Building an Auditory Lexicon

Samuel David Jones

Supervisors: Dr Silke Brandt and Dr Patrick Rebuschat

Department of Linguistics and English Language
Lancaster University

Thesis submitted for the degree of

Doctor of Philosophy

January 2020

For Emma and Yuma

Declaration

The contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This thesis does not exceed the permitted maximum of 80,000 words including footnotes and appendices. All Lancaster University ethical requirements were strictly upheld throughout this project.

Samuel David Jones

Authorship statement

This thesis contains published work conducted in collaboration with my supervisor Dr. Silke Brandt between May 2017 and July 2019. The thesis is being submitted under the Alternative Format framework to enable me to build a publication portfolio prior to graduation. Each of the empirical studies included in this thesis have been accepted for publication in peer-reviewed journals (see *Published papers*). I conceived, conducted, and wrote up each of the studies presented. Dr. Brandt contributed to the design of the studies presented in chapters four, five, six, and eight, and provided comments on the text for all studies.

Samuel David Jones

Acknowledgements

Thank you to my supervisors Silke Brandt and Patrick Rebuschat. It has been a pleasure to study with you. Thank you to Barbara Dodd, Shelley Gray, Eileen Haebig, Suze Leitão, Franck Ramus, Phaedra Royle, Richard Schwartz, Bill Wells, Cori Williams, and Jennifer Windsor for their feedback on the first study of this thesis. Thank you to Daniel Quintana, Andy Field, and Wolfgang Viechtbauer, for providing suggestions on methodology, and to Paul-Christian Bürkner for advice on Bayesian model fitting using the brms package in R. Thank you to Gert Westermann and Sam Kirkham for their comments on earlier versions of these papers, and to Ben Ambridge for inspiring discussions about exemplar models. Finally, thank you to my family, Saori, Emma, and Yuma, for your support and patience over the last three years.

Published papers

Chapter Four

Jones, S. D., & Brandt, S. (2018). Auditory lexical decisions in developmental language disorder: A meta-analysis of behavioral studies. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1766–1783. https://doi.org/10.1044/2018_JSLHR-L-17-0447.
Repository: https://osf.io/2cvnm/

Chapter Five

Jones, S., & Brandt, S. (2019a). Do children really acquire dense neighbourhoods? *Journal of Child Language*, 1–14. https://doi.org/10.1017/S0305000919000473
Repository: https://osf.io/zfy2p/

Chapter Six

Jones, S. D., & Brandt, S. (2019b). Neighborhood density and word production in delayed and advanced learners. *Journal of Speech, Language, and Hearing Research*, 1–8. https://doi.org/10.1044/2019_JSLHR-L-18-0468.
Repository: https://osf.io/p8ax4/

Chapter Seven

Jones, S. D. (2019). Accuracy and variability in early spontaneous word production. *First Language, 40*(2), 128–150. Repository: https://osf.io/w9y27/

Chapter Eight

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. *Cognitive Science, 44*: Article e12812. Repository: https://osf.io/2qk5j/

Abstract

The ability to form mental representations of word sounds is central to language comprehension and production, and provides the basis of grammatical development and literacy. However, this process remains poorly understood. The five empirical studies presented in this thesis address open questions related to the formation and use of word sound memories. Chapter four presents a meta-analysis of studies using the auditory lexical decision task to measure the quality of word sound representations in typically developing children and children with DLD (developmental language disorder). This chapter also provides baseline data and recommendations of use to both researchers and clinicians. The study presented in chapter five uses Bayesian multi-level regression to model large-scale parental report data from CDI’s (communicative development inventories). This study provides insight into the role that high phonological neighbourhood density plays in early word production, though not comprehension, relative to factors including word length, frequency, concreteness, and relevance to infants. The study in chapter six uses the same methodology to evaluate individual differences in the importance of neighbourhood density as a predictor of word production, and presents results with implications for the development of clinical interventions. The study in chapter seven presents a quantitative corpus analysis examining spoken word accuracy and variability rates in typically developing children recorded over a three-year period. I report the effects of age, frequency, and neighbourhood density on accuracy and variability rates, and argue against the view that such rates may provide a reliable marker of speech sound disorder. Finally, in chapter eight, I present a neural network simulation of the high neighbourhood density learning advantage reported in studies two, three, and four, and present an account of network performance that can also accommodate contradictory behavioural evidence of low-density word learning advantages. All studies are integrated within an exemplar-based framework of auditory-lexical development emphasising the mechanism of analogous generalisation. In the interests of transparency this thesis is accompanied by an online repository containing pre-registration protocols and the data and code required to reproduce each analysis: osf.io/u3qsc.

Contents

List of Figures

List of Tables

Chapter 1 General Introduction

1.1 The problem

One of the defining characteristics of early language development is a bias towards learning words from dense phonological neighbourhoods, i.e. words that sound similar to many other words (Stokes, 2014; Storkel, 2004; Vitevitch & Storkel, 2013). This observation might seem trivial, but it is counterintuitive. You are, for instance, more likely to remember that one person you saw dressed as a clown on the underground during your commute to work than any of the hundreds of people you saw in suits. Similarly, try remembering the following words: *Match*, *Mitch, clinch*, *hitch*, *penguin, finch*, *stitch*, *watch, touch*. It is likely that *penguin* stood out from the list, and that you found this word easier to remember. Such distinctiveness biases – also known as isolation effects – are well documented in cognitive science, where they are commonly explained in terms of greater attention prompting greater depth of processing and this in turn supporting better encoding in memory (Hunt & Worthen, 2006).

The robustness of the association between stimulus distinctiveness and successful memorisation prompts the following question, which is at the centre of this thesis:

Why do young children learn dense rather than distinctive words?

This thesis is not the first attempt to tackle this question. The high-density bias in early word learning has previously been explained in terms of short-term memory advantages (Hoover, Storkel, & Hogan, 2010; Stokes, Kern, & Dos Santos, 2012), long-term memory advantages (Storkel, 2004), the relative ease of determining the meanings of dense words (Smith & Yu, 2008), verbal practice effects (Čeponienė & Keren-Portnoy, 2011), and sensitivity to sub-lexical sound-pattern probabilities (Fourtassi, Bian, & Frank, 2018). This thesis does, however, present a novel answer to the question above, namely:

Because the auditory lexicon is built through analogous generalisation.

The notion of analogous generalisation is developed fully in chapter two, in which I also review literature relevant to the specific aims of the five empirical studies presented in this thesis. In short, three closely related processes are essential to the theory proposed. The first is the memorisation of auditory words from the ambient language. The child hears the words *catch*, *hat*, *mat*, *can*, *sat*, *match*, and *bat* in the speech around them and these words are encoded in memory, giving the child an implicit understanding of the sound structure of their target language. The second process is a short-term memorisation advantage for high-density target words with sound features consistent with those frequently encoded. The child hears the novel target word *cat*, for instance, and this is held in short-term memory easily because it contains familiar sounds; namely those attested in *catch*, *hat*, *mat*, *can*, *sat*, *match*, and *bat*. Note that this supporting neighbourhood provides the basis of analogous generalisation to *cat w*hether or not the child can understand or produce these words (i.e. whether the neighbourhood constitutes implicit or explicit knowledge). The third process is the explicit long-term memory advantage that this short-term memory advantage entails. Continuing the example, the word *cat* is held in short-term memory easily and this supports the formation of a well-detailed long-term memory trace. The child has learned the word *cat* through analogous generalisation.

Two points of clarification are necessary. First, when I say that the theory I propose is novel, I acknowledge that previous studies have identified components of the account outlined here as integral to the high-density bias, perhaps most notably short-term memory advantages (e.g. Hoover et al., 2010; Stokes et al., 2012). My claim, however, is that existing explanatory accounts of the high-density bias are fragmentary, and that when defined in terms of a combination of fundamental processes – implicit, short-term, and long-term memory advantages – analogous generalisation provides a powerful unified framework for explaining this bias in the face of isolation effects (i.e. distinctiveness-driven learning advantages). Despite a large literature on analogy-based learning, including both mathematical and verbal models of early language development (Ambridge, 2019; Johnson, 1997; Skousen, 1989), I am not aware of any theoretical account describing the high-density bias central to early word learning in these terms.

Second, my claim is not that analogous generalisation across dense neighbourhoods is the only factor driving growth of the emerging auditory lexicon. As emphasised throughout this thesis, children are likely to remember the sounds of words that are, for instance, highly frequent, highly concrete, and highly relevant to their lives. The learning preference for high-density words may be overridden when any of these conditions are met. Just imagine a child readily remembering the phonologically complex name of their favourite monster or alien cartoon character. Rather my claim is that all else being equal high-density words will be acquired more easily than low-density words because they are easier to generalise to.

The account outlined so far provides the thread linking the five empirical studies that follow. However, each empirical study comes with its own set of goals and methodology, and aims to make its own contribution to the field. Open science principles are upheld throughout this thesis, with pre-registration protocols and all data and code required to produce each analysis available via an online repository: osf.io/u3qsc. The empirical studies are as follows.

In chapter four, *Auditory lexical decisions in developmental language disorder*, I present a meta-analysis of studies testing children with developmental language disorder in the auditory lexical decision task. In this task, participants are required to provide a ‘yes’/‘no’ or non-linguistic response (i.e. button press) to identify auditory word and non-word stimuli. The advantage of the auditory lexical decision task over related tasks such as naming or non-word repetition is that it minimises the possibility that an observed deficit (e.g. heightened inaccuracy) is attributable to expressive or oral-motor factors. The contribution of this study is to provide summary effect size estimates for accuracy and response time measures for comparisons to age- and language-matched control groups, and to make concrete recommendations for future research and clinical practice.

Chapter five, *Do children really acquire dense neighbourhoods?*, presents results from a Bayesian hierarchical regression model in which rates of word understanding and production among 300 children aged 12 to 25 months were predicted by: (i) phonological neighbourhood density, (ii) frequency, (iii) word length, (iv) babiness rating, (v) concreteness, (vi) valence, (vii) arousal, and (viii) dominance. The contribution of this study is to examine the effect of high phonological neighbourhood density on both word understanding and word production when neighbourhood density is modeled alongside a large inventory of predictor variables. This analysis is valuable because prior studies investigating the high-density bias have modeled neighbourhood density alongside only a handful of variables (e.g. frequency or length), and properties that appear to facilitate acquisition in relative isolation may prove to have only a limited impact when considered alongside a more representative range of explanatory factors (Braginsky, Yurovsky, Marchman, & Frank, 2019).

In chapter six, *Neighbourhood density and word production in delayed and advanced learners*, I use a similar methodology to examine individual differences in the importance of ambient-language phonological neighbourhood density as a predictor of word production in 442 children aged 18-months, with productive lexicon sizes between zero and 517 words. The contribution of this study is to re-examine the hypothesis that a difficulty forming memories of words comprising uncommon sound sequences (i.e. low phonological neighbourhood density words) is a central determinant of delayed expressive vocabulary development (e.g. Stokes, 2014).

Chapter seven, *Accuracy and variability in early spontaneous word production*, examines factors explaining rates of accuracy and variability in 244,459 spontaneous word productions from five typically developing children recorded over a three-year period (0:11-4;0). High rates of error and variability in early word production have been proposed as a marker of speech sound disorder (e.g. Holm, Crosbie, & Dodd, 2007), however this approach has been challenged by studies reporting high rates of error and variability in the typically developing population (Sosa, 2015). Chapter seven evaluates this debate, and provides recommendations for future research and clinical practice. In addition, I report the effects of age, frequency and neighbourhood density on accuracy and variability rates.

The overarching theory presented in this thesis is that the auditory lexicon is built through analogous generalisation across dense neighbourhoods. One challenge to this account comes from behavioural evidence that high-density words are also more likely to be misunderstood as instances of known words, and therefore that high-density may under certain circumstances complicate the learning process (e.g. Swingley & Aslin, 2007). In chapter eight, *Density and distinctiveness in early word learning*, I present a neural network simulation illustrating how these apparently contradictory density and distinctiveness effects can emerge from a common learning mechanism.

Chapter nine includes an integrated discussion of these empirical studies, and revisits the overarching principle of analogy-based learning driving growth of the auditory lexicon. This chapter also summarises the major contributions of this thesis and outlines directions for future research.

The study of early word learning provides a springboard to many vital questions; from the identification and treatment of speech and language disorders, to the architecture of the brain and mind, to the engineering of machine intelligence. My hope is that this comprehensive account of the development of the early auditory lexicon is of interest to both researchers and clinicians.

Chapter 2 Background

It is almost customary for studies of early word learning to begin with a description of Quine's (1960) *gavagai* problem. In this thought experiment, a linguist studying an undocumented language hears the word *gavagai* shouted by a native speaker as a white rabbit runs passed. What does the linguist infer about the word *gavagai*? Quine's problem encourages us to think about the subtle challenges facing young word learners. Does *gavagai* refer to the rabbit as a whole or to a particular part of the rabbit? Does it mean *white* or *running*? Or might it mean *catch it!* Existing solutions to this problem – often called the problem of indeterminacy – fall on a continuum from nativist accounts emphasising innate biases such as the assumption that labels extend to whole objects, to constructivist accounts emphasising socio-pragmatic cues such as shared attention, the ability to follow speaker gaze, and the ability to infer speaker states of mind (Ambridge & Lieven, 2011).

Research related to the problem of indeterminacy dominates developmental cognitive science and the study of early word learning more specifically. However implicit in the *gavagai* thought experiment are a series of more primary problems, which the learner must overcome prior to inferring the meaning of the spoken word *gavagai* and its appropriate patterns of use. Perhaps most general is the question of how the learner develops a mental representation of the spoken word *gavagai.* Crucially, this word-sound representation needs to be flexible enough to support accurate recognition when the word is produced by different speakers and at different rates, while at the same time detailed enough to support accurate production by the learner. The overarching aim of this thesis is to develop a comprehensive account of this process. In short, I will argue that the auditory lexicon is built principally through analogy-based learning. However, the five empirical studies in this thesis cover a lot of ground. Staying with the *gavagai* example, this thesis will address questions including:

1. Is the word *gavagai* easy or difficult to remember and to produce relative to, say, the word *elephant*? Would the word *gavagai* be easier or more difficult to remember and produce if the target language contained many or few similar sounding words?

2. With respect to overall learnability, how important is the sound of the word *gavagai* relative to alternative factors including the concreteness of the inferred referent (i.e. a rabbit), the relevance of the label-referent pairing to the life of the learner, and the frequency with which the learner hears the label?

3. What does the representation of *gavagai* look like in the mind? And what is the most appropriate simplification of that mental representation that allows us to study and communicate about it? Relatedly, how best can we operationalise the associations between the mental representations of *gavagai* and other known words?

