained or not –, the presence of channel interaction was associated with degraded speech recognition performance. But for the pre-trained networks only, speech recognition performance *also* improved more slowly when transitioning to spectrograms with added channel interaction, relative to spectrograms without channel interaction.

In CI users, channel interaction is associated with poor speech recognition performance. The fact that we were able to model this suggests that neural networks can be used to investigate other facets of speech processing with CIs. The pattern we obtained with the pre-trained networks, in particular, has ramifications for implant research: Given that pre-training on high-resolution spectrograms leads to slower adaptation to CI-like input in the presence of channel interaction, we should expect a similar effect in CI users who are implanted after a period of normal hearing. Such postlingually deaf CI users adapt to high-resolution signals, delivered through the intact cochlea, during normal hearing. Once implanted, channel interaction should

slow adaptation to CIs in this population, but not in subjects who were born deaf – i.e., in subjects who never adapted to processing high-resolution speech.

6.2 Testable predictions

Computational models can be thought of as specifications over a set of assumptions. In designing them, one is forced to spell out underlying assumptions about what the relevant variables are, as well as the manner in which they interact with one another. Considered by itself, it can be a useful exercise to create models that explain the available data, in that it enforces a degree of specificity not necessarily required by verbal theories.

But in the spirit of the scientific method, computational models should also generate predictions about data that have not yet been collected. Compared to matching existing data, correct predictions present more powerful evidence for the validity of a given model. After all, one can tweak parameters until one's model fits the available data, but the predictions that follow from this cannot be constrained in a similar fashion. Aside from this, novel predictions can inform the scientific process by opening up new avenues of inquiry and prodding future research into directions that would not otherwise have been considered. Here, we present predictions, corresponding to the two areas covered in the thesis, that can be tested with human subjects.

First, the results presented in chapter 4 suggest that short syllable sequences, compared to frequent or internally predictable ones, are (a) well-suited for predicting when their component words are learned and (b) likely to correspond to individual words. Based on this, we suggested that infants and young children preferentially extract short, word-like syllable sequences from unsegmented speech, to be stored as hypothesized linguistic units.

Accordingly, we predict that infant subjects, in an artificial word segmentation task similar to the paradigm used in Saffran et al. (1996)'s seminal study, should give precedence to short instead of predictable or frequent syllable sequences. If empirically verified, this would imply that transitional probabilities between syllables, thought to be an important segmentation cue (Saffran et al., 1996; Aslin et al.,

1998; Aslin, 2017), are less crucial than sequence length; and that whole-sequence frequency, thought to determine the degree to which multi-word sequences become entrenched in adult memory (Arnon and Snider, 2010; Arnon and Priva, 2014), plays a less important role in determining which types of chunks are stored by children.

Next to speech segmentation, we also investigated speech processing with cochlear implants. We reported, in chapter 5, that the presence of channel interaction in an approximated implant-delivered signal slows learning in deep neural networks that were pre-trained on input with a higher spectral resolution, but not in networks that were only trained on CI-like input. This implies that the presence of channel interaction slows adaptation to CIs in postlingually deaf CI users (born normally hearing), but not in congenitally deaf CI users (born deaf).

A straightforwardly derived prediction for implant research is that the reduction of channel interaction (Landsberger and Srinivasan, 2009) will accelerate the transition to CIs in postlingually deaf CI users, but not in congenitally deaf CI users. Such a pattern would show that the spectral degradation caused by channel interaction has two distinct effects on speech processing in CI users: (1) a negative effect on overall speech recognition performance, which is already well-established (Stickney et al., 2006); and (2) a negative effect on the transition process, from normal hearing to CIs, in postlingually deaf listeners. This second effect would be an additional incentive, beyond boosting speech recognition performance in implanted subjects, to reduce channel interaction as much as possible.

Appendix A

Supplementary material, chapter 2

Correlation coefficients

Table A.1: Full and partial correlation coefficients with 95 % confidence intervals (in parentheses) for all correlations reported in analyses II – IV.

		ADS-#Freq	CDS-#Freq	ADS-#MWUs	CDS-#MWUs	#baseline
full	RTs	-0.30 (-0.28, -0.31)	-0.24 (-0.22, -0.25)	-0.30 (-0.28, -0.31)	-0.25 (-0.23, -0.26)	-0.30 (-0.29, -0.31)
	AoFP	-0.30 (-0.29, -0.32)	-0.45 (-0.44, -0.46)	-0.30 (-0.29, -0.32)	-0.45 (-0.44, -0.46)	-0.45 (-0.44, -0.46)
partial	RTs	-0.08 (-0.07, -0.09)	-0.05 (-0.04, -0.06)	-0.09 (-0.08, -0.10)	-0.09 (-0.08, -0.10)	-0.06 (-0.05, -0.06)
	AoFP	-0.08 (-0.07, -0.09)	-0.14 (-0.13, -0.15)	-0.09 (-0.08, -0.10)	-0.14 (-0.13, -0.15)	-0.05 (-0.04, -0.06)

CHILDES corpora

1. CHILDES corpora used for CDS corpus:
   Belfast (Henry, 1995), Fletcher (Fletcher and Garman, 1988), Manchester (Theakston et al., 2001), Thomas (Lieven et al., 2009b), Tommerdahl (Tommerdahl and Kilpatrick, 2013), Wells (Wells, 1981), Forrester (Forrester, 2002), Lara (Rowland and Fletcher, 2006)

2. CHILDES corpora used for AoFP corpus:
   Bates (Bates et al., 1991), Bernstein-Ratner (Ratner, 1986), Bliss (Bliss, 1988), Bloom 1970 (Bloom et al., 1974), Bloom 1973 (Bloom, 1976), Bohannon (Bohannon III and Marquis, 1977), Braunwald (Braunwald, 1971), Brent (Brent and Siskind, 2001), Brown (Brown, 1973), Carterette (Jones and Carterette, 1963), Clark (Clark, 1978), Cornell (no citation provided), Demetras (Demetras, 1986), Ervin-Tripp (no citation provided), Evans (no citation provided), Feldman (Feldman and Menn, 2003), Garvey (Garvey and Hogan, 1973), Gathercole (Gathercole, 1980), Gleason (Masur and Gleason, 1980), HSLLD (Beals, 1993), Hall (Hall et al., 1984), Higginson (Higginson, 1985), Kuczaj (Kuczaj, 1977), MacWhinney (MacWhinney, 2000b), McCune (McCune, 1995), McMillan (no citation provided), Morisset (Morisset et al., 1995), Nelson (Nelson, 1989), NewEngland (Ninio et al., 1994), Peters/Wilson (Peters, 1987), Post (Demetras et al., 1986), Providence (Song et al., 2013), Rollins (Rollins, 2003), Sachs (Sachs, 1983), Snow (MacWhinney and Snow, 1990), Soderstrom (Soderstrom et al., 2008), Sprott (no citation provided), Suppes (Suppes, 1974), Tardif (no citation provided), Valian (Valian, 1991), Van Houten (Van Houten, 1986), Van Kleeck (no citation provided), Warren-Leubecker (Warren-Leubecker and Bohannon III, 1984), Weist (Weist and Zevenbergen, 2008)

Appendix B

Supplementary material, chapter 4

Child-Produced Speech Statistics

measure	BE	NA
# children	247	743
mean child age (months)	32.66 (SD = 9.25)	41.39 (SD = 23.45)
# utterances	873,623	846,894
mean utterance length (words)	3.45 (SD = 2.95)	2.51 (SD = 1.88)
# tokens	3,016,863	2,130,946
# types	43,510	24,322

Table B.1: Statistics for child-produced speech used to estimate corpus-derived AoFP.

CHILDES corpora

The aggregated corpora of transcribed adult- and child-produced speech used in chapter 4 contain material from all (British and North American) English corpora on the CHILDES data base (MacWhinney, 2000a) where the target children were normally developing[1].

1. CHILDES corpora used for the BE corpus (10):

   Belfast (Henry, 1995), Fletcher (Fletcher and Garman, 1988), Forrester (Forrester, 2002), Howe (Howe, 1981), Lara (Rowland and Fletcher, 2006), MPI-EVA-Manchester (Lieven et al., 2009a), Manchester (Theakston et al., 2001), Thomas (Lieven et al., 2009a), Tommerdahl (Tommerdahl and Kilpatrick, 2013), Wells (Wells, 1981)

2. CHILDES corpora used for the NA corpus (41):

   Bates (Bates et al., 1991), Bernstein-Ratner (Ratner, 1986), Bliss (Bliss, 1988), Bloom 1970 (Bloom et al., 1974), Bloom 1973 (Bloom, 1976), Bohannon (Bohannon III and Marquis, 1977), Braunwald (Braunwald, 1971), Brent (Brent and Siskind, 2001), Brown (Brown, 1973), Clark (Clark, 1978), Cornell (no reference provided), Demetras-Trevor (Demetras, 1986), Evans (no reference provided), Feldman (Feldman and Menn, 2003), Garvey (Garvey and Hogan, 1973), Gathercole (no reference provided), Gleason (Masur and Gleason, 1980), HSLLD (Beals, 1993), Hall (Hall et al., 1984), Higginson(no reference provided), Kuczaj (Kuczaj, 1977), MacWhinney (MacWhinney, 2000b), McCune (McCune, 1995), McMillan (no reference provided), Morisset (Morisset et al., 1995), New England (Ninio et al., 1994), Post (Demetras et al., 1986), Providence (Song et al., 2013), Rollins (Rollins, 2003), Sachs (Sachs, 1983), Sawyer (Sawyer, 1997), Snow (MacWhinney and Snow, 1990), Soderstrom (Soderstrom et al., 2008), Sprott (no reference provided), Suppes (Suppes, 1974), Tardif (no reference provided), Valian (Valian, 1991), Van Houten (Van Houten, 1986), Van Kleeck (no reference provided), Warren-Leubecker (Warren-Leubecker and Bohannon III, 1984), Weist (Weist and Zevenbergen, 2008)

[1] Available online: `https://childes.talkbank.org/`. Data were downloaded August 2018

Correlations between Predictors

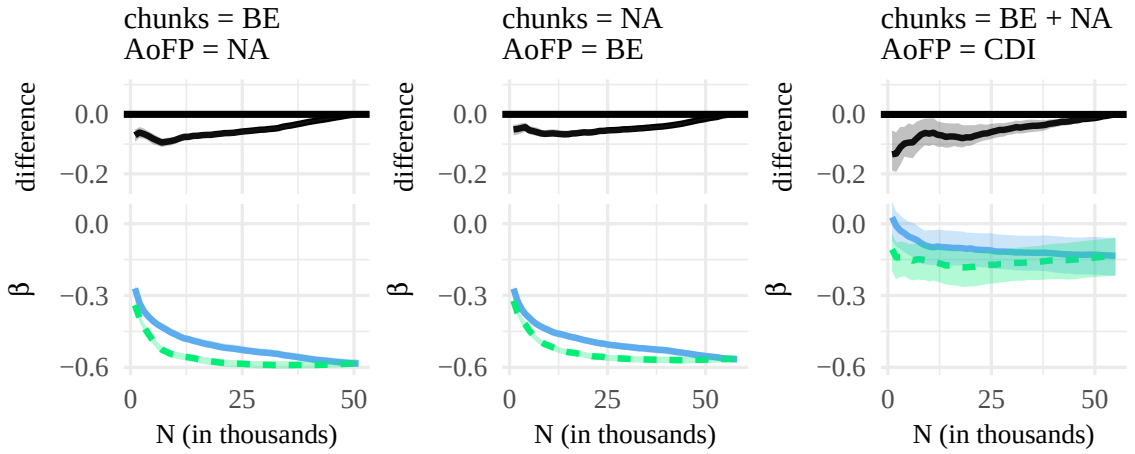
Table B.2 shows pairwise correlations between the predictors used in the linear regression analyses. Correlations between $\#MSU$ counts and the co-variates are mostly weak to moderate, but we also find a few stronger correlations (e.g. word frequency is strongly positively correlated with all $\#MSU$ counts). Including collinear predictors in regression models can lead to unstable results. It is thus important that similar results are obtained when the co-variates are excluded. Results for analyses without co-variates are reported on the following pages.