4. How might we measure the quality of the word sound memory of *gavagai*? Is the degree of detail in this mental representation high from the outset, or does it improve with exposures? Does a two-year-old child exposed to *gavagai* form a memory trace of similar quality to that formed by a ten-year-old child?

5. If we exposed a large number of children to the spoken word *gavagai*, we might expect individual differences in retention and production accuracy scores at test. What explains this variation? What is the relation between cognitive and oral-motor development and auditory word memorisation and production? How might neurological disorder impede memorisation and production? And how can we best identify and treat children with language disorder?

This review chapter provides background to the key ideas outlined in the questions above. It develops a theory of early auditory-lexical development centred on the mechanism of analogous generalisation. Each empirical study in this thesis contains its own focused literature review, and the purpose of this chapter is to provide a bird's eye view of the field. The structure of this chapter is as follows. I begin with a comparison of abstractionist, exemplar, and hybrid accounts of word sound representation, before describing different methodological approaches to quantifying the associations between representations. I then evaluate theories of early auditory word learning and discuss how this process may be affected by neurological disorder. Finally, I summarise the explanatory framework linking the five empirical studies that follow, namely the notion that the auditory lexicon is built through analogous generalisation.

2.1 What do words look like in the mind?

2.1.1 Prototypes

A dominant and perhaps intuitive view of word sound memory is that word representations in the mind resemble the phonemic transcriptions accompanying orthographic dictionary entries (see Port, 2007; Ramscar & Port, 2016, for discussion). That is, our memory of the sound of the word *cat*, for instance, might broadly resemble the international phonetic alphabet (IPA) form /kæt/ stored alongside possible meanings and patterns of use. This, put crudely, is the view of prototype theories of word sound representation (see Ambridge, 2019, for review). Two features are essential here. First, auditory word representations in the mind are assumed to be abstractions from the variation inherent in natural speech; a product of factors including speaker age, gender, dialect, and speech rate. Prototype theories entail that such information is jettisoned during acquisition, and see this normalisation process as an advantage enabling learners to generalise word knowledge across speakers and contexts. Second, word prototypes are assumed to be composed of elementary segments, such as phonemes rendered in the IPA or alternate symbolic systems representing articulatory features (e.g. Chomsky & Halle, 1968). These, alongside similarly abstract combinatorial rules, provide the basis of spoken word recognition and the building blocks of spoken word production.

The appeal of prototype theory is that it proposes an apparently efficient mode of word storage supporting cross-contextual recognition and productivity. However, the principles of abstraction and segmentation essential to the prototype framework each face challenges. Perhaps most damagingly, the notion of an abstract word-level prototype is difficult to reconcile with evidence of speaker effects that the process of abstraction should wash out. For instance, recognition memory among adults and children is better when test stimuli are presented in voices similar (e.g. of people of the same biological sex) to that in which they were taught (e.g. Houston & Jusczyk, 2000). A second and closely related challenge for prototype theories of word representation is to specify the form of the relevant prototype given that discrete systems of linguistic representation (e.g. orthographic letters, phonemes, morphemes, words) map poorly to continuous speech (Ramscar & Port, 2016). For instance, the phoneme [æ] – a candidate sub-lexical prototype – may be rendered differently in terms of quality and duration depending on speaker and context. Do learners then form different prototypes of all possible renderings of [æ] or a single prototype with rules linking that prototype to its various possible realisations? Similarly, at the word-level, does the word *butter* spoken in cockney dialect (/bʌʔə/) and received pronunciation (/bʌtər/) correspond to two distinct prototypes or a single prototype with additional transformation rules linking to these different pronunciations (Ambridge, 2019)?

2.1.2 Exemplars

An alternative to the prototype framework known as exemplar theory aims to address these specific criticisms (Ambridge, 2019; Johnson, 1997; Skousen, 1989). Under an exemplar account, spoken word exposures are stored in rich auditory code alongside associated speaker- and context-specific details. Broadly, if under prototype theory word representations are conceived of as *types*, then under exemplar theory word representations are conceived of in terms of clouds of *tokens*. Every spoken word exposure is held to contribute to the on-going development of the auditory lexicon concretely and without abstraction. Units such as phonemes and indices of manner and place relations are therefore considered tools for the study and communication of language science that have no cognitive reality. It is held that word recognition and production involve on-the-fly generalisation across multiple exemplars to the target item rather than, for instance, the search for a uniqueness point corresponding to a unique prototype or the placing together of abstract segments (e.g. phonemes) according to abstract combinatorial rules to form a motor plan for production (Ambridge, 2019).

Notable objections to exemplar theories of language development are that it is implausible that the mind stores such a large amount of information, and also that the framework sits uncomfortably with dominant models of memory, for instance the semantic/episodic divide. That is, exemplar-based word learning is often associated with episodic memory (i.e. spoken word exposures are stored as rich episodes), while many existing word learning frameworks focus on the semantic memory system (i.e. the division of long-term memory dedicated to abstract ideas and concepts). This criticism might be particularly damaging to exemplar theory as applied to early word learning, given evidence that episodic memory may not come fully online until after the onset of word comprehension and production (Ghetti & Lee, 2011). However, proponents of the exemplar framework point out that each of these criticisms suffers due to on-going controversy in the basic literature (Ambridge, 2019). There is disagreement, for instance, regarding the amount of information the mind can store (note also that exemplar theory accommodates forgetting), and the classification of memory systems (e.g. procedural, episodic, semantic, etc.) and more importantly the specific role these systems play in early language development remains an area of considerable debate.

More serious for exemplar theory is the charge that the analogous generalisation mechanism essential to the framework remains underspecified. In lieu of a satisfactory account of how this mechanism operates in the absence of abstract information, it is perhaps unsurprising that some theorists, advocates of so-called hybrid models, argue that both abstract and prototypic representations must be involved in language representation and use. Pierrehumbert (2016), for instance, acknowledges that strict abstractionist models are too minimal to account for evidence of speaker effects, but nevertheless defends the principle that the ability to generalise during both spoken word recognition and production necessitates abstract prototypes including phonemes and combinatorial rules. From an exemplar-based perspective, however, the significant limitation of the hybrid position remains that it is hard to specify the form that such prototypes may take in the face of the variation inherent in natural speech (e.g. in idiolect, dialect, sociolect, and speech rate and loudness).

2.1.3 Section summary

Evidence of sensitivity to speaker effects apparently rules out strict abstractionist accounts of auditory word representation (section 2.1.1), while the difficultly of specifying functional prototypes at any level of linguistic representation stands against both abstractionist and hybrid accounts (section 2.1.2). In contrast, exemplar-based theories of word sound representation find relatively good support in the existing language development literature, for instance studies reporting speaker effects (Houston & Jusczyk, 2000). Work is, however, required to clarify how the exemplar framework relates to dominant models of early memory and language development. It will also be valuable to develop the verbal account of analogous generalisation in order to supplement existing mathematical and computational models of this process (e.g. Johnson, 1997; Skousen, 1989). Each of these aims is pursued throughout this thesis.

2.2 Association networks

2.2.1 Lexical competition effects

The current review has so far described auditory word representations in isolation. However, word representations interact to affect target word processing on a range of paradigms (see Weber & Scharenborg, 2012, for review). Interestingly, contrasting effects are observed whether the experimental outcome of interest is word production or recognition. Longer recognition latencies and higher error are observed for target words that sound similar to many other known words, theoretically because a large cohort of mental representations sharing initial auditory features becomes activated and significant target word sound information is required to reach an identification point (e.g. the cohort model; Marslen-Wilson & Tyler, 1980; Sommers & Lewis, 1999; Vitevitch & Luce, 1998). Conversely, retrieval and production advantages are observed for words that sound similar to many other words, a finding that has been explained in terms of both articulatory practice effects (i.e. words that sound similar to many other words are more likely to be regularly produced) and excitatory feedback between lexical and phonemic levels of processing (see Dell & Gordon, 2003, p. 11).

The degree of auditory similarity between words therefore has important implications for accurate and rapid word recognition, retrieval, production, and – as detailed at length below – successful word acquisition. Various approaches to quantifying degrees of auditory word-form similarity exist. Vitevitch (2008), for instance, applied the principles of network analysis (i.e. the degree metric and clustering coefficients) to the study of word retrieval. However, a dominant approach has been to use measures of string edit distance such as Levenshtein distance based on IPA or similar representations of target words (e.g. SAMPA; the Speech Assessment Methods Phonetic Alphabet). This metric – broadly termed phonological neighbourhood density – is central to four out of five of the empirical studies in this thesis, and is therefore discussed in detail in the remainder of this section.

2.2.2 Phonological neighbourhood density

2.2.2.1 The plus-minus-one-phoneme criterion

Phonological neighbourhood density constitutes a general principle rather than fixed operational definition. Broadly, words that sound similar to many other words are termed high density, and words that sound similar to few other words are termed low density. Specific operational definitions of neighbourhood density differ between studies. However, following Luce and Pisoni (1998; experiment two), the dominant approach in adult and child language research has been to define neighbourhood density as the number of words in a given corpus that can be formed by the addition, substitution, or elimination of a single phoneme in a target word. Under this so-called plus-minus-one phoneme metric, the word *cat*, for instance, neighbours *catch, hat,* and *at*. A significant limitation of the plus-minus-one-phoneme method is that there is no grading of phonological distance within neighbour and non-neighbour categories. For instance, the words *bag*, *map*, and *hippocampus* are all non-neighbours of the target *cat*, despite different degrees of phonological similarity from this target. Conversely, *cat* neighbours both *can* and *hat*, suggesting equivalent phonological distance between these items. In this way, the use of plus-minus one neighbourhood density entails the loss of information about degrees of word-level phonological similarity between words.

Such loss similarly occurs at the corpus level, and this may be particularly damaging in child language research or when quantifying the phonological structure of small corpora more generally. While at the word-word level the plus-minus one criterion is categorical – that is, two words are either neighbours or not – a graded picture of phonological structure should emerge through density counts across the corpus, i.e. some words will have greater neighbour density counts than other words. In adult samples – the population for which the plus-minus one criterion was originally developed (Luce & Pisoni, 1998) – there may be a large range of positive density counts across the corpus (e.g. some words may have one neighbour while others will have hundreds of neighbours) and the number of frequently produced words with zero neighbourhood density may be very low. However, because young children know considerably fewer words than adults, a significant proportion of words in a child corpus may be ascribed zero neighbourhood density, while the range of positive density counts across the corpus may be limited (Fourtassi, Bian, & Frank, 2018). For instance, in a preliminary analysis conducted during the development of pre-registration protocols for the empirical studies presented in this thesis, I found that

48.31% of words listed in the UK communicative development inventory (UK-CDI; Alcock, Meints, & Rowland, 2017) had zero plus-minus-one-phoneme neighbourhood density, with a range of zero to nine neighbours. Thus a substantial proportion of words in a representative child lexicon – including frequently produced words – constitute lexical hermits under a plus-minus-one-phoneme criterion of neighbourhood density.

2.2.2.2 The plus-minus-two-phoneme criterion

One suggested way to address information loss when quantifying auditory similarity structure is to increase string edit distance from one to two phonemes (Fourtassi et al., 2018). Under a plus-minus-two phoneme metric of word similarity the target *cat*, for instance, not only neighbours *hat* and *can*, but also *bag* and *map*. Adopting a plus-minus-two-phoneme criterion of similarity considerably reduces the number of words ascribed zero neighbourhood density within a given corpus, while expanding count ranges across the corpus. For instance, in the preliminary analysis described above, I found that an average of 12.92% of words listed in the UK-CDI (Alcock et al., 2017) had zero plus-minus-two-phoneme neighbourhood density, with a range of zero to 63 neighbours across the corpus. Information loss therefore appears reduced relative to the plus-minus one phoneme criterion, resulting in a dataset that may be relatively more powerful in statistical analysis (i.e. higher sensitivity, inferences may be made with fewer cases). However, the usefulness of two-phoneme neighbourhood density remains questionable because the perceptual similarity of words classed as neighbours under this criterion may not be immediately clear, particular for short words with few phonemes such as *cat* and *bag*, which dominate the emergent lexicon. Note, for instance, that other two-phoneme neighbours of *cat* include (via *caught*): *cawed, bought*, and *fought*. Furthermore, in the recent manuscript employing this approach (Fourtassi et al., 2018), words are apparently scored as neighbours whether they are one or two phonemes from a target, as in the case of *can* (one phoneme) and *bag* (two phonemes) relative to the target *cat*.

Information loss therefore remains an issue for categorical measures of phonological similarity even when string edit distance is increased.

2.2.2.3 Continuous measures of word-form similarity

One way to alleviate these issues is to adopt a continuous criterion of word-form similarity. Continuous measures of word-level phonological similarity may be operationalised in a number of ways. In two out of the five empirical studies that follow I used a continuous metric of word form similarity called phonological Levenshtein distance, or PLD20, defined as the mean number of additions, substitutions, or eliminations of phonemes required to change a particular word into its nearest twenty phonological neighbours (Suárez, Tan, Yap, & Goh, 2011, p. 606). In contrast to classic definitions and operationalisations of word-form similarity (e.g. Luce & Pisoni, 1998), a smaller PLD20 indicates greater phonological similarity (i.e. high density), while a high PLD20 indicates greater phonological distance (i.e. high distinctiveness). The major advantage of this continuous measure is that every word in any given corpus is attributed a density value, and this supports the identification of neighbourhood effects – such as the inhibition and facilitation effects described at the beginning of this section – for words that would be classed as lexical hermits under a categorical criterion (Suárez et al., 2011).

Despite advantages of decreased information loss and associated increases in predictive power, continuous measures of word-level phonological similarity have not been widely adopted in developmental research, where their application may be especially useful because young children know few categorical neighbours. Criterion selection is ultimately question dependent, and researchers may have justification to adopt a categorical measure of word-form similarity. The question of criterion may arguably matter less, for instance, when quantifying large-scale input corpus densities, where frequently produced words might have many attested categorical neighbours. Generally speaking, however, the use of categorical criteria of phonological word-form similarity may be unwarranted, particularly given the recent development of software packages supporting computation using continuous criteria and open-source datasets listing pre-computed values (e.g. stringdist; van der Loo, 2014).

2.2.3 Section summary

In the previous and current sections, I described auditory word representations in terms of clouds of exemplars, which may be characterised as forming complex association networks on the basis of auditory distance. The use of IPA word representations and ideally continuous measures of string edit distance (e.g. PLD20) constitute dramatic simplifying approaches essential to studying and describing the mental lexicon. In the following section I situate this system in a developmental framework, with special emphasis on development prior to literacy.