corpus	1. S	2. F	3. P	4. Freq	5. Con	6. Nsyl	7. PhonN
BE	1. —	0.65 ***	0.62 ***	0.69 ***	0.04 ***	-0.24 ***	0.26 ***
	2. —	—	0.69 ***	0.69 ***	-0.07 ***	-0.08 ***	0.14 ***
	3. —	—	—	0.62 ***	-0.02	-0.03 *	0.02 *
	4. —	—	—	—	-0.04 ***	-0.18 ***	0.22 ***
	5. —	—	—	—	—	-0.03 **	0.01
	6. —	—	—	—	—	—	-0.74 ***
NA	1. —	0.64 ***	0.59 ***	0.70 ***	0.03 **	-0.26 ***	0.27 ***
	2. —	—	0.68 ***	0.71 ***	-0.07 ***	-0.08 ***	0.16 ***
	3. —	—	—	0.62 ***	-0.01	-0.01	0.02 *
	4. —	—	—	—	-0.06 ***	-0.20 ***	0.26 ***
	5. —	—	—	—	—	-0.02	-0.01
	6. —	—	—	—	—	—	-0.74 ***
BE + NA	1. —	0.76 ***	0.61 ***	0.81 ***	-0.38 ***	-0.63 ***	0.59 ***
	2. —	—	0.83 ***	0.92 ***	-0.46 ***	-0.25 ***	0.34 ***
	3. —	—	—	0.81 ***	-0.43 ***	-0.11 ***	0.19 ***
	4. —	—	—	—	-0.52 ***	-0.33 ***	0.40 ***
	5. —	—	—	—	—	0.20 ***	-0.18 ***
	6. —	—	—	—	—	—	-0.72 ***

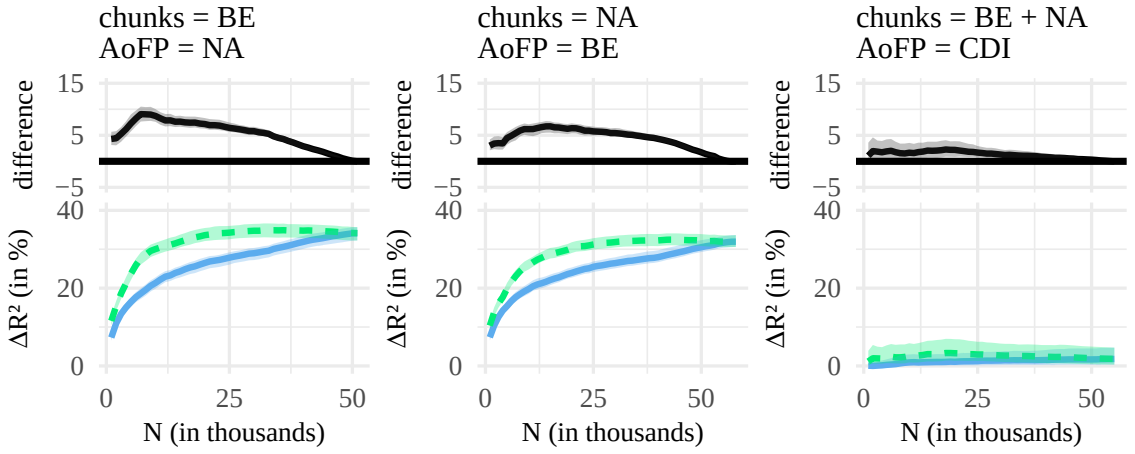
Table B.2: Pairwise correlations (Spearman's ρ) for predictors used in regression analyses. $S = \#MSU\text{-}S$, $F = \#MSU\text{-}F$, $P = \#MSU\text{-}P$. ***: $p \leq 0.001$. **: $p \leq 0.01$. *: $p \leq 0.05$.

Results without Covariates

Figures B.1, B.2, and B.3 are fashioned after figures 4, 5, and 6 from analysis III, except that we do not control for co-variates. In contrast to analysis III, we find that *MSU-F* performs slightly better than *MSU-P* when the co-variates are excluded (figure B.3). Importantly, however, we replicate the key finding from analysis III: As long as N is not very small or very large, high *MSU-S* counts are more strongly predictive of early AoFP than either high *MSU-F* (figure B.1) or high *MSU-P* (figure B.2) counts.