2.3 Building an auditory lexicon

2.3.1 Auditory-linguistic sensitivity in newborns

One of the astonishing claims of developmental cognitive science is that auditory word learning starts in utero (Partanen et al., 2013). The fetal brain undergoes dramatic changes, including extensive synapse production (i.e. an increase in the number of connections between neurons), the myelination of axons (i.e. the insulation of neuron projections), and the organisation of the auditory cortex in response to external stimuli. Such plasticity is continuous with a capacity to learn before birth, and the sounds coming through the intrauterine walls, including the sound of caregiver speech, provide a dominant stimulus prompting learning during this time. Accordingly, a number of studies have demonstrated newborn sensitivity to properties of the ambient language, including listening preferences (e.g. identified using a sucking rate habituation paradigm) to the target language, the mother's voice, and storybooks read during pregnancy (see Aslin, Jusczyk, & Pisoni, 1998, for review). The important observation here is that substantial information about the sound structure of the target language is learned prior to the onset of word learning as commonly defined, for example as the ability to recognise or label a white rabbit as *gavagai*.

2.3.2 First words

Infants commonly understand a limited number of words by six to nine months (Bergelson & Swingley, 2012), with first words emerging around the first birthday and exponential growth in productive lexicon size thereafter (Tomasello, 2005). As in most areas of development, there are substantial individual differences in early receptive and productive vocabulary size. For instance, 18-month-old children in the 95th centile may produce up to 240 words, while age-matched peers below the 10th centile may produce as few as five words (Alcock et al., 2017). Nevertheless, despite common differences in size, the content of children's early lexicons is remarkably consistent between children both within and across language communities (Braginsky, Yurovsky, Marchman, & Frank, 2019). To demonstrate this, Braginsky et al. (2019) modeled ages of acquisition for 400 words recorded in large-scale parental report data as a function of a range of predictor variables previously associated with variance in learning outcomes, for instance word length, frequency, concreteness, and relevance to babies and infants. Results indicated that, across ten languages, early-acquired words tended to be short, high frequency, highly concrete, and highly relevant to the lives of babies and infants.

2.3.3 Phonological neighbourhood density and the early lexicon

Braginsky et al. (2019) emphasise that the relatively comprehensive list of predictors included in their statistical model of age of acquisition is incomplete. Accordingly, to build on this work in the interest of better understanding the development of the auditory lexicon, Jones and Brandt (2019a, chapter five) fitted a modified version of Braginsky et al.'s (2019) model with the addition of ambient language phonological neighbourhood density as a predictor variable. The aim of this analysis was to determine the importance of neighbourhood density relative to other predictors of word comprehension and production (e.g. frequency, concreteness, babiness) in 300 children aged 12;0 to 25;0. We considered this an important analysis given the insights the study of neighbourhood density effects has provided into auditory word representation and association network growth. Jones and Brandt

(2019a, chapter five) replicated Braginsky et al.'s (2019) major findings, reporting learning advantages for high-frequency, concrete words with high relevance in infancy and early childhood. In addition, we reported that high phonological neighbourhood density strongly predicted word production but not word comprehension – with which frequency, concreteness, and relevance to babies were more strongly associated. The results also suggested that the high-density word production advantage was stronger in younger children. Each of these findings is consistent with prior work reporting that high-density words are learned developmentally earlier and on fewer or noisier experimental exposures than low-density words, and, furthermore, that high neighbourhood density appears to confer specific advantages on word production (e.g. Stokes, 2010; Stokes, Kern, & Dos Santos, 2012; Storkel, 2004, 2006, 2011; Takac, Knott, & Stokes, 2017; Vitevitch & Storkel, 2013; Vitevitch, Storkel, Francisco, Evans, & Goldstein, 2014). Also of relevance here is the observation that word production accuracy and stability are often better for high neighbourhood density words, with low-density words commonly produced inaccurately and inconsistently (i.e. differently aross multiple productions; e.g. Sosa & Stoel-Gammon, 2012; Jones, 2019, chapter seven).

The early high-density word learning advantage is intriguing in light of the aforementioned high-density word competition effects observed during recognition tasks (Weber & Scharenborg, 2012), and similarly in light of the distinctiveness advantages – i.e. isolation effects – described in the *General Introduction*. On the other hand, high-density has been associated with word production advantages (Weber & Scharenborg, 2012), and Jones and Brandt (2019a, chapter five) similarly report a substantial production though not comprehension effect in early acquisition. Together, these findings suggest that it might be more accurate to talk about a high-density expressive lexicon advantage, rather than a high-density word learning advantage per se. Putting this question aside for now, it is apparent that a bias towards the acquisition of high-density words is a defining characteristic of the emerging auditory lexicon, though the mechanism underlying this advantage remains poorly understood (Gierut & Morrisette, 2012). The aim of this thesis is to propose a theory of the high-density bias centred on the principle of analogous generalisation. In the paragraphs that follow I begin to build the account of this process that will ground the empirical studies that follow.

2.3.4 Building an auditory lexicon through analogous generalisation

It is first important to clarify the idea of generalisation in the context of auditory word learning. Studies of early auditory word learning commonly refer to the acquisition of dense neighbourhoods. This is true, for instance, of chapter five of this thesis and the landmark study by Storkel (2004) from which that chapter takes its name (*Do children acquire dense neighbourhoods?*). This title may, however, be somewhat misleading. For instance, it has been shown by Schwartz and colleagues that young children acquire and produce test words that contain sounds that the child has previously produced more easily than test words containing previously unattested sounds (e.g. Schwartz & Leonard, 1982; Schwartz, Leonard, Frome Loeb, Swanson, & Loeb, 1987). Nevertheless, this is often not what is meant by the acquisition of dense neighbourhoods. Instead, many studies, including those of the current thesis, demonstrate that children learn target words that are high density in the input language, whether or not they have explicit knowledge of those words' phonological neighbours (i.e. whether or not they are able to recognise and produce those neighbours).

This distinction is made clear in the aforementioned study by Fourtassi et al. (2018), in which two hypotheses of early auditory word learning are compared. The first hypothesis is of a learning preference for novel auditory words associated with many other words in the child's lexicon (i.e. internal connectivity; as in the work by Schwartz and colleagues cited above). The second hypothesis is of a learning preference for novel auditory words associated with many other words in the input or ambient language (i.e. external connectivity; as in Storkel, 2004). Using a network growth model, Fourtassi et al. (2018) provide evidence that, outside of the lab, phonological representation networks grow predominantly on the basis of external connectivity. That is, Fourtassi et al. (2018) argue it is connectivity in the ambient language, and not connectivity to words in the child's lexicon, that makes high-density target words relatively easy to learn, and which guides the development of the early auditory lexicon.

2.3.4.1 Long-term auditory priming and conspiracy effects

Fourtassi et al.'s (2018) argument requires some qualification. Crucially, for any neighbourhood density learning effect to exist, the phonological association structure of the ambient language would have to be represented in the mind of the child. Without such representation, the high-density bias is seemingly inexplicable. My position – supported by evidence of children's sensitivity to the association structure of the ambient language from the neonatal period (e.g. Church & Fisher, 1998; Goldinger, 1996; Jusczyk, Luce, & Charles-Luce, 1994) – is that it is not word-level phonological connectivity in the ambient language per se that drives auditory-lexical growth, but that early auditory-lexical growth is driven by the representation of ambient language connectivity across spoken word exemplars stored in the mind of the child. This position may be characterised by adopting Church and Fisher's (1998) label "long-term auditory priming". Under this account, spoken word exposures are encoded in memory and this supports the recognition and production of a target word with auditory features identical or similar to primes even after considerable delay, for instance a week (see Fisher, Church, & Chambers, 2004, for an extensive review). Interestingly, such priming effects (i.e. higher accuracy or shorter response time) are reduced when training and test voices differ, as predicted under exemplar accounts of early word learning that emphasise context-sensitive encoding (Ambridge, 2019).

Analogous generalisation may, then, be characterised as the child successfully learning, for instance, the word *cat*, after exposure to words including *catch*, *hat*, *mat*, *can*, *sat*, *match*, and *bat*. Exposure to such words supports analogous generalisation to *cat* whether or not the child has functional knowledge of the words in the supporting neighbourhood (e.g. *catch*, *hat*, *mat*, etc.). That is, the representations that support analogous generalisation to the novel target word may be conceptual, i.e. stored with semantic details including referential information, or they may be perceptual, i.e. stored without semantic details (Fisher et al., 2004). The encoding of a large number of exemplars with similar auditory features (e.g. *catch*, *hat*, *mat*, etc.) may be described as a form of *conspiracy effect* (Rumelhart, McClelland, and the PDP Research Group, 1986), which is a computational modelling term used to describe the process by which the connection weights in a neural network become biased in the direction of frequent input patterns during training, supporting low-error generalisation to similar targets (Jones & Brandt, 2020, chapter eight; see also chapter three).

2.3.4.2 Short-term and long-term memory advantages

Central to the account of analogous generalisation developed here is the idea that this conspiracy effect (i.e. the implicit and explicit encoding of dominant patterns of the auditory ambient language structure; Jusczyk et al., 1994) comes with short-term memory advantages. Short-term memory refers to the temporary storage (seconds to a couple of minutes) of information without transformation. The ability to accurately repeat a phone number, for instance, is a short-term memory skill. Tasks in which transformations are involved, for instance repeating a phone number backwards, tap working memory skills. The dominant model of these faculties is Baddeley & Hitch's (1974) working memory model, and at the core of this model and of its subsequent refinements (e.g. Baddeley, 2000) is a component termed the phonological loop, in which untransformed speech information is held and sub-vocally rehearsed. There is a substantial literature on the role of the phonological loop in early language development (see Vance, 2008, for review). This literature indicates that the ability to successfully repeat non-words or a string of digits is strongly associated with linguistic competence, and that deficits in such tasks are often associated with language-delay (Baddeley, Gathercole, & Papagno, 1998; Gathercole, Hitch, Service, & Martin, 1997; Gathercole & Baddeley, 1990; Gathercole, Service, Hitch, Adams, & Martin, 1999).

Short-term memory is differentiable from long-term memory. While short-term memory involves temporary storage, long-term memory can include representation for the lifespan. Short-term memory does, however, act as a gatekeeper to long-term memory, in the sense that stimuli such as words that are held in short-term memory accurately will be passed to long-term memory and stored there in greater detail (Vance, 2008). In addition, the established lexicon – defined as the total store of spoken word exemplars in long-term memory – confers top-down effects on the short-term storage of spoken words. That is, target high-density words that contain sound patterns attested in many other exemplars stored implicitly and explicitly in long-term memory are held in short-term memory more accurately than low-density words, and this in turn supports the subsequent formation of highly detailed long-term word memory traces (e.g. Gathercole, Frankish, Pickering, & Peaker, 1999).

2.3.4.3 Word representation and production

The auditory lexicon is therefore built by the learner implicitly following high-density pathways through the ambient language. Analogous generalisation may be characterised specifically in terms of the short- and long-term memory advantages that occur given a large number of stored spoken word exemplars with close proximity to the target word (Gathercole et al., 1999). The specific production advantages for high-density words reported above – e.g. earlier age-of-first-production and heightened production accuracy and stability – then emerge as a by-product of this primary cognitive advantage. Many variables support entry to the receptive lexicon. For instance, young children are highly likely to recognise concrete words that they hear frequently and that are highly relevant to their lives, for instance the word *pushchair*. Words with such characteristics may be recognised despite complex phonology. However, entry to the productive lexicon is dependent on the ability to form an accurate motor plan, understood not as an abstract representation but as a generalisation made across existing auditory word exemplars on the fly (Ambridge, 2019). Fuzzy stored exemplars of words with complex phonology may not be amenable to early production, leaving such words anchored initially in the receptive lexicon. As the range of stored exemplars increases as a function of language exposure, however, children become better able to represent and use words with relatively complex phonology; a factor reflected in the weakening of the association between high neighbourhood density and word production over time (Jones & Brandt, 2019a, chapter five).

One criticism of this account might concern how we know the high-density productive advantage is cognitive and not oral-motor in nature. For instance, could it not be that the early exemplar representations of *pushchair* are in fact highly detailed, but that the child's immature oral-motor skills simply prevent them from producing this word accurately? This is an apparently viable position because oral-motor skills do develop markedly during the early years, and there is evidence that these skills confer effects on language ability independent of the child's general cognitive ability (Alcock, 2006). Nevertheless, there is also strong evidence that when you remove the necessity of a verbal response from the word processing task demands, as in the auditory lexical decision task described at length in chapter four of the current thesis, performance still varies as a function of age and language proficiency (Jones & Brandt, 2018, chapter four; see Claessen, Heath, Fletcher, Hogben, & Leitão, 2009, for review). Similarly, clinical evidence indicates that children diagnosed with expressive language deficits usually also score poorly on measures of receptive language, a factor prompting Leonard (2009) to question the validity of pure expressive language deficit as a diagnostic category.

Separating out oral-motor and representational accounts of both typical and atypical development remains an important challenge. There is little doubt that oral-motor skills develop in early childhood, and that production practice effects may sharpen the representation of word sounds. Early-learned, high-density words contain sounds that are produced more frequently and which may require minimal articulatory recourses. However, the existing empirical evidence – particularly the large body of evidence probing early word sound representation quality, and evidence that expressive language deficits are in general attributable to underlying representational problems (see section 2.4) – stands against the idea that the high-density word learning bias central to the development of the auditory lexicon may be explicable in terms of a pure oral-motor effect.

2.3.5 Controversies and complications

2.3.5.1 Emergentism and accessibility

It is important to acknowledge controversy regarding the level of detail in early auditory word representations (see Ainsworth, Welbourne, & Hesketh, 2016, for review). As touched on above, newborns come into the world with a number of auditory-perceptual biases that set the stage for early language development (Aslin et al., 1998) including preferences for speech over non-speech sounds (Vouloumanos & Werker, 2004), infant- or child-directed speech (Cooper & Aslin, 1990), and familiar voices (DeCasper & Fifer, 1980). However, despite these apparently sophisticated auditory-perceptual skills, infants commonly make errors in auditory word recognition that suggest insensitivity to fine-grained word sound details. For instance, infants and young children may fail to identify mispronunciations of known words (Van Der Feest & Fikkert, 2015; cf. White & Morgan, 2008), or may fail to map minimally different non-words including *bih* and *dih* to novel objects (Pater, Stager, & Werker, 2004; Stager & Werker, 1997).