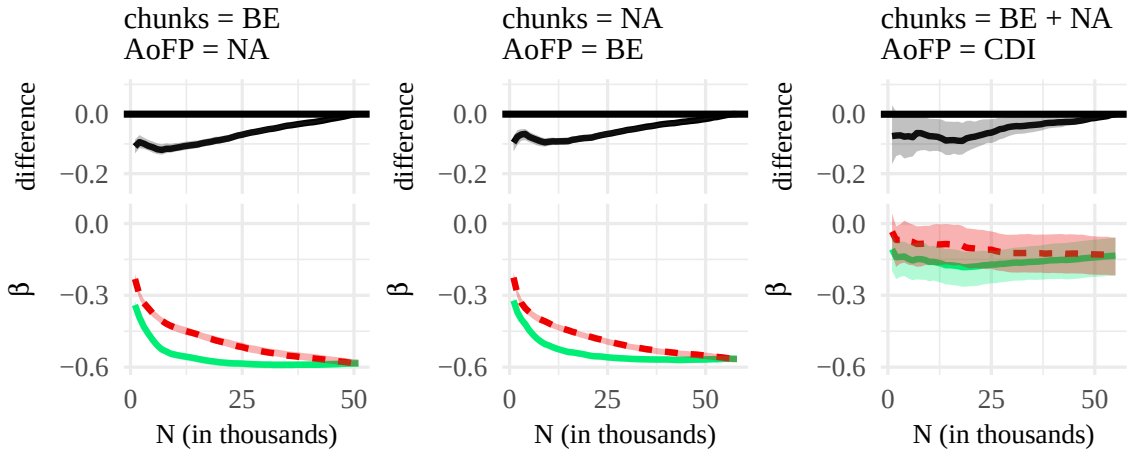


(a) Bottom: Regression coefficients (β). Top: difference between coefficients.

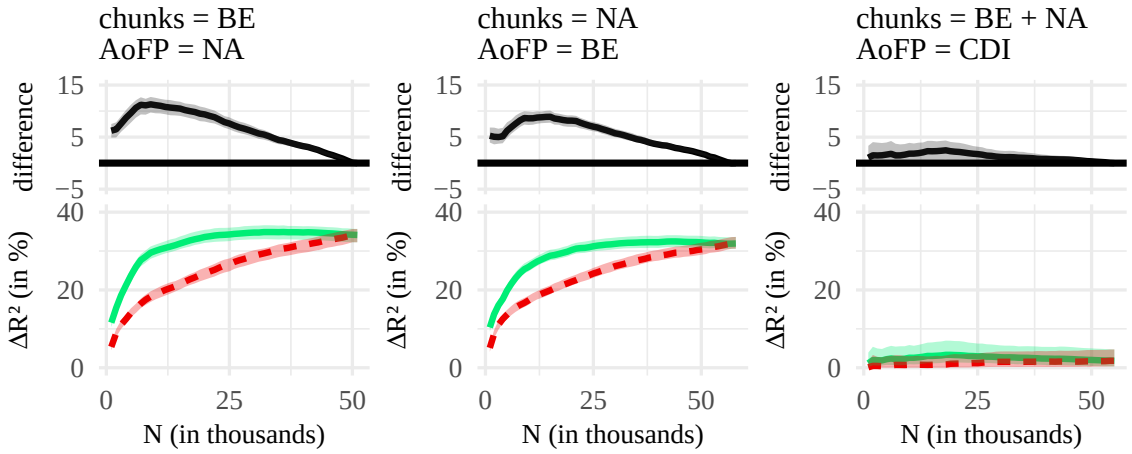


(b) Bottom: amount of variance in AoFP (ΔR^2 in %) . Top: difference between R^2 values.

Figure B.1: Comparison of *#MSU-S* (green line) and *#MSU-F* (blue line).

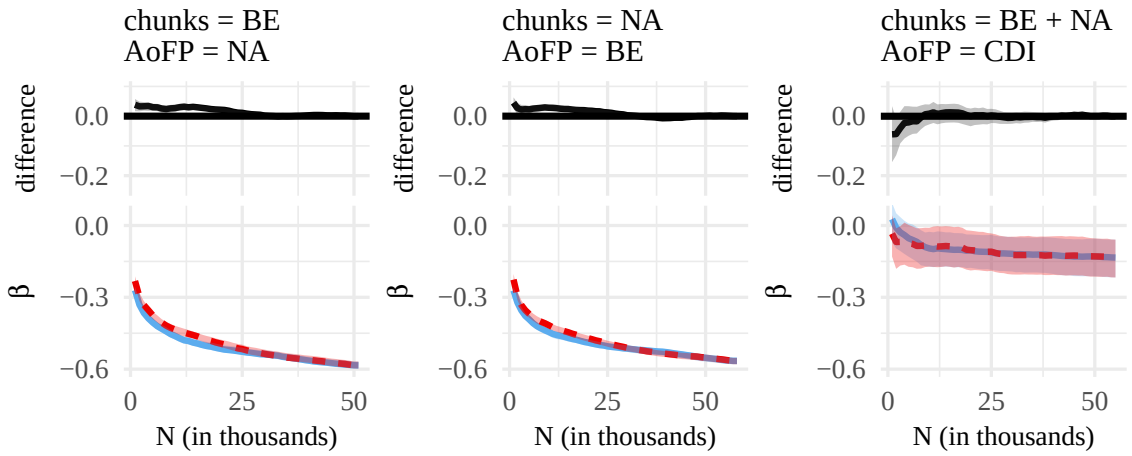


(a) Bottom: Regression coefficients (β). Top: difference between coefficients.