Such findings have motivated emergentist accounts of early auditory word representation, perhaps most prominently the lexical restructuring model (Metsala & Walley, 1998). Under this account, the small size of the early lexicon makes gestalt word sound representation possible, enabling children to focus on establishing a rudimentary lexical base. Growth of the lexicon, however, renders this strategy increasingly implausible and inefficient, and infants then apply their skills of auditory-perception to the task of developing rich word sound representations, which are then organised on the basis of fine-grained phonemic similarity networks. Support for this position comes from experimental tasks assumed to probe the detail of word sound representations (see Claessen, Heath, Fletcher, Hogben, & Leitão, 2009, for review), including the gating paradigm (Grosjean, 1980) in which participants aim to identify auditory target words on the basis of clipped segments of increasing length. Young children typically identify familiar high-density target words upon exposure to shorter segments, suggesting that these items are better detailed (Walley, 1993). Support for the restructuring of the lexicon also comes from studies reporting lexical competition effects – comparable to the adult-study effects reported above – at 24 but not 18 months (see Nivedita & Borovsky, 2018, p. 63, for review). These findings suggest that at 18 months auditory word representations exist in relative isolation, but that by two years of age, perhaps as a function of heightened detail, word sound representations are restructured in similarity neighbourhoods. Note that such competition effects are also present in children younger than two years who have large vocabularies, supporting the idea that restructuring is a product of lexicon size rather than age (Walley, 1993).

One challenge for emergentist accounts such as the lexical restructuring model is to accommodate evidence of sensitivity to minimal changes or aberrations in the target item during lexical perception tasks from as early as 14 months (e.g. Swingley, 2005; White & Morgan, 2008). Swingley (2005), for instance, observed an early preference for accurately produced known words over inaccurately produced known words using a preferential looking paradigm. Such findings have motivated the development of accessibility accounts of early word representation, such as Werker & Curtin's (2005) influential processing rich information from multidimensional interactive representations or PRIMIR model. Under PRIMIR, rich auditory word information is encoded from infancy, but access to this detail during recognition or production varies as a function of developmental stage, exposure context, and experimental task demands. PRIMIR therefore provides a unifying framework accommodating findings underpinning both strict emergentist accounts and strict accessibility accounts of early word representation (Ainsworth et al., 2016). In addition, PRIMIR accommodates findings from the aforementioned literature motivating exemplar theories of auditory word representation, for instance by emphasising the context-sensitive encoding of early word sound exemplars. Furthermore, closely in keeping with exemplar principles, PRIMIR assumes phonemes have no initial cognitive reality, for instance as the building blocks of word representations. Instead, phonemes are assumed to emerge slowly as fuzzy representations from the distributional analysis of stored word exemplars (Ainsworth et al., 2016), before being sharpened as a product of later developing literacy skills including the awareness of orthographic and phonemic associations. The ability to accommodate such factors constitutes an important improvement on the lexical restructuring model, and makes PRIMIR a robust theoretical framework for understanding early word sound representation.

2.3.5.2 Multicolinearity and interactions

The empirical studies that follow illustrate a number ways in which the general account of high neighbourhood density effects and analogous generalisation presented here becomes more complicated. First, as described above, neighbourhood density is just one of many variables associated with variance in early word learning, including word frequency, length, babiness rating, and alternative word sound variables such as phonotactic probability, which is a measure of the co-occurrence probability of a sequence of phonemes. Sub-lexical phonotactic probability effects have been of considerable interest in prior work in this general area, given that the study of such effects can inform understanding of how children learn the sound structure of the ambient language. However, given high levels of correlation between predictors – a factor often resulting in multicolinearity – it is often difficult to include phonotactic probability in statistical models designed to assess early word-level density effects. This unfortunately makes it impossible to determine the influence of the excluded variable. Second, and relatedly, neighbourhood density interacts with other lexical variables, perhaps most importantly word frequency. In particular, there is some evidence that high neighbourhood density is more strongly associated with word production for low-frequency words, with high frequency apparently nullifying this effect (Hollich, Jusczyk, & Luce, 2002; Storkel, 2004). Focused discussions of multicollinearity and interaction effects appear in the empirical chapters that follow.

2.3.5.3 Learning advantages for distinctive stimuli

A further complication to the explanatory account presented here is the finding that high neighbourhood density may in certain contexts impede word learning. As argued in chapter eight of this thesis, understanding this apparent contradiction depends on looking at word learning sub-process (e.g. Leach & Samuel, 2007). For instance, some studies identify an initial *triggering* of learning, in which the mismatch between an auditory target stimulus and stored exemplars is large enough for that target to be identified as novel, and a *configuration* stage, in which a word sound representation is established (Hoover, Storkel, & Hogan, 2010; McKean, Letts, & Howard, 2014; Storkel & Lee, 2011). Low neighbourhood density (i.e. high distinctiveness) has been associated with triggering stage advantages, while high neighbourhood density has been associated with configuration stage advantages. Storkel and Lee (2011), for instance, report an immediate naming and referent identification advantage for low-density stimuli (also low phonotactic probability stimuli), which is attributed to a heightened triggering effect, though better delayed test performance for high-density stimuli in the absence of further training, which is attributed to the formation of a detailed and robust long-term memory trace.

Similar effects are seen when the learning environment is made competitive through the presentation of high-density auditory stimuli (e.g. *bih*, *dih*) in the absence of additional cues associated with successful word learning, such as variance in syntactic or semantic class, pragmatic information, or related gaze or gesture cues (Stager & Werker, 1997; Swingley & Aslin, 2007). In such cases, the target word may be so similar to known neighbours that the child processes the target stimulus as an instance, perhaps a mispronunciation, of a known word. That is, learning is not triggered. Despite constituting poor task performance, this behaviour is generally adaptive because – as noted in the *General introduction* – recognition mechanisms must be liberal enough to support cross-contextual comprehension on the fly, for instance when a learner encounters an unfamiliar dialect (Church & Fisher, 1998). Furthermore, as described in the previous section on association networks, the number of minimally different words that young children know and hear in the speech directed to them is limited (Guevara-Rukoz et al., 2018). Therefore, the prior probability that a sound sequence that is very similar to a known word refers to a distinct referent is low, making it reasonable to classify that sound sequence as an instance of a known word (Swingley & Aslin, 2007).

Nevertheless, existing studies demonstrate that when complementary cues are present, the issue of the mis-perception of a novel neighbour as a known rather than unknown word may not arise. Dautriche, Swingley, and Christophe (2015), for instance, report that children aged 18 months were unable to learn a novel noun that neighboured a well-known noun (as in Swingley & Aslin, 2007), but successfully learned a novel noun that neighboured a well-known verb. Such results highlight that children's understanding of the similarity of a novel target word to stored exemplars is multi-dimensional, in this case involving syntactic class in addition to phonological features. Thus the potentially inhibitive effect of close phonological proximity is likely to be over-ridden in naturalistic learning environments, where multiple cues (e.g. gaze, gesture, pragmatic information) are present (Roy, Frank, DeCamp, Miller, & Roy, 2015), allowing learners to capitalise on phonological string similarity and acquire high-density words through analogous generalisation.

2.3.6 Section summary

In this section, I evaluated the literature on the emergence of the auditory lexicon from pre-birth to pre-literacy. The account presented can be summarised as follows. Spoken word exposure results in the formation of perceptual and conceptual memory traces, which as a whole represent the sound structure of the ambient language in the mind of the child. This process may be characterised as a form of what in computational research has been termed a *conspiracy effect* (Rumelhart et al., 1986). This effect confers memory and processing advantages, which themselves have been described by Church and Fisher (1998) in terms of long-term auditory priming. Unpacking this a little, high-density target words exhibiting auditory features consistent with the dominant features of stored exemplars are held in short-term memory more accurately and passed to long-term memory in greater detail. This is evident in better performance for high-density words on a range of tasks considered to probe the quality of word sound memories (e.g. the auditory lexical decision task), and in a marked production advantage for high neighbourhood density words. The term *analogous generalisation* provides a useful shorthand for describing this combination of fundamental memory processes. In the final section of this review, I consider how this process may be affected by neurological disorder.

2.4 The auditory lexicon in atypical development

Outside of the considerable variation observed in typical language development there are a subset children – estimated at up to 7.5% of the English-speaking population (Norbury et al., 2016) – who present language-learning difficulties severe enough to interfere with their education, career prospects, and general quality of life. Such children are heterogeneous in terms of the patterns of impairment they present, and this has contributed to the study and diagnosis of language disorder becoming something of a terminological minefield. Affected children may, for instance, be referred to variously as having developmental dysphasia, language delay, or specific language impairment. As a result of the recent CATALISE consortium on language impairment, however, developmental language disorder is now the generally agreed on term for children displaying significant language learning difficulties in the absence of a clear biomedical cause (Bishop, Snowling, Thompson, & Greenhalgh, 2016).

2.4.1 Representational deficits in developmental language disorder

While problems with the acquisition and accurate use of syntax are characteristic of developmental language disorder, a substantial number of affected children also show word learning difficulties. Such problems affect both semantic and phonological development, though the focus of this review is on the latter of these domains. Children with developmental language disorder commonly not only know fewer words than their age-matched peers, but also present experimental performance profiles suggesting that the long-term auditory word representations they form lack sufficient detail (Bishop, 2014). Across a range of tasks held to tap the quality of underlying word sound representation – including auditory lexical decision, non-word repetition, gating, naming, and eye-tracking – children with developmental language disorder commonly perform worse than age-matched though not language-matched peers, indicating a developmental delay rather than deviance (Kan & Windsor, 2010; Claessen et al., 2009; Claessen & Leitão, 2012; Maillart, Schelstraete, & Hupet, 2004; see Leonard, 2014, for review). Studies demonstrating such deficits in the absence of elicited verbal responses have been instrumental in linking these children's performance profiles to poor quality long-term word sound representations, and ruling out an explanation in terms of pure expressive impairments. For instance, in a meta-analysis of studies using the auditory lexical decision task, Jones and Brandt (2018, chapter four) reported that even when no verbal response (or only a simple yes/no response) was required, children with developmental language disorder were significantly less accurate than age-matched peers at identifying whether an auditory string was a word or non-word (see also Claessen et al., 2009; and discussion of Leonard, 2009, above).

2.4.2 Explaining representational deficits

2.4.2.1 The temporal processing deficit hypothesis

There are a number of prominent explanatory accounts of auditory word representation deficits in children with developmental language disorder. One such account is Tallal and colleagues' temporal processing deficit hypothesis (e.g. Tallal & Piercy, 1973), according to which the deficits observed in developmental language disorder are explicable in terms of a general impairment in processing auditory information. This position was bolstered by evidence that children with developmental language disorder often performed poorly on tasks testing the processing of verbal and non-verbal stimuli presented either rapidly (i.e. each stimulus presentation is short) or in quick succession (i.e. presentations follow each other quickly) (e.g. Tallal, Stark, & Curtiss, 1976; Tallal & Piercy, 1973; cf. Mody, Studdert-Kennedy, & Brady, 1997). Later attempts to establish a direct causal influence of auditory-perceptual processing on language development led to the release of a targeted intervention program that claimed to train auditory processing and produce improvements that transferred to language processing (Merzenich et al., 1996; Tallal et al., 1996).

Despite the initial promise of research in this direction, the temporal processing deficit hypothesis has, as summarised by Bishop (2014), fared less well in recent years. While evidence that auditory-processing deficits are sometimes associated with language impairment is generally robust, evidence that the training of auditory-perceptual deficits may improve language ability has been contradicted in more recent randomised controlled trials (Strong, Torgerson, Torgerson, & Hulme, 2011). This, along with evidence that auditory-perceptual deficits do not appear to be heritable, has revived concerns regarding the causal impact of auditory-perceptual deficits in developmental language disorder, with recent studies arguing that such deficits may be the outcome rather than origin of children's language difficulties (Bishop, Hardiman, & Barry, 2012).

2.4.2.2 Short-term memory

A second branch of explanatory research that maintains considerable influence in the study of early atypical word representation emphasises phonological short-term memory deficits (e.g. Gathercole & Baddeley, 1990). Children with developmental language disorder often perform significantly worse than typically developing children on measures of short-term memory, most prominently the non-word repetition task in which participants must verbally repeat a nonsense auditory string such as *hampent* or *dopelate*. Performance on this task is positively associated with vocabulary growth, providing suggestive evidence that the word learning deficits observed in developmental language disorder may be attributable to a difficulty holding target words – particularly items of three syllables or above – in phonological short-term memory, with this impeding the subsequent formation of accurate long-term word memories (Bishop, North, & Donlan, 1996; Dollaghan & Campbell, 1998; Gathercole & Baddeley, 1990). This may be because the capacity of short-term memory is limited or because its contents are subject to abnormally rapid decay.

Putting short-term memory at the heart of the lexical deficits often observed in developmental language disorder is appealing because poor performance on such tasks is heritable and highly robust, replicating with small sample sizes and stimulus inventories (Bishop et al., 1996). Nevertheless, despite being a strong marker of language disorder in general, it is clear that a range of sub-skills are involved in short-term memory tasks such as non-word repetition (e.g. auditory-perception, encoding, and motor planning), making an interpretation of one-to-one correspondence with a particular domain of deficit such as the phonological loop unwarranted (Coady & Evans, 2008). Furthermore, causal directionality is again an issue, with evidence that non-word repetition is better for relatively word-like targets indicating that the established lexicon plays an important top-down role in task performance. This has led some researchers to argue that poor non-word repetition performance is the outcome rather than origin of limited auditory lexicon size (Melby-Lervåg et al., 2012; Snowling, Chiat, & Hulme, 1991). In turn, however, evidence against a top-down explanation of non-word repetition task performance comes from Bishop et al. (1996), who demonstrated that task deficits remained in children with a history of language impairment but typical vocabulary size. Were non-word repetition task performance causally attributable to the top-down influence of the established lexicon this pattern would not be expected. One additional possibility is, of course, that non-word repetition task performance and vocabulary development are associated via a third factor, such as the aforementioned deficits in auditory-perception or relatedly in speech encoding. Summarising the literature in this area, for instance, Bishop (2014) writes:

> The available evidence could be parsimoniously explained in terms of a primary auditory deficit in speed of encoding information that affects the development of phonological classification, so that children persist in using immature strategies of encoding speech, and hence have inefficient organisation of phonological representations in the lexicon. The memory

> difficulties would be seen then as secondary to atypical encoding of phonological information. (p. 131)

Bishop's (2014) account centers on a primary impairment to mechanisms involved in speech encoding which results in a protracted period of gestalt lexical representation, as touched on in the prior comparison of the lexical restructuring hypothesis and PRIMIR.

2.4.2.3 The procedural learning deficit hypothesis

There has also been significant interest in the role that impairment to implicit memory systems may play in language disorder. One dominant account in this area is the procedural learning deficit hypothesis (Ullman & Pierpont, 2005), under which early language disorder is attributed to a difficulty unconsciously or automatically abstracting rule-like information from natural speech (e.g. Lum, Conti-Ramsden, Morgan, & Ullman, 2014; Lum, Conti-Ramsden, Page, & Ullman, 2012; Tomblin, Mainela-Arnold, & Zhang, 2007). Just as the non-word repetition task has been central to studies of phonological short-term memory, the serial response time task has been central to studies of procedural memory. Tomblin et al. (2007), for instance, asked adolescents with and without developmental language disorder to press one of four buttons in order to identify which of four squares a creature appeared in on a computer screen. In some blocks of trials the creature appeared at random across the four possible squares, while in other blocks of trials the creature appeared in a systematic (though difficult to discern) pattern. In such tasks, procedural learning is evidenced by a decrease in reaction time during patterned trials. Children with developmental language disorder are often reported to perform poorly on the serial reaction time task, with considerable systematic pattern exposure required to prompt a reaction time decrease similar to typically developing peers (e.g. Tomblin et al., 2007).

While in its initial formulation the procedural learning deficit hypothesis was linked principally to grammatical impairment, more recent work by Gupta and colleagues has emphasised the role of procedural learning mechanisms in the acquisition of word phonology (e.g. Gupta, 2012; Gupta & Cohen, 2002; Gupta & Tisdale, 2009). The account finds support in neuroanatomical and electrophysiological data indicating that children with developmental language disorder sometimes show abnormalities in the brain structures and patterns of activity associated with procedural learning (see Leonard, 2014, for review). However, while it is uncontroversial that implicit learning mechanisms play a role in language development, the position that a deficit in this area plays a causal role in developmental language disorder has recently come under intense criticism. Notably, West, Vadillo, Shanks, and Hulme (2017) reported low reliability across measures of procedural learning and no association between performance on such tasks and language and literacy outcomes.

2.4.3 Section summary

The etiology of developmental language disorder is complex, and homing in on a single explanatory account or pitting apparently distinct explanatory accounts against each other is unwarranted: Auditory processing, implicit memory, and short-term memory may all be implicated to some degree in the lexical representation deficits observed in certain children affected by developmental language disorder. That said, evidence of short-term memory deficits from studies using the non-word repetition task currently provides arguably the best proximal explanation of the auditory word learning difficulties observed in this population. In contrast to work in other domains (e.g. auditory-processing, implicit memory) such findings replicate widely and task performance is demonstrated to be heritable. Concerns regarding causal directionality remain well justified, as do concerns regarding the specific interpretation of non-word repetition data (e.g. its relation to a distal causal mechanism such as speech-encoding) and its relation to data from closely related paradigms such as auditory-processing tasks. In general, the literature reviewed in this section suggests that much more work is required in order to develop our understanding of the origin of such deficits. As highlighted by West et al. (2017), improving task reliability and increasing statistical power by working with larger participant samples will be central to this process, as will attempting to reach consensus regarding the cognitive process tapped in particular tasks and relatedly the interpretation of resulting data. Better understanding of the auditory word processing deficits observed in some children affected by developmental language disorder is essential because the auditory lexicon provides the basis for intelligible speech and later literacy and grammatical development, which in turn affect educational, career, and psychosocial outcomes (Conti-Ramsden, Durking, Toseeb, Botting, & Pickles, 2018).

2.5 Conclusion

The aim of this chapter has been to ground the five empirical studies that follow in a theoretical framework explaining the acquisition and use of dense word sound memories. I began by describing a series of questions implicit in Quine's (1960) *gavagai* thought experiment, each of which is given fuller attention in the empirical studies that follow. I then evaluated prototype, exemplar, and hybrid theories of word sound representation, arguing in favour of an exemplar-based framework in which spoken word exposures are stored in rich auditory code alongside non-linguistic speaker- and context-specific features. I also discussed lexical competition effects, and evaluated categorical and continuous methods of quantifying the association structure such effects imply. I then traced the emergence of this system of stored exemplars and networks of association from pre-birth to pre-literacy. This section involved the description of early implicit learning – a biological *conspiracy* effect – and the development of the auditory lexicon via the primary process of analogous generalisation across dense phonological neighbourhoods in the ambient language – a process to which I argued short-term memory advantages were central. Successful analogous generalisation during auditory word learning is therefore a function primarily of short-term memory advantages (that transfer to long-term memory), which are attributable to mechanisms of both implicit learning and the top-down influence of the established lexicon. This is realised as a learning advantage for high-density words, which is the defining characteristic of the emerging auditory lexicon. Finally, I looked at auditory word learning in developmental language disorder, and reported consensus that this population often shows auditory word representation deficits. Such deficits were linked to underlying impairments in auditory processing and implicit and short-term memory. However, it was argued that the literature on short-term memory is to date most reliable. In the following chapter, I provide an overview of the main methodological approaches used in this thesis.

Chapter 3 Methodology

The auditory lexicon can only be studied indirectly, and so inquiry in this area benefits from adopting multiple converging methodological approaches. The purpose of this chapter is to provide an outline of the primary methodological approaches used throughout this thesis. I begin by describing the principles of meta-analysis, which is central to chapter four. I then provide an overview of Bayesian parameter estimation, which is used in chapters five, six, seven, and eight. Finally, I describe autoencoder neural networks, which are used in chapter eight. This chapter provides a brief outline of key principles, and I provide recommendations for further reading throughout.

3.1 Principles of meta-analysis

The aim of meta-analysis is to provide aggregated data summaries. If we have a series of similar studies reporting differing degrees of support for a particular effect, it can be useful to pool this evidence to arrive at a summary estimate of the true population effect. In chapter four I apply this method to studies using the auditory lexical decision task. The motivation for this is that the auditory lexical decision task provides a good index of the quality of word sound representations when the confounding influence of retrieval or motor planning processes are removed. However, many of the existing studies using this paradigm have small sample sizes and may therefore be underpowered. Pooling estimates in a meta-analysis therefore provides one way to get a more reliable picture of the true effect in the population.

The crucial stages of conducting a meta-analysis are those prior to fitting the statistical model. An informal theorem often applied to meta-analysis is *junk in, junk out*, and avoiding a junk out scenario with unreliable population estimates begins by defining search terms that identify all empirical studies of central interest to the question at hand. It is also vital to employ a search strategy that can counteract the impact of publication bias. Published studies are likely to report stronger effects than unpublished studies, and so not including available unpublished studies – so-called grey literature – may distort results and provide a biased estimate of the population effect. On the other hand, unpublished studies may be of varying quality, having not been subject to peer review. This brings us to the second essential stage of meta-analysis prior to model fitting. Studies selected on the basis of pre-defined search terms must be filtered according to strict inclusion and quality criteria. In chapter four, for instance, this entails the removal of a large number of studies not reporting essential diagnostic information, studies not using control groups, and studies not reporting the statistics required to compute the population estimate.

With a cohort of applicable high-quality studies in hand, the meta-analysis can be conducted using a variety of software packages. In chapter four I use the metafor package (Viechtbauer, 2010) in R (RStudio Team, 2016). Digging into the statistical procedure used is outside of the scope of this chapter, and readers are referred to Field (2013) for a detailed account. Essentially, the aim is to take the mean score (e.g. reaction time/accuracy), the standard deviation, and the sample size for each group of interest (e.g. in chapter four this is groups with and without developmental language disorder), and in addition the total study cohort size, and then to use these statistics to calculate effect sizes and the standard errors of effect sizes for each empirical study. It is then possible to fit a statistical model that summarises the effect sizes across studies, providing the estimate of the true population effect that we are interested in. It is also possible to add moderators to this model to predict variance in effect sizes. For instance in chapter four I used a moderator analysis to predict individual study effect sizes on the basis of group identity (i.e. with and without language disorder) and experimental outcome (i.e. response time and accuracy).

There are, then, a number of methods for testing for publication bias post analysis. In chapter four I used the fail-safe *N* method, which provides an estimate of the number of studies reporting null effects that would be required to nullify the estimated population effect. If one or two studies could nullify the estimated population effect then the effect is not particularly robust. However, a fail-safe *N* in the hundreds or thousands would suggest a more substantial effect. It is important to acknowledge that the fail-safe *N* method faces criticism. Field (2013, p. 327), for instance, writes; 'because significance testing the estimate of the population effect size is not really the reason for doing a meta-analysis, the fail-safe *N* is fairly limited.' I am writing this *Methodology* chapter after the publication of chapter four of the current thesis, and in retrospect it is likely that I would have selected an alternative method of publication bias estimation, such as funnel plot visualisation (see Field, 2013). One advantage of making all data and code associated with the study in chapter four available via an online repository is that readers unhappy with the decision to use fail-safe *N* can easily compute their own preferred measure.

To summarise, meta-analysis provides a useful way to pool similar studies and estimate a true population effect, mitigating the issues of small sample sizes and measurement error. The folk theorem of meta-analysis is *junk in, junk out*. This emphasises how important it is to define clear search terms, apply strict eligibility and quality criteria, and to test for publication bias. Adopting open science principles such as pre-registration, the use of a PRISMA protocol (see chapter four and repository), and providing open data and code, also contributes to a persuasive meta-analysis.

3.2 Bayesian parameter estimation

The goal of statistical modelling is to summarise datasets into interpretable forms. Say we have a dataset of one million land sizes in squared meters and associated land values. Rather than scrolling through all this data in an attempt to make sense of it, it would be useful to summarise the data into a simple formula that would tell us the average value of one squared meter of land and then the average increase in the value of land associated with each square meter increase in size. On a chart with land size on the horizontal *x*-axis and land value on the vertical *y*-axis, the results of such an analysis may approximate an upward-sloping line (see Figure 3.1). This would enable us to quickly provide estimates of the value of particular land sizes, whether or not we had value data for the land size of interest. The same goal is at the heart of statistical modelling in the apparently more complex projects common in developmental cognitive psychology, where we might want to predict response reaction time or accuracy as a function of age or clinical profile, or as in chapters five, six, seven, and eight, to predict proportions of word production and comprehension using neighbourhood density.

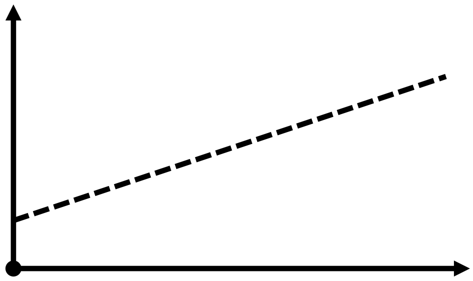

Figure 3.1: Line illustrating the relationship between two variables. This could be land size (*x*-axis) and land value (*y*-axis), or neighbourhood density (*x*-axis) and proportions of children who produce a given word (*y*-axis).

Two essential parameters define the summary line shown in Figure 3.1. The first is the point where the line touches the vertical *y*-axis. This is known as alpha, *α*. The second is the gradient or slope of the line with each unit of increase on the *x*-axis. This is known as beta, *β*. Beta is often of particular interest because it tells us the relationship between the predictor variable and the response variable. For instance, does response time on a given task increase or decrease as a function of age? Or does the chance that a child will produce a given word increase or decrease as a function of that word's neighbourhood density? In classical, so-called frequentist statistics, the outcome of many statistical tests is a value for *β*, positive for an upward-sloping line and negative for a downward-sloping line, and a *p*-value which tells you how likely it is that you would observe a *β* value at least that extreme if there was actually no effect. If the *p*-value is very low, it is unlikely that you would ever observe an effect that large if in reality there was no association between the variables assessed.

In contrast to this approach, the outcome of Bayesian statistical modelling is a probability distribution that describes the plausibility of different values of the parameter of interest. For instance, how plausible is it that *β* is 2.11, or -3.87, or 0.00?

As I describe in the empirical papers that follow, a probability distribution for *β* bound above zero (e.g. *β*=0.3 to 0.7) indicates a positive association between variables. That is, as the predictor value on the *x*-axis increases so does the response value on the *y*-axis. A distribution for *β* bound below zero (e.g. *β*=-0.3 to -0.1) suggests a negative association between the predictor and outcome variables, i.e. as the predictor value increases the response value decreases. And a distribution for *β* encompassing zero (e.g. *β*=-0.5 to 0.3) suggests that no linear relationship between predictors is plausible, i.e. a flat regression line.

To arrive at a *β* value that explains the relationship between variables of interest we can use a parameter estimation algorithm. To set this up for our neighbourhood density and word production example using the brms package (Bürkner, 2018) we can load the relevant data and packages and type the following code into R:

```
model <- brm(produces ~ neighbourhood_density,
                        data = master,
                        family = 'binomial',
                        prior = set_prior('normal(0, 1)',
                                class = "b"))
```

The code above fits a statistical model (`model <-`) using the brms package in R (`brm`). In this model word production is predicted by neighbourhood density (`produces ~ neighbourhood_density`), as recorded in the master dataset (`data = master`). The family argument refers to the likelihood, formally understood as the conditional density of the data given the parameters. In this example, we are looking at parental report data including aggregated 'produces' and 'does not produce' responses, and the binomial distribution is appropriate under these conditions (`family = 'binomial'`) (see chapter five). Finally, we set priors, which provide the algorithm with a starting point for estimation. Here I have set a prior for the *β* parameter with a normal distribution centred on zero and a standard deviation of 1 (`prior = set_prior('normal(0, 1)', class = "b")`). Note that tight, informative priors are more important when you are working with small samples. In large-scale projects such as chapters five and six of this thesis priors will often be overwhelmed by the data.

We can also add levels to our model. For instance, linguistically advanced children might be expected to do well in most experimental trials. We may therefore add a child identification variable to the syntax above to indicate that we expect the responses of each individual child to be correlated. In the same way, we might add age as a grouping variable, indicating, for instance, that two year olds will in general perform similarly to other two year olds, while five year olds will in general perform similarly to other five year olds. Such grouping information is used throughout the empirical studies of this thesis, and is added to the syntax introduced above by using the following bracketed arguments (identified by the bold black arrow):

```
model <- brm(produces ~ neighbourhood_density
                        + (1| child) + (1| age),        ←
                        data = master,
                        family = 'binomial',
                        prior = set_prior('normal(0, 1)',
                                class = "b"))
```

Inputting this code starts the parameter estimation algorithm running; a process called sampling. The outcome of this process is a series of visual chains, which represent the algorithm stepping around the parameter space (i.e. around all possible values of β) to find the most plausible value of the parameter. Figure 3.2 shows the raw output from this process with respect to proportions of word production and neighbourhood density. As you can see in the right-hand panel, the chains output from sampling resemble the output from a polygraph lie detector test, with the parameter value (e.g. β) on the y-axis and the sample number on the x-axis. It is important that the chain does not pulse or wave up and down over time and also that it is walking around and not stuck in a trough. This is what the terms stationary and well mixing refer to in the empirical studies that follow. The rhat diagnostic that can be retrieved by calling the fitted model in R provides an indication of sampling quality. The ideal rhat is 1 and 1.1 is acceptable, but higher values of rhat may indicate sampling problems.