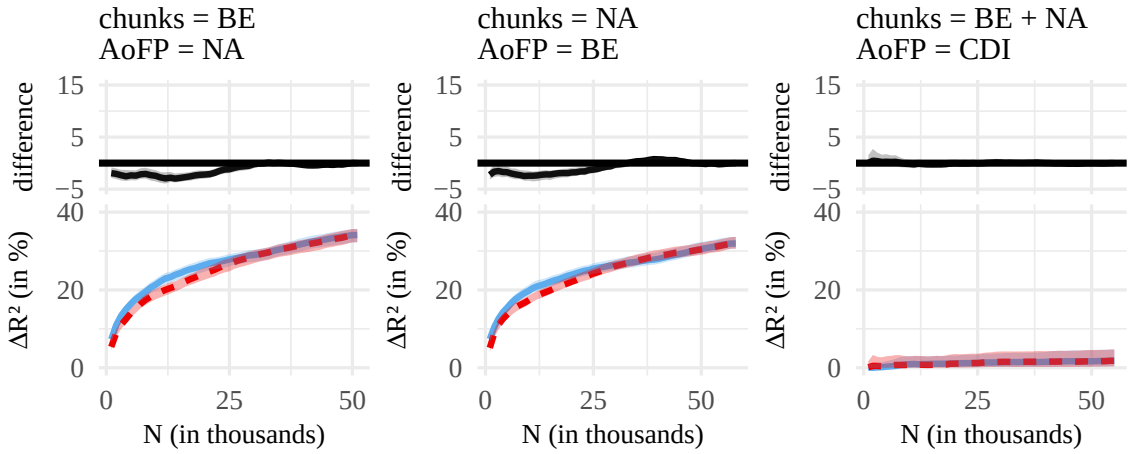


(b) Bottom: amount of variance in AoFP (ΔR^2 in %) . Top: difference between R^2 values.

Figure B.2: Comparison of $\#MSU\text{-}S$ (green line) and $\#MSU\text{-}P$ (red line).



(a) Bottom: Regression coefficients (β). Top: difference between coefficients.



(b) Bottom: amount of variance in AoFP (ΔR^2 in %) . Top: difference between R^2 values.

Figure B.3: Comparison of $\#MSU\text{-}P$ (red line) and $\#MSU\text{-}F$ (blue line).

Bibliography

Adelman, J. S., Brown, G. D., and Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. John Wiley & Sons, New York, 3rd edition.

Altmann, G. and Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4):583–609.

Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(02):239–273.

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96:703–719.

Anderson, J. R. and Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6):396–408.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4):245–266.

Anderson, M. L., Brumbaugh, J., and Şuben, A. (2010). Investigating functional cooperation in the human brain using simple graph-theoretic methods. In Chaovalitwongse, W., Pardalos, P. M., and Xanthopoulos, P., editors, *Computational Neuroscience*, pages 31–42. Springer, New York, NY.

Arnon, I. and Christiansen, M. H. (2017). The role of multiword building blocks in explaining l1–l2 differences. *Topics in Cognitive Science*, 9(3):621–636.

Arnon, I. and Clark, E. V. (2011). Why brush your teeth is better than teeth–children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, 7(2):107–129.

Arnon, I., McCauley, S. M., and Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, 92:265–280.

Arnon, I. and Priva, U. C. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon*, 9(3):377–400.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1-2).

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

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30:1–47.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, 62(5):874–890.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., and Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology*, 133(2):283.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, 39(3):445–459.

Bannard, C., Lieven, E., and Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, 106(41):17284–17289.

Bannard, C. and 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.

Barry, C., Hirsh, K. W., Johnston, R. A., and Williams, C. L. (2001). Age of acquisition, word frequency, and the locus of repetition priming of picture naming. *Journal of Memory and Language*, 44(3):350–375.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, 22(04):637–660.

Bartlett, S., Kondrak, G., and Cherry, C. (2009). On the syllabification of phonemes. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 308–316. Association for Computational Linguistics.

Bates, E., Bretherton, I., and Snyder, L. (1991). *From first words to grammar: Individual differences and dissociable mechanisms*. Cambridge University Press, Cambridge.

Beals, D. E. (1993). Explanatory talk in low-income families' mealtime conversations. *Applied Psycholinguistics*, 14(04):489–513.

Behrens, H. (2009). Usage-based and emergentist approaches to language acquisition. *Linguistics*, 47(2):383–411.

Bengio, Y. (2012). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36.

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

Biber, D., Conrad, S., and Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405.

Bijeljac-Babic, R., Bertoncini, J., and Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, 29(4):711–721.

Bizley, J. K. and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693.

Bliss, L. (1988). The development of modals. *Journal of Applied Developmental Psychology*, 9:253–261.

Bloom, L. (1976). *One word at a time: The use of single word utterances before syntax*. Walter de Gruyter, Berlin.

Bloom, L., Hood, L., and Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, 6:380–420.

Bloomfield, L. (1933). *Language*. Holt, New York.

Bohannon III, J. N. and Marquis, A. L. (1977). Children's control of adult speech. *Child Development*, 80:1002–1008.

Bonin, P., Barry, C., Méot, A., and Chalard, M. (2004). The influence of age of acquisition in word reading and other tasks: A never ending story? *Journal of Memory and language*, 50(4):456–476.

Borensztajn, G., Zuidema, W., and Bod, R. (2009). Children's grammars grow more abstract with age–evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, 1(1):175–188.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., and Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4):298–304.

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pages 1691–1690.

Braunwald, S. R. (1971). Mother-child communication: The function of maternal-language input. *Word*, 27(1-3):28–50.

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

Brooke, J., Hammond, A., Jacob, D., Tsang, V., Hirst, G., and Shein, F. (2015). Building a lexicon of formulaic language for language learners. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 96–104, Denver, Colorado. Association for Computational Linguistics.

Brooke, J., Tsang, V., Hirst, G., and Shein, F. (2014). Unsupervised multiword segmentation of large corpora using prediction-driven decomposition of n-grams. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 753–761, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press, Cambridge, Massachusetts.

Brysbaert, M. and Cortese, M. J. (2010). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology*, 64(3):545–559.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Burnard, L. (2007). Reference guide for the british national corpus (xml edition). (accessed on: 12.07.2016).

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., and Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, 63(4):i–174.

Carroll, J. B. and White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1):85–95.

Cartwright, T. A. and Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, 63(2):121–170.

Chatterjee, M., Deroche, M. L., Peng, S.-C., Lu, H.-P., Lu, N., Lin, Y.-S., and Limb, C. J. (2018). Processing of fundamental frequency changes, emotional prosody and lexical tones by pediatric ci recipients. In *Proceedings of the International Symposium on Auditory and Audiological Research*, volume 6, pages 117–125.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., and Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing Research*, 322:151–162.