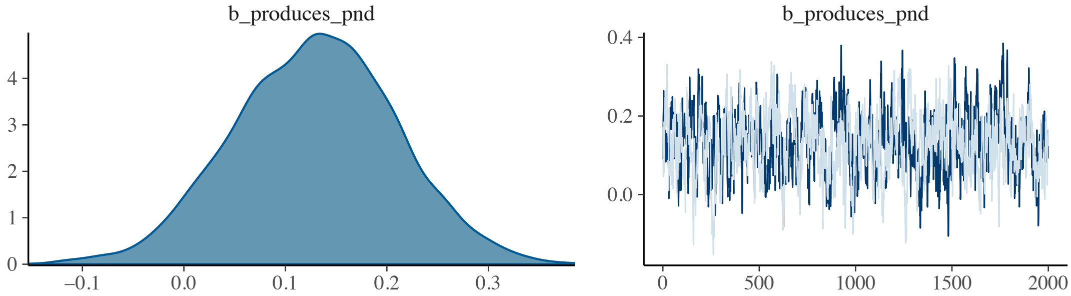

Figure 3.2: Raw probability distribution (left) and chain (right) for the association between neighbourhood density and proportions of word production. The left panel shows the density distribution of plausible values of β. The right panel shows the chain from which this distribution is derived.

The left-hand panel of Figure 3.2 shows the chain in the right-hand panel flipped on its side as a probability distribution. That is, the x-axis of the left panel is the y-axis of the right panel. Just eyeballing this density distribution you can see that most of the mass is positive (high density is associated with greater rates of word production), with a peak at around β=0.14. However note that the left-hand tail of this distribution crosses zero, indicating that it is plausible, though highly unlikely, that the true effect is zero. By putting central emphasis on a posterior distribution as shown in Figure 3.2, Bayesian statistics propagates uncertainty in the data more strongly than an emphasis on point estimates such as p-values. It is also possible, however, to summarise the posterior distribution shown by calculating a posterior mean or density interval.

In summary, Bayesian parameter estimation is one of many statistical approaches that enable us to summarise large data frames into interpretable forms. This overview focussed on the estimation of the beta, β, parameter, which along with alpha, α, describes the relationship between variables (e.g. neighbourhood density and word comprehension or production). Bayesian statistics propagates uncertainty in parameter value estimates via the posterior probability distribution, which can also be summarised into means and intervals. McElreath (2016) is a superb resource detailing the approach summarised here.

3.3 Autoencoder neural networks

Autoencoders are a class of artificial neural network that are trained to output the data they are given as input. This might seem trivial: *If we already have the input data, what use is there in training a network to output it?* Perhaps more puzzling is the fact that autoencoders are designed to be bad at copying their input to output. This feature is, however, the key asset of the architecture. Autoencoders are constrained to be unable to copy their input perfectly and as a result they learn to represent only the dominant characteristics of the input. For this reason, autoencoders are well suited to compression and denoising tasks in which the aim is to strip away extraneous detail and extract core features.

A simplified autoencoder architecture is shown in Figure 3.3, which is taken from chapter eight of this thesis. The autoencoder has three layers, an input x to the left, a hidden layer h in the middle, and an output or reconstruction layer r to the right. Labelled to the bottom left of Figure 3.3 is the encoder f that passes an input data representation such as string of 0s and 1s describing the sound features of a given word to the hidden layer h (see chapter eight for examples). To the bottom right of Figure 3.3 is the decoder, g, which tries to recreate the input x. The lines between layers illustrate weights, which are scalars that increase or reduce the influence of the signal they receive. The aim of the autoencoder is to map the input to the output via the hidden layer, that is: $g(f(x)) = x$. However, as h is constrained to be smaller than x, the network is forced to extract only dominant features of x.

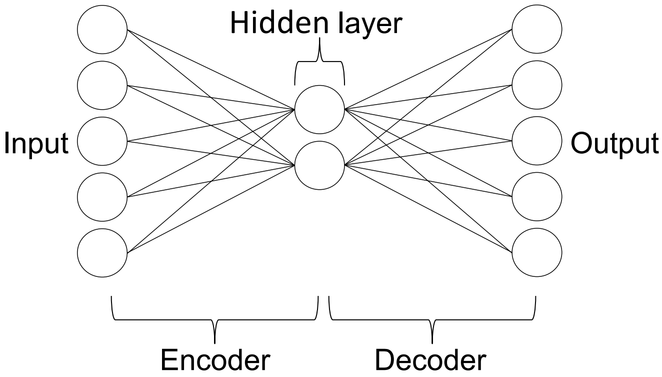

Figure 3.3: A simplified autoencoder architecture.

Prior to application the network must be trained by exposing it to data. During training the weights connecting network layers adapt gradually in order to minimise the difference between input x and output $g(f(x))$. This discrepancy, sometimes called the reconstruction error, may be measured using mean squared error, which is the average squared distance between input and output values. During training, connection weights will increase to amplify the signal from features that decrease the mean squared error, and decrease to de-amplify the signal from features that increase the mean squared error.

Mean squared error is useful because it tells us which properties of the input are easy or difficult for the network to represent. This allows us to train the autoencoder, present test items, and then use the test-phase error rates to make inferences about the quality of the internal representations formed for particular items. In chapter eight, for instance, I train an autoencoder neural network on a large corpus of child-directed speech. At test, I then present words from the MacArthur-Bates communicative development inventory (Fenson et al., 2007), and record the mean squared error for each word. It is then possible to use Bayesian regression to fit a statistical model in which test word mean squared error is predicted by lexical characteristics such as word length, exposure frequency, and neighbourhood density. Modelling results can then be validated against data from real communicative development inventory administrations.

Central to the interpretation of network performance presented in chapter eight is the principle that autoencoders have two major applications. The first, as we have seen, is feature extraction, which is central to denoising and compression tasks. When working with the numerical representations of spoken words, the network is able to form a representation of the dominant features of the sound structure of the input across the connection weights. This process is sometimes called a conspiracy effect in the literature; a term used at a number of points in this thesis (Rumelhart, McClelland, and the PDP Research Group, 1986). Presenting a new word at test that has features similar to the dominant trained features will result in a low mean squared error, as the conspiracy effect has primed the model to generalise easily to this high-density item. This property provides the basis of the theory of learning by analogous generalisation developed throughout this thesis. In contrast, presenting a test word with features orthogonal to those represented across the network's weights prompts a spike in error rate. Chapter eight suggests a parallel between this spike in error rate, a signal of anomaly detection, and the learning advantages reported for highly distinctive words in the behavioural literature (Swingley & Aslin, 2007). On this basis, it is argued that autoencoders provide a neat computational analogy to both the density and distinctiveness advantages that have been reported in studies of early word learning, showing that these apparently contradictory effects can emerge from a common architecture and learning algorithm.

In summary, autoencoders are a class of artificial neural network that aim to reconstruct a given input. Due to constraints on hidden layer size they are unable to do this perfectly, resulting in a degraded internal representation and reconstruction error. Using statistical modelling, reconstruction error can tell us which properties of the input data are easy or difficult for the network to represent. It is possible, for instance, to model reconstruction error as a function of lexical characteristics such as word length, exposure frequency, and neighbourhood density. Autoencoder modelling provides a useful computational analogy for understanding the formation of word sound representations (chapter eight), the quality of which we can tap using a range of experimental paradigms. Results can then be validated against real-world data. Goodfellow, Bengio, and Courville's (2016) *Deep learning* – the go-to textbook on the subject – is available in full online: https://www.deeplearningbook.org.

Chapter 4 Auditory Lexical Decisions in Developmental Language Disorder: A Meta-Analysis of Behavioural Studies

Linking statement: The empirical studies in this thesis are connected by the theme of word sound representation and use. The aim of this study is to set the stage for those that follow by demonstrating group differences in the ability to represent spoken words.

4.1 Abstract

Despite the apparent primacy of syntactic deficits, children with developmental language disorder (DLD) often also evidence lexical impairments. In particular, it has been argued that this population have difficulty forming lexical representations that are detailed enough to support effective spoken word processing. In order to better understand this deficit, a meta-analysis of studies testing children with DLD in the auditory lexical decision task was conducted. The objective was to provide summary effect size estimates for accuracy and response time measures, for comparisons to age- and language-matched control groups. Two thousand three hundred and seventy-two (2372) records were initially identified through electronic searches and expert consultation, with this cohort reduced to nine through duplicate removal and the application of eligibility and quality criteria. The final study cohort included 499 children aged 3;8-11;4. Multivariate analysis suggests that children with DLD were significantly less accurate in the auditory lexical decision task than age-

matched controls. For the response time estimate, however, confidence intervals for the same group comparison crossed zero, suggesting no reliable difference between groups. Confidence intervals also crossed zero for language-matched control estimates for both accuracy and response time, suggesting no reliable difference between groups on either measure. Results broadly support the hypothesis that children with DLD have difficulty forming detailed lexical representations relative to age- though not language-matched peers. However, further work is required to determine the performance profiles of potential subgroups and the impact of manipulating different lexical characteristics, such as the position and degree of non-word error, phonotactic probability, and semantic network size.

4.2 Introduction

Children with developmental language disorder (DLD; also specific language impairment, or SLI), show severe language deficits in the absence of frank neurological damage, acquired epileptic aphasia, autism-like behavior, sensory-neural hearing loss, or genetic conditions such as Down syndrome or cerebral palsy (Bishop, Snowling, Thompson, & Greenhalgh, 2016). While morpho-syntactic deficits are the hallmark of DLD, spoken word processing is also commonly impaired (see Kan & Windsor, 2010, for review). Affected children may, for instance, have difficulty repeating non-words accurately (Graf Estes, Evans, & Else-Quest, 2007), or may require longer auditory strings than age-matched controls in order to recognise a word in the gating paradigm (e.g. Dollaghan, 1998; Montgomery, 1999).

The current meta-analysis looks at the auditory lexical decision task, in which participants are required to provide a ‘yes’/‘no’ or non-linguistic (i.e. button press) judgement response to auditory word and non-word stimuli. For instance, in response to the word *dinosaur* (/daɪnəsɔː/) the participant is required to make an affirmative response, while in response to the non-word *dinokor* (/daɪnəkɔː/) the participant is required to reject the stimulus. Accuracy and response time may be recorded, and word and non-word stimuli are normally manipulated in line with primary research aims. This may include, for instance, controlling target word frequency, phonotactic probability, the number of semantically associated words (i.e. semantic network size), the position of non-word error, e.g. *dinokor* (/daɪnəkɔː/) versus *kinosaur* (/kaɪnəsɔː/), and the degree of non-word divergence, e.g. *dinokor* (/daɪnəkɔː/) versus *kinokor* (/kaɪnəkɔː/).

In its conventional form, the auditory lexical decision task is argued to measure ‘the quality or precision of stored phonological representations at the whole-word level’ (Claessen & Leitão, 2012, p. 215), with accurate rejection of a non-word taken as evidence that the corresponding word-level, phonological representation is appropriately detailed. As such, the lexical decision paradigm constitutes a useful tool to examine the hypothesis that children with DLD have difficulty forming detailed lexical representations in long-term memory, potentially as a result of underlying auditory processing or short-term memory deficits (Bishop, 1997). This pattern of development constitutes a delay rather than deviance, with young, typically developing children also apparently forming relatively holistic lexical representations prior to the emergence of a system of phonemic representation that supports the retention, and accurate and rapid processing of minimally different words (e.g. /kæt/ and /kæʧ/); a transition interacting closely with growth of the lexicon (Walley, 1993; see, however, Ainsworth, Welbourne, & Hesketh, 2016, for an interpretation of early underspecification-like performance in terms of the complexity of task demands).

In this context, the auditory lexical decision task has a number of advantages over other paradigms. First, the task arguably resembles natural spoken word recognition more closely than alternatives such as gating or non-word repetition, and so results may be more generalisable. Second, in requiring only a button touch or minimal verbal response, the task minimises the possibility that performance deficits stem from the motor output level rather than underspecification of the lexicon; an interpretation not ruled out by paradigms requiring more complex verbal responses, for instance naming and non-word repetition.

Superficially, there may be little question that children with DLD perform worse than age-matched controls on the auditory lexical decision task. However, previous meta-analyses of associated paradigms (e.g. non-word repetition; Graf Estes et al., 2007) indicate that there may exist heterogeneity in effect sizes that is masked by a general emphasis on statistical significance. The meta-analytic approach facilitates the fine-grained assessment of such heterogeneity, enabling researchers to examine which particular clinical profiles or task design features are associated with smaller or larger performance discrepancies. In doing so, results may improve our understanding of factors inhibiting spoken word processing in this population, and provide a platform for the development of evidence-based practice. Better understanding of this deficit is important because protracted lexical underspecification may have a detrimental impact on various areas of linguistic development and behaviour, including not only vocabulary learning and spoken word recognition and production, but also grammatical development and literacy (Claessen & Leitão, 2012; Goodman & Bates, 1997).

Given the extensive use of the lexical decision task in clinical and non-clinical contexts, this report may be of interest to both researchers and practitioners. The population effect size estimates may provide a useful benchmark for future research, for instance when conducting prospective power analyses or for researchers adopting a Bayesian analytical framework in which priors must be specified. Data aggregation is particularly valuable in the field of DLD given the prevalence of studies with low sample sizes, often entailing low statistical power and a high false positive rate, i.e. small samples are more likely to produce extreme values (Robey & Dalebout, 1998). The question examined is:

> *What are the estimated population effect sizes of the discrepancies in performance (response accuracy and latency) between children with DLD and age- and language-matched controls on the auditory lexical decision task?*

A substantial literature documenting lexical processing deficits across a range of paradigms (see Kan & Windsor, 2010) suggests population estimates will indicate age-matched controls regularly outperform children with DLD, with higher accuracy rates and lower response times. However, given that evidence of lexical underspecification is held to reflect a developmental delay rather than deviance (Bishop, 1997), it may be reasonable to expect little difference in estimates between children with DLD and language-matched controls.

4.3 Method

This study was pre-registered with the Open Science Framework on June 9th, 2017, with a protocol available from the associated project page (see https://osf.io/2cvnm/). The study fulfils Preferred Reporting Items for Systematic reviews and Meta-Analysis guidelines (PRISMA, see http://prisma-statement.org), with a completed checklist also available from the Open Science Framework project page.

4.3.1 Eligibility criteria

4.3.1.1 Participants

The population of primary interest was atypically developing children and adolescents, defined as those prior to or in full-time education, with age- and language-matched control groups included on the basis of provision in primary studies. Atypically developing was defined as children with DLD, as described by Bishop et al. (2016) and repeated in the introduction to the current study. A summary of the CATALISE statement on diagnostic terminology can be found at: https:// naplic.org.uk/sites/default/files/Summary%20of%20CATALISE%20%28v3%29.pdf. Participants were not distinguished on the basis of age, gender, socio-economic status, ethnicity, language, or geographical location.

4.3.1.2 Experimental design

Studies of interest were those using the auditory lexical decision task to test children with DLD. Studies were required to use experimental and control groups. No single-subject case studies were included, though there was no lower boundary on cohort size.