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In Sinclair, R. J. A. and Levelt, W., editors, *The Child's Conception of Language*, pages 17–43. Springer Verlag, Berlin.

Clark, E. V. (2009). *First Language Acquisition*. Cambridge University Press, Cambridge, UK.

Clark, R. (1974). Performing without competence. *Journal of Child Language*, 1:1–10.

Coath, M. and Denham, S. L. (2007). The role of transients in auditory processing. *Biosystems*, 89(1-3):182–189.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1):204.

Cullington, H. E. and Zeng, F.-G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects a. *The Journal of the Acoustical Society of America*, 123(1):450–461.

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

Dale, P. S. (1991). The validity of a parent report measure of vocabulary and syntax at 24 months. *Journal of Speech, Language, and Hearing Research*, 34(3):565–571.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

DeCasper, A. J. and Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, 208(4448):1174–1176.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The neuronal recycling hypothesis. In Dehaene, S., Duhamel, J. R., Hauser, M. D., and Rizzolatti, G., editors, *From Monkey Brain to Human Brain: A Fyssen Foundation Symposium*, pages 133–157. MIT Press, Cambridge, Massachusetts.

Dehaene, S. and Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262.

Demetras, M. J., Post, K. N., and Snow, C. E. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, 13(02):275–292.

Demetras, M. J.-A. (1986). Working parents conversational responses to their two-year-old sons. Working Paper. University of Arizona.

Durieux, G. and Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test. In Weissenborn, J. and Höhle, B., editors, *Approaches to bootstrapping: Phonological, syntactic and neurophysiological aspects of early language acquisition*, pages 189–232. John Benjamins Publishing Company, Amsterdam.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582.

Estes, K. G., Evans, J. L., Alibali, M. W., and Saffran, J. R. (2007). Can infants map meaning to newly segmented words? statistical segmentation and word learning. *Psychological Science*, 18(3):254–260.

Feldman, A. and Menn, L. (2003). Up close and personal: A case study of the development of three english fillers. *Journal of Child Language*, 30(04):735–768.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., and Thal, D. J. (2007). *The MacArthur-Bates Communicative Development Inventories User's*

Guide and Technical Manual, Second Edition. Paul H. Brookes Publishing Company.

Fisher, C., Gertner, Y., Scott, R. M., and Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149.

Fitneva, S. A., Christiansen, M. H., and Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, 36(5):967–997.

Fletcher, P. and Garman, M. (1988). Normal language development and language impairment: Syntax and beyond. *Clinical Linguistics & Phonetics*, 2(2):97–113.

Forrester, M. A. (2002). Appropriating cultural conceptions of childhood participation in conversation. *Childhood*, 9:255–276.

François, C., Cunillera, T., Garcia, E., Laine, M., and Rodriguez-Fornells, A. (2017). Neurophysiological evidence for the interplay of speech segmentation and word-referent mapping during novel word learning. *Neuropsychologia*, 98:56–67.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, 110(2):1150–1163.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Garvey, C. and Hogan, R. (1973). Social speech and social interaction: Egocentrism revisited. *Child Development*, 44:562–568.

Gathercole, V. (1980). *Birdies like birdseed the bester than buns: A study of relational comparatives and their acquisition*. Unpublished Doctoral Dissertation.

Gaudrain, E. and Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *The Journal of the Acoustical Society of America*, 137(3):1298–1308.

Gaudrain, E. and Baskent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear and Hearing*.

Gerken, L., Wilson, R., and Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32(2):249–268.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press, Oxford.

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

Grimm, R., Cassani, G., Gillis, S., and Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical processing and word learning: A computational investigation. *Frontiers in Psychology*, 8.

Hall, W. S., Nagy, W. E., and Linn, R. L. (1984). *Spoken words, effects of situation and social group on oral word usage and frequency*. Lawrence Erlbaum, Hillsdale, NJ.

Harris, Z. S. (1954). *Methods in Structural Linguistics*. University of Chicago Press, Chicago, Illinois.

Harrison, R. V., Gordon, K. A., and Mount, R. J. (2005). Is there a critical period for cochlear implantation in congenitally deaf children? analyses of hearing and speech perception performance after implantation. *Developmental Psychobiology*, 46(3):252–261.

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. Oxford University Press, New York.

Higginson, R. P. (1985). *Fixing: Assimilation in language acquisition*. Unpublished Doctoral Dissertation.

Hills, T. T., Maouene, J., Riordan, B., and Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, 63(3):259–273.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., and Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, 20(6):729–739.

Howe, C. (1981). *Acquiring language in a conversational context*. Academic Press, New York.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., and Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, 87:38–58.

Johns, B. T., Dye, M., and Jones, M. N. (2014). The influence of contextual variability on word learning. In Bello, P., Guarini, M., McShane, M., and Scassellati, B., editors, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 242–247, Austin, TX. Cognitive Science Society.

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., and Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, 132(2):EL74–EL80.

Johnson, E. K. and 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.

Jolly, H. R. and Plunkett, K. (2008). Inflectional bootstrapping in 2-year-olds. *Language and Speech*, 51(1-2):45–59.

Jones, G. L., Ho Won, J., Drennan, W. R., and Rubinstein, J. T. (2013). Relationship between channel interaction and spectral-ripple discrimination in cochlear implant users. *The Journal of the Acoustical Society of America*, 133(1):425–433.

Jones, M. H. and Carterette, E. C. (1963). Redundancy in children's free-reading choices. *Journal of Verbal Learning and Verbal Behavior*, 2:489–493.

Jones, M. N., Johns, B. T., and Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 66(2):115–124.

Jusczyk, P. W. and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1):1–23.

Jusczyk, P. W. and Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, 23(5):648.

Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16(5):589–600.

Lakoff, G. (1987). *Women, fire, and dangerous things*. University of Chicago Press, Chicago.

Landsberger, D. M. and Srinivasan, A. G. (2009). Virtual channel discrimination is improved by current focusing in cochlear implant recipients. *Hearing Research*, 254(1-2):34–41.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10):1995.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

Lee, H., Ekanadham, C., and Ng, A. Y. (2007). Sparse deep belief net model for visual area v2. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S., editors, *Proceedings of Neural Information Processing Systems 20*, pages 873–880.

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

Lieven, E., Salomo, D., and Tomasello, M. (2009a). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

Lieven, E., Salomo, D., and Tomasello, M. (2009b). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3):481–507.

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

Lignos, C. and Yang, C. (2010). Recession segmentation: simpler online word segmentation using limited resources. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 88–97. Association for Computational Linguistics.

MacWhinney, B. (2000a). *The CHILDES project: The database*. Psychology Press, Oxfordshire.

MacWhinney, B. (2000b). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26:657–657.

MacWhinney, B. and Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17(02):457–472.

Mandel, D. R., Jusczyk, P. W., and Nelson, D. G. K. (1994). Does sentential prosody help infants organize and remember speech information? *Cognition*, 53(2):155–180.

Mao, Y. and Xu, L. (2017). Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants. *International Journal of Audiology*, 56(sup2):S23–S30.

Marchetto, E. and Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cognitive Psychology*, 67(3):130–150.

Martin, A., Peperkamp, S., and Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124.

Masur, E. F. and Gleason, J. B. (1980). Parent–child interaction and the acquisition of lexical information during play. *Developmental Psychology*, 16(5):404–409.

Matychuk, P. (2005). The role of child-directed speech in language acquisition: a case study. *Language Sciences*, 27(3):301–379.

McCauley, S. M. and Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The cappuccino model. In Carlson, L. A., Hölscher, C., and Shipley, T. F., editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1619–24, Austin, TX. Cognitive Science Socitety.

McCauley, S. M. and Christiansen, M. H. (2014). Acquiring formulaic language: A computational model. *The Mental Lexicon*, 9(3):419–436.

McCauley, S. M., Monaghan, P., and Christiansen, M. H. (2015). Language emergence in development. In Brian, M. and O'Grady, W., editors, *The Handbook of Language Emergence*, pages 415–436. Wiley-Blackwell, Hoeboekn, NJ.

McCune, L. (1995). A normative study of representational play in the transition to language. *Developmental Psychology*, 31(2):198.

McDermott, J. H. and Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5):926–940.

McDonald, S. A. and Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3):295–322.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. In Hirsh-Pasek, K. and Golinkoff, R. M., editors, *Action meets word: How children learn verbs*, pages 31–63.

Moberly, A. C., Lowenstein, J. H., Tarr, E., Caldwell-Tarr, A., Welling, D. B., Shahin, A. J., and Nittrouer, S. (2014). Do adults with cochlear implants rely on different acoustic cues for phoneme perception than adults with normal hearing? *Journal of Speech, Language, and Hearing Research*, 57(2):566–582.

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

Morisset, C. E., Barnard, K. E., and Booth, C. L. (1995). Toddlers' language development: Sex differences within social risk. *Developmental Psychology*, 31(5):851.

Morrison, C. M., Chappell, T. D., and 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.

Naigles, L. G. and Kako, E. T. (1993). First contact in verb acquisition: Defining a role for syntax. *Child Development*, 64(6):1665–1687.

Nelson, K. (1989). *Narratives from the Crib*. Harvard University Press, Cambridge, Massachusetts.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., and 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.

Ninio, A., Snow, C. E., Pan, B. A., and Rollins, P. R. (1994). Classifying communicative acts in children's interactions. *Journal of Communication Disorders*, 27(2):157–187.

Noreen, E. W. (1989). *Computer-intensive methods for testing hypotheses*. Wiley, New York.

O'Donnell, T. J. (2015). *Productivity and Reuse in Language: A theory of Linguistic Computation and Storage*. MIT Press.

Oh, S.-h., Kim, C.-s., Kang, E. J., Lee, D. S., Lee, H. J., Chang, S. O., Ahn, S.-h., Hwang, C. H., Park, H. J., and Koo, J. W. (2003). Speech perception

after cochlear implantation over a 4-year time period. *Acta Oto-Laryngologica*, 123(2):148–153.

Papafragou, A., Cassidy, K., and Gleitman, L. (2007). When we think about thinking: The acquisition of belief verbs. *Cognition*, 105(1):125–165.

Parker, M. D. and Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (mlum) and mean length of utterance in words (mluw). *First Language*, 25(3):365–376.

Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2):244–247.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge University Press, Cambridge, New York.

Peters, A. M. (1987). The role of imitation in the developing syntax of a blind child. *Text-Interdisciplinary Journal for the Study of Discourse*, 7(3):289–309.

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

Pine, J. M. and Lieven, E. V. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language*, 20(3):551–571.

Räsänen, O., Doyle, G., and Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, pages 3204–3208.

Räsänen, O., Doyle, G., and Frank, M. C. (2018). Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171:130–150.

Räsänen, O. and Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, 122(4):792.

Ratner, N. B. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics*, 14:303–309.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Rollins, P. (2003). Caregiver contingent comments and subsequent vocabulary comprehension. *Applied Psycholinguistics*, 24:221–234.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., and Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences*, 104(17):7295–7300.

Rowland, C. F. and Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, 33:859–877.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's Language*, 4:1–28.

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

Sawyer, K. (1997). *Pretend play as improvisation*. Erlbaum, Mahwah, New Jersey.

Saxton, M. (2009). The inevitability of child directed speech. In Foster-Cohen, S., editor, *Language Acquisition*, pages 62–86. Palgrave Macmillan, New York.