4.3.1.3 Outcome measures

The values of interest were the means and associated standard deviations of typical and atypical group performance on the auditory lexical decision task. This could be an accuracy rate (percentage or raw score) and/or a response time (in milliseconds; note that response times are typically only included for accurate responses in the primary literature). Standardised mean differences and variances were calculated from these primary statistics, in addition to group sizes. Throughout this study, negative effect sizes for accuracy outcomes indicate that children with DLD were less accurate than controls, while positive effect sizes for response time indicate that children with DLD were slower to respond.

4.3.1.4 Types of study

Journal articles, research reports, book chapters, and grey literature, including conference abstracts and unpublished theses and datasets were considered for inclusion. Accommodating grey literature is crucial to mitigating the impact of publication bias, whereby significant results are more likely to be published than non-significant results. Newspapers, magazine articles, and blogs were excluded. There was no restriction on the date of publication.

4.3.2 Defining and piloting search terms

Initial scope searches using the free text strings *specific language impairment, developmental language disorder,* and *lexical decision* were conducted on June 1st, 2017, using the databases PubMed, PsychINFO, Web of Knowledge, and Linguistics and Language Behavior Abstracts. These searches returned eleven studies testing clinical populations using the auditory lexical decision task, from which specific search terms were extracted from keyword lists (see Table 4.1).

Table 4.1: Keywords extracted from initial scope searches

1. Developmental language disorder (DLD)
2. Specific language impairment (SLI)
3. Language impairment
4. Phonological representation
5. Lexical representation
6. Auditory lexical decision
7. Auditory lexical judgement

These initial scope searches revealed that the paradigm was referred to variously as the auditory lexical decision task and the auditory lexical judgement task. In addition, there were anticipated differences between diagnostic labels, prominently: SLI, DLD, and language impairment. Main search terms were defined to accommodate this diversity. In particular, a strategy was developed using Boolean operators to link variations in diagnostic terminology to variations in paradigm terminology. An example search strategy in simplified (i.e. no field specification or MeSH terms) PubMed format is:

> (specific language impairment OR developmental language disorder OR language impairment) AND (auditory lexical decision OR auditory lexical judgement)

Piloting this strategy on PubMed on June 5th, 2017 returned 67 results. The number of records retrieved did not increase with the inclusion of alternative diagnostic labels including primary language impairment, developmental dysphasia, or language disability. Note that none of the finalised search terms listed above differ in British and American English spelling.

4.3.3 Main search strategy

Four approaches were used in evidence gathering: Electronic database searches, journal searches, bibliographic searches, and expert consultation. First, the following seven electronic databases were searched using the strategy specified above: Scopus, PubMed, Web of Science, LLBA, JSTOR, OVID, and ERIC. Second, forty-six journals in child language, psycholinguistics, speech-language therapy, and developmental psychology were hand searched using the aforementioned search terms and associated free text strings. The journals examined were identified during prior electronic database searches, and are listed in full on the Open Science Framework page associated with this project (see https://osf.io/2cvnm/). Third, the literature reviews and reference sections of retrieved papers were hand searched for further relevant papers. Fourth, 55 researchers were contacted regarding overlooked studies and the availability of unpublished datasets. The email sent included a link to the pre-registration protocol and a spreadsheet of studies retrieved prior to consultation, specifying author, year, title, and DOI with hyperlinks to the primary sources. The pre-registered stop search date for all data gathering was August 29th, 2017.

4.3.4 Quality, strength of evidence, and bias risk assessment

The strength of the body of evidence collected using these four search strategies was assessed according to the following criteria. Only papers that met these criteria were included; there was no ranked quality index.

1. Studies lacking an appropriate control group or data required to compute standardised mean differences and sampling variances (e.g. *M*, *SD, N*/*n*) after author consultation were excluded.

2. Studies lacking primary diagnostic or linguistic data (e.g. age, non-verbal IQ, standardised language test scores) for experimental or control groups were

excluded.

3. Studies in which authors declared conflict of interest were excluded.

Statistics from studies meeting the above quality criteria were extracted for inclusion in the meta-analysis.

4.4 Meta-analysis

4.4.1 Data extraction

Studies were attributed numeric IDs and coded by: (a) author(s); (b) year of publication; (c) DLD group mean chronological age; (d) DLD group mean language age; (e) control type (i.e. age- or language-matched); (f) outcome measure (i.e. accuracy or response time); (g) stimulus type (i.e. words, non-words); (h) stimuli sub-classification, commonly unique to the aims of original study (e.g. word initial or final manipulation in non-word formation); (i) mean scores (typically a percentage for accuracy outcomes, with response times specified in milliseconds), standard deviations, and sample sizes of DLD groups; and (j) mean scores, standard deviations, and sample sizes of control groups. Coding was conducted by the first author, with a random sample of five studies then repeated by a trained coder. Disagreements were resolved through re-examination until agreement was 100%. The complete dataset can be built using the R code available from the Open Science Framework page associated with this project (see https://osf.io/2cvnm/).

4.4.2 Software package and model selection

The meta-analysis was conducted using the Metafor package in R (Viechtbauer, 2010). This package was chosen because it is freely available and the associated code can be easily disseminated in the interests of quality assessment and replication. Metafor is also able to manage complex datasets like that analysed in the current study, with multiple control groups and dependent measurements. Given that a number of studies include both accuracy and latency outcomes (i.e. multiple-endpoints; Gleser & Olkin, 2009), as well as two types of control group (age- and language-matched), the decision was taken to fit a multivariate, random-effects model, which would accommodate stochastically dependent effect sizes while providing an overall estimate for each control group and outcome pairing.

4.4.3 Procedure

With the data frame in R, standardised mean differences (Hedges' *g*: Hedges, 1981) and sampling variances were computed using `metafor::esclac()`. In the current study, there are four comparisons of interest (see Table 4.2).

Table 4.2: Group and outcome comparisons of interest

Group comparison	Outcome measure
DLD - Age-matched controls	Accuracy
DLD - Language-matched controls	Accuracy
DLD - Age-matched controls	Response time
DLD - Language-matched controls	Response time

In order to retrieve estimates for each of these combinations (i.e. groups (age-matched, language-matched) with outcomes (accuracy, response time)), dummy variables were created and plugged into the linear model specified within the `rma.mv()` function as moderators. The model was then passed to the `robust.rma.mv()` function, which provides a robust estimate of the variance-covariance matrix of model estimates and computes tests and confidence intervals of coefficients using a small-sample adjustment. Adopting the same procedure, two additional models were fitted which specified identical moderators plus random effects at (a) study level (denoted 'author'; see model 2), and (b) both study and outcome levels (i.e. accuracy and response time; see model 3). This reflects the assumption that the underlying true effects within these levels will be more similar than the underlying true effects from different levels (see http://www.metafor-project.org/doku.php/analyses:konstantopoulos2011). Model fit was then compared using `fitstats()` to retrieve Akaike information criterion values, before identifying potential outliers calculating standardised residuals; `rstandard()`. Publication bias risk was assessed using fail-safe N, which provides an estimate of the number of additional studies reporting negligible effects required in order to nullify a summary effect (Rosenthal, 1979; Orwin, 1983; Rosenberg, 2005). If this number is relatively large, it may be inferred that the estimate is unlikely to be compromised by publication bias.

4.4.4 Search results and study selection

Figure 4.1 shows the number of studies retrieved through searches and expert consultation, and the number excluded during preliminary screening and quality and eligibility assessment. A total of 2340 records were retrieved through electronic database searches. A full record of our electronic database searches is available from the Open Science Framework page associated with this project (see https://osf.io/2cvnm/). Twelve unique records were then retrieved through bibliographic and hand searches using the aforementioned search terms and associated free text strings. The response rate to expert consultation emails was 20%, with eleven contributing author comments received and twenty studies not previously identified recommended for inclusion. In all, 2372 records were retrieved, with 2335 then excluded through duplicate removal and the screening of abstracts in line with the aforementioned criteria. This brought the number of studies sent to full-text quality and eligibility assessment to thirty-seven. At this stage the cohort included four articles considered grey literature: One pre-print, one poster, one doctoral thesis, and one research report. The bottom right panel of Figure 4.1 lists the rationales for excluding 28 studies during full-text appraisal and quality assessment. Nine studies were ultimately included in the meta-analysis, all of which were published in peer-reviewed journals between 1994 and 2016. No contributing authors declared a potential conflict of interest.

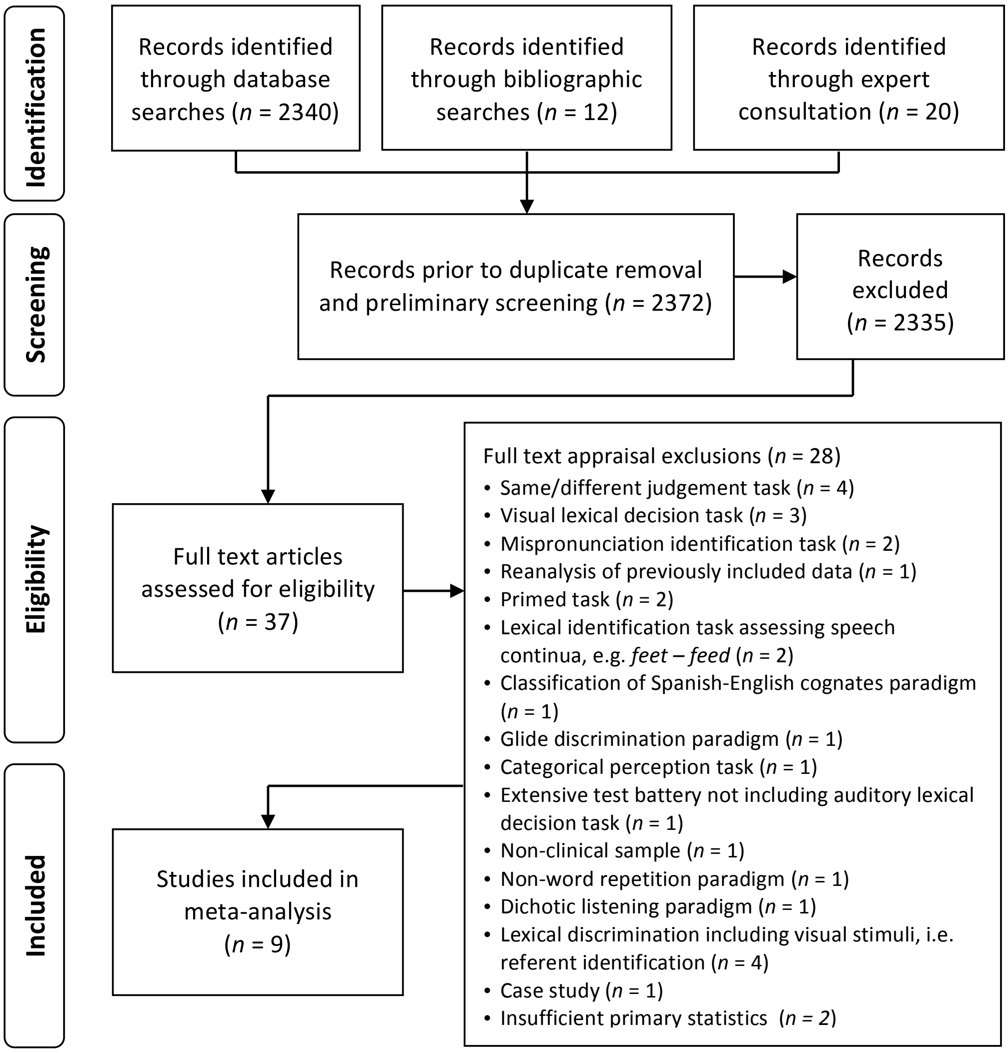
Identification
Screening
Eligibility
Included
Records identified through database searches (n = 2340)
Records identified through bibliographic searches (n = 12)
Records identified through expert consultation (n = 20)
Records prior to duplicate removal and preliminary screening (n = 2372)
Records excluded (n = 2335)
Full text articles assessed for eligibility (n = 37)
Full text appraisal exclusions (n = 28)
• Same/different judgement task (n = 4)
• Visual lexical decision task (n = 3)
• Mispronunciation identification task (n = 2)
• Reanalysis of previously included data (n = 1)
• Primed task (n = 2)
• Lexical identification task assessing speech continua, e.g. feet – feed (n = 2)
• Classification of Spanish-English cognates paradigm (n = 1)
• Glide discrimination paradigm (n = 1)
• Categorical perception task (n = 1)
• Extensive test battery not including auditory lexical decision task (n = 1)
• Non-clinical sample (n = 1)
• Non-word repetition paradigm (n = 1)
• Dichotic listening paradigm (n = 1)
• Lexical discrimination including visual stimuli, i.e. referent identification (n = 4)
• Case study (n = 1)
• Insufficient primary statistics (n = 2)
Studies included in meta-analysis (n = 9)

Figure 4.1: Study search and selection flow diagram.

4.4.5 Description of selected studies

An extensive summary of the nine studies included in the meta-analysis is presented in Appendix A.1, which details: (a) author, (b) year, (c) type(s) (i.e. age- or language-matched), ages, and sample sizes of experimental and control groups, (d) the standardised tests used to determine DLD, age-matched, and language-matched control groups[1], (e) stimulus type and number (including sub-classifications), (f) response type (verbal or non-verbal), and (g) outcome measure (accuracy or response time).

4.4.5.1 Participants

The nine studies involved a total of 499 participants: 191 with DLD (age range 3;8-11;4), 120 age-matched control participants (age range 7;3-11;4), and 188 language-matched control participants (age range 4;1-9;7). One study included only age-matched controls, while three studies included only language-matched controls. The remaining five studies included both age- and language-matched controls. Five studies included participants whose first language was English, while three tested French-speaking children, and one tested Brazilian-Portuguese speakers. Five studies specified that participants were monolingual, with monolingual or multilingual status unclear in the remaining studies.

The diagnostic criteria used in each study are specified in Appendix A.1. Diagnosis commonly worked on the basis of a verbal/non-verbal IQ discrepancy. In two studies (Edwards & Lahey, 1996; Windsor & Hwang, 1999) data from standardised diagnostic tests was use to identify subgroups, namely expressive-only (termed SLI-expressive) and expressive-receptive DLD (termed SLI-mix). In one study (Crosbie, Howard, & Dodd, 2004), subgroups not initially identified through standardised tests were defined post-hoc on the basis of auditory lexical decision task data. Befi-Lopes, Pereira, and Bento (2010) and Maillart, Schelstraete, and Hupet (2004) also include language-impaired, lexical-age subgroups based on receptive vocabulary test performance. The remaining studies did not differentiate subgroups. Note, relatedly, that the test group of James, Van Steenbrugge, and Chiveralls (1994) comprised language-disordered children with concurrent central auditory processing deficits.