Saxton, M. (2010). *Child language: Acquisition and development*. Sage Publications, London.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523.

Shannon, R. V. (1983). Multichannel electrical stimulation of the auditory nerve in man. ii. channel interaction. *Hearing Research*, 12(1):1–16.

Shera, C. A., Guinan, J. J., and Oxenham, A. J. (2002). Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. *Proceedings of the National Academy of Sciences*, 99(5):3318–3323.

Soderstrom, M., Blossom, M., Foygel, R., and Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, 35(04):869–902.

Song, J. Y., Demuth, K., Evans, K., and Shattuck-Hufnagel, S. (2013). Durational cues to fricative codas in 2-year-olds' american english: Voicing and morphemic factors. *The Journal of the Acoustical Society of America*, 133:2931–2946.

Sprenger, S. A., Levelt, W. J., and Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, 54(2):161–184.

Stickney, G. S., Loizou, P. C., Mishra, L. N., Assmann, P. F., Shannon, R. V., and Opie, J. M. (2006). Effects of electrode design and configuration on channel interactions. *Hearing Research*, 211(1-2):33–45.

Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *The Journal of the Acoustical Society of America*, 116(2):1081–1091.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(2):201–221.

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.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, 29(2):103–114.

Tang, Q., Benítez, R., and Zeng, F.-G. (2011). Spatial channel interactions in cochlear implants. *Journal of Neural Engineering*, 8(4):046029.

Testolin, A., Stoianov, I., and Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, 1(9):657.

Theakston, A. L., Lieven, E. V., Pine, J. M., and Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28(01):127–152.

Thiessen, E. D., Hill, E. A., and +, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, 7(1):53–71.

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

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Tomasello, M. (2009). *Constructing a Language*. Harvard University Press.

Tommerdahl, J. and Kilpatrick, C. D. (2013). The reliability of morphological analyses in language samples. *Language Testing*, 31:3–18.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Valian, V. (1991). Syntactic subjects in the early speech of american and italian children. *Cognition*, 40(1–2):21–81.

Van Houten, L. J. (1986). The role of maternal input in the acquisition process: The communicative strategies of adolescent and older mothers with the language learning children. Paper presented at the Boston University Conference on Language Development, Boston.

Warren-Leubecker, A. and Bohannon III, J. N. (1984). Intonation patterns in child-directed speech: Mother-father differences. *Child Development*, 55:1379–1385.

Weist, R. M. and Zevenbergen, A. A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, 4(4):291–308.

Wells, G. (1981). *Learning through interaction: The study of language development*. Cambridge University Press, Cambridge.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.

Yurovsky, D., Yu, C., and Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: scaffolding from child-directed speech. *Frontiers in Psychology*, 3.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Dutch Summary

Wanneer mensen de vaardigheid verwerven om taal te begrijpen en te produceren, moeten ze veel taken oplossen die sterk met elkaar verweven zijn. Ze moeten bijvoorbeeld continue spraak opdelen in discrete eenheden, en een betekenis toekennen aan die eenheden. Het beschikbare bewijsmateriaal toont aan dat kennis verworven door vooruitgang in één taak de vooruitgang in andere domeinen in stand houdt en vergroot – of bootstrapt.

We maken computationele modellen van bootstrapprocessen in taalontwikkeling, die informatie uit één domein gebruiken om taken uit een ander domein op te lossen. In drie studies modelleren we eerst hoe kinderen kennis gebruiken uit het perceptuele domein om taalkundige eenheden te ontdekken in niet-gesegmenteerde spraak. Hierna volgt een vierde studie die modelleert hoe volwassen luisteraars middelen gebruiken uit het domein van normaal gehoor om categorieperceptie met een cochleair implantaat op te lossen.

Met betrekking tot het eerste selecteren we *chunks*, zowel op woord- als op lettergreepniveau, uit grote corpora van spraak gericht aan kinderen, en voorspellen we de leeftijd waarop kinderen voor het eerst woorden produceren, gebaseerd op het aantal chunks waarin elk woord voorkomt. Deze aanpak veronderstelt dat als een bepaald woord in veel niet-gesegmenteerde chunks voorkomt die zijn opgeslagen in het langetermijngeheugen van kinderen, dat kinderen dat woord dan makkelijk zouden moeten ontdekken, en het vroeg in hun ontwikkeling zouden moeten beginnen te gebruiken.

We tonen aan dat korte sequenties, in tegenstelling tot frequente of intern voorspelbare sequenties, het meest geschikt zijn voor het voorspellen van het aanleren van woorden. Bovendien hebben korte sequenties een grotere kans om woorden te zijn – wat suggereert dat de vroege protolexicons van kinderen korte, woordachtige chunks

bevatten.

Na het onderzoek rond chunks beschrijven we een laatste studie die de spraakverwerking simuleert van volwassenen met cochleaire implantaten (CI's) – neurale protheses die worden gebruikt om een beschadigd binnenoor deels te vervangen. Veel CI-gebruikers moeten de overgang maken van het verwerken van signalen in hoge resolutie, aangeleverd via het binnenoor, naar signalen in lagere resolutie, aangeleverd via het implantaat.

We modelleren dit via diepe neurale netwerken, en leggen de focus op interferentie tussen aangrenzende kanalen in CI's (*kanaalinteractie*). We ondervinden dat neurale netwerken die eerst getraind zijn op spraak in hoge resolutie, voorafgaand aan training op data in lage resolutie, trager leren wanneer gesimuleerde kanaalinteractie aanwezig is in input in lage resolutie. De spectrale degradatie veroorzaakt door kanaalinteractie kan dus bijkomende afstelling van bestaande neurale circuits vereisen, wat de overgang naar CI's na normaal gehoor vertraagt.