[1] Note that the standardised measures used to determine language-matched control groups differed between studies (see Appendix A.1).

4.5 Results

Three robust, multivariate, random effects models were fitted. Model one specified moderators only (i.e. dummy variables specifying comparison and outcome pairing; see Table 4.2); model two specified moderators and random effects at study level; and model three specified moderators and random effects at both study and outcome levels. These models were compared using Akaike information criterion (AIC); a parsimony-adjusted measure of relative model fit based on out-of-sample deviance (McElreath, 2016). The results of this process are summarised in Table 4.3.

Table 4.3: Akaike information criterion (AIC) by model.

	Moderators only	Moderators and random effects at study level	Moderators and random effects at study and outcome levels
AIC	216.28	148.97	144.44

Decreases in AIC indicate that model fit is improved considerably by the specification of random effects at study level, and marginally by the additional specification of random effects at outcome level. Computing internally standardised residuals for model three suggested no significant outliers according to a +/-2 threshold (though a small number of cases approached this figure; see R code). Accordingly the estimates and confidence intervals reported here are derived from model three. Table 4.4 presents a full model summary.

Table 4.4: Model three summary statistics, showing condition, effect size (Hedges' g) estimate, standard error (SE), t-value, p-value, and 95% confidence interval (CI) lower (L) and upper (U) bounds. RT = reaction time.

Control type - Measure	Estimate	SE	*t*-value	*p*-value	L-CI	U-CI
Age - Accuracy	-0.88	0.19	-4.63	0.00	-1.37	-0.39
Age - RT	0.53	0.21	2.52	0.05	-0.01	1.06
Language - Accuracy	-0.46	0.23	-1.99	0.10	-1.06	0.13
Language - RT	0.22	0.13	1.74	0.14	-0.10	0.54

Four primary observations can be taken from the output shown in Table 4.4. First, children with DLD are substantially less accurate than age-matched controls in the lexical decision task, with an estimated Hedges' *g* of -0.88 (SE=0.19) and a confidence interval bound below zero (95%CI = -1.37 to -0.39), suggesting a robust population effect. Second, despite a moderate effect size, the confidence interval for the estimate reflecting response time discrepancies between children with DLD and age-matched controls marginally crosses zero (Hedges' *g* = 0.53; SE=0.21; 95%CI = -0.01 to 1.06), suggesting zero or values approaching zero are a reasonable possibility for the population effect. Third, the confidence interval for the moderate effect size reflecting accuracy discrepancies between children with DLD and language-matched controls crosses zero (Hedges' *g* = -0.46; SE=0.23; 95%CI = -1.06 to 0.13). Fourth, the confidence interval for the small effect size reflecting response time discrepancies between children with DLD and language-matched controls crosses zero (Hedges' *g* = 0.22; SE=0.13; 95%CI = -0.10 to 0.54). Forest plots visualising case and summary effect sizes and confidence intervals grouped by control type and outcome measure are presented in Appendix A.2. These plots indicate considerable variability between both studies and cases. For instance, Figure A.2.3 shows differing estimates for cases 87 (Hedges' *g* = -0.26; 95%CI = -0.78 to 0.26) and 88 (Hedges' *g* = 0.15; 95%CI = -0.36 to 0.67), both of which record discrepancies in accuracy judgements to real words between children with DLD and language-matched controls in the Haebig, Kaushanskaya, and Ellis Weismer (2015) study. Two factors potentially contributing to between- and within-study variability, namely sample heterogeneity and differences in the manipulation of experimental stimuli, are considered at length in the discussion section, where we also justify our decision not to include stimuli sub-classifications or posited subgroups as moderators.

4.5.1 Risk of publication bias

Standard fail-safe N methods do not generalise to multiple dependent outcomes (e.g. Rosenthal, 1979; Orwin, 1983; Rosenberg, 2005). In such conditions, sub-setting is required. Table 4.5 shows fail-safe Ns for Rosenthal, Orwin, and

Rosenberg methods, grouped by control group and outcome measure combination.

Table 4.5: Rosenthal, Orwin, and Rosenberg method fail-safe N by group and outcome pairing. DLD = developmental language disorder. See primary studies for details of the technical differences between methods.

Group comparison	Measure	Rosenthal	Orwin	Rosenberg
DLD - Age-matched	Accuracy	297	17	193
DLD - Language-matched	Accuracy	2178	55	937
DLD - Age-matched	Response time	262	15	187
DLD - Language-matched	Response time	20	17	5

The above estimates vary considerably, with a range of N=5 to N=2178 and substantial variation between group and outcome combinations across methods. Given a total cohort of nine studies, however, it may be reasonable to tentatively assume a low risk of publication bias significantly impacting the results reported above. Nevertheless, there does exist unobtainable data that may have shifted the estimates presented in Table 4.4: Two applicable studies were not included in the meta-analysis due to insufficient data to calculate standardised mean differences and sampling variances (Pizzioli & Schelstraete 2007, 2011), and an additional unpublished dataset was identified though not retrieved through expert consultation.

4.6 Discussion

This meta-analysis examined studies testing children with DLD on the auditory lexical decision task. Two thousand three hundred and seventy-two (2372) records were initially retrieved through electronic database searches, bibliographic searches, and expert consultation, with nine studies then selected for inclusion on the basis of eligibility and quality criteria. The final cohort included 499 children aged 3;8-11;4. The question examined was:

What are the estimated population effect sizes of the discrepancies in performance (response accuracy and latency) between children with DLD and age- and language-matched controls on the auditory lexical decision task?

This question was addressed using a multivariate, random effects model. Estimates shown in Table 4.4 suggest children with DLD were considerably less accurate than age-matched controls at identifying auditory words and rejecting auditory non-words, with a strong effect size estimate in this condition. However, the response time estimate for the same group comparison was less conclusive, with a confidence interval marginally crossing zero. This does not demonstrate no population effect, but indicates that zero or effect sizes approaching zero are a reasonable possibility for the underlying true effect. Thus while children with DLD appear considerably less accurate in the auditory lexical decision task than their age-matched peers, the current estimates suggest that they may not, in general, be significantly slower in making their responses. It is worth noting here that four primary studies investigated but found no evidence of a speed-accuracy trade-off, in which response accuracy may be compromised by a concern to provide a rapid response, or, alternatively, accuracy is high among participants who take considerable time planning a response (Crosbie et al., 2004; Edwards & Lahey, 1996; Pizzioli & Schelstraete, 2013; Quémart & Maillart, 2016). Note also that Edwards and Lahey (1996) and Crosbie et al. (2004) included a measure of auditory-vocal reaction time (AVRT), in which participants were required to say 'yes' immediately upon hearing a tone. In each study, analysis indicated no significant difference between experimental and control groups, suggesting between-group discrepancies in responses to lexical stimuli may not be attributable to difficulty identifying general forms of signal, formulating and articulating a verbal response, or the general complexity of task demands.

Confidence intervals for estimates of comparisons to language-matched controls crossed zero for both accuracy and response time outcomes, again suggesting zero is a reasonable possibility for the underlying true effect in these conditions. The observation that the accuracy deficit in particular 'disappears' when groups are matched by language ability is consistent with the view that the development of affected children is delayed though not deviant (Bishop, 1997), and provides tentative support for accounts specifying a causal association between vocabulary growth and increasingly detailed lexical representations (e.g. Walley, 1993; Walley, Metsala, & Garlock, 2003). A 'delayed but not deviant' account of results should, however, be taken with some caution. As commented by an anonymous reviewer, the current study looks at just one type of task, and it is plausible that the same participants are delayed by different time intervals in different types of task; one year in task A, though two years in task B, etc. Indeed, below we discuss posited subgroups whose receptive vocabulary skills appear unimpaired despite receptive grammatical deficits warranting diagnosis. The term delayed may therefore be something of an oversimplification because the linguistic profile of a particular language-impaired child is unlikely to correspond to a discrete age range in typical development. The term delayed also suggests that these children will eventually catch up with peers, which is unclear from the data at hand.

In summary, results of the current meta-analysis are in line with previous reports of performance deficits among children with DLD relative to age-matched controls in tasks held to measure the accuracy of lexical representations (e.g. Borovsky, Burns, Elman, & Evans, 2013; Farquharson, Centanni, Franzluebbers, & Hogan, 2014; Ramus, Marshall, Rosen, & van der Lely, 2013; Rispens & Baker, 2012). However, this conclusion requires some qualification. The forest plots shown in Appendix A.2 indicate variance in effect sizes both between and within studies, and the meta-analysis included a number of studies presenting results that contradict the overall estimates presented in Table 4.4. Befi-Lopes et al., (2010) and Maillart et al. (2004), for instance, report accuracy discrepancies between children with DLD and language-matched controls, while Edwards and Lahey (1996) and Pizzioli and Schelstraete (2013) report significant differences in response time between children with DLD and age-matched controls. In the sections that follow we discuss two broad factors that may contribute to such variability in outcomes: (a) sampling variation, within-group heterogeneity, and the presence of possible sub-groups, and (b) study-specific differences in stimulus manipulation.

4.6.1 Sampling variation and sub-groups

Children with DLD differ widely in the specific problems they have with language. Unsurprisingly, then, primary studies included in the meta-analysis often reported relatively large variances among experimental groups (e.g. Crosbie et al., 2004), while in three studies subgroups associated with different performance profiles were formally identified (Crosbie et al., 2004; Edwards & Lahey, 1996; Windsor & Hwang, 1999). Such sub-group analyses are valuable because they may help explain how children with particular patterns of impairment approach the problem of spoken word recognition. Two studies sub-classified experimental-group participants on the basis of standardised test data (Edwards & Lahey, 1996; Windsor & Hwang, 1999). Edwards and Lahey (1996) report that children they describe as having expressive-only deficits performed considerably better than children with so-called mixed, or expressive-receptive deficits. This may be expected given that the auditory lexical decision task is itself a receptive measure. However, the authors note that their sub-group analysis is of questionable validity given a lack of appropriate statistical power, which post-hoc analysis estimated at just 26% (with a type I error rate of $\alpha = .05$ and a type II error rate of $\beta = .20$, power should be .80, or 80%). Analysis of the same sub-groups by Windsor and Hwang (1999) was statistically inconclusive, and results were omitted from the published manuscript. Given sample sizes comparable to those in Edwards and Lahey (1996), however, it is likely that Windsor and Hwang's (1999) analysis was similarly underpowered. On analysis of their auditory lexical decision task data, Crosbie et al. (2004) identify a sub-group of children described as having pronounced 'lexical' deficits, who performed worse than age-matched controls, and a 'post-lexical', syntactic or integration deficit sub-group who performed in line with age-matched controls. As Crosbie et al. (2004) note, however, the post-hoc identification of sub-groups is unsatisfactory, particularly given this approaches close association with questionable research practices such as *p*-hacking (or data dredging), in which data is mined for significant patterns not included in pre-specified hypotheses. In summary, the few existing attempts to accommodate experimental group heterogeneity and sub-groups are insufficient, and this prevented against the inclusion of linguistic sub-group as a moderator in our statistical model. Further studies pre-registering sub-groups of interest and conducting prospective power analyses are required to determine the impact of linguistic sub-profile on auditory lexical decision task performance. That said, a number of researchers have emphasised the need to look at DLD in terms of dimensions of impairment rather than discrete subtypes (e.g. Bishop, 2006, p. 220). One useful direction, therefore, may be to use standardised assessment scores as continuous predictors of task performance in a linear regression model, rather than defining categorical subgroups (e.g. lexical versus post-lexical) for use in *t*-tests or ANOVAs.

4.6.2 Study-specific stimulus manipulation

There was broad consensus that the rejection of non-words was slower and less accurate across groups than responses to words, with this pattern typically pronounced in DLD groups relative to age-matched controls (Edwards & Lahey, 1996; Haebig et al., 2015; Pizzioli & Schelstraete, 2013). This finding may be attributable to the absence of long-term representations corresponding to non-words prompting extended lexical searches, though positive response bias may also play a role.

Word and non-word stimuli were often further manipulated in line with the primary study aims. Haebig et al. (2015), for instance, manipulated semantic network size to examine the role of semantics in spoken word processing by children with DLD and autism spectrum disorder. In this study, stimuli comprised twenty target words with a high number of semantically associated words, and twenty target words with a low number of semantically associated words. There is no question that such manipulations contribute to variability in effect sizes. For instance, the discrepancy highlighted above between cases 87 and 88 from Haebig et al. (2015) may be attributable to the use of high- and low-semantic network words respectively. However, because such variables were often study specific, we considered their formal inclusion as moderators in our model to be of questionable value. Position and degree of non-word manipulation (see introduction for examples) were included as independent variables in two out of nine studies (Befi-Lopes et al., 2010; Maillart et al., 2004), which we again considered insufficient to warrant the inclusion of these variables as moderators. Interestingly, however, these studies showed considerable disagreement. Maillart et al. (2004), for instance, report that children with DLD were relatively less accurate when non-word manipulations occurred in initial or final (though not medial) positions, while Befi-Lopes et al. (2010) report a word initial manipulation advantage relative to language-matched controls for children with DLD lexical age 5;0. It is plausible that this disagreement is attributable in part to assessing samples with different first languages (French- and Brazilian-Portuguese-speaking respectively), with the word regions most amenable to segmentation apparently moderated by the phonological system of a particular language (van der Feest & Fikkert, 2015). This example illustrates the complex interaction between fine-grained differences in stimulus type and the sampling variation discussed above. In summary, lack of an appropriate number of studies incorporating comparable stimulus sub-classifications made it unclear what conclusions could be drawn had we included these variables as moderators, though identifying the specific characteristics that make a non-word difficult for children with DLD to accurately reject undoubtedly constitutes an important part of the future research agenda. Researchers interested in examining the variables discussed in this section in more detail may consult the associated R file, in which all sub-classifications are coded as part of the master dataset.

4.6.3 Limitations of the study cohort

This section addresses what we consider limitations of the study cohort, with the aim of improving future research using the auditory lexical decision task. First, we are aware of no paper including formal reliability and validity estimates with respect to the auditory lexical decision task. A useful model for this line of inquiry is a recent paper by West, Vadillo, Shanks, and Hulme (2017), who conducted reliability analyses into a range of tasks thought to measure procedural learning. These researchers report low task reliability and a prevalence of so-called ‘extreme groups’, which may overestimate the extent of linear relationships between variables in the population (p. 11). It is unclear whether the lexical underspecification literature suffers from similar issues, though a partial replication of West et al.’s (2017) study in this domain would be welcome considering the diversity of paradigms argued to converge on the quality of lexical representations (e.g. gating, naming, non-word repetition, and lexical decision), as well as widespread inconsistency in outcomes.

Second, a number of studies provided no conclusive evidence that the words used at test were known to the participant (e.g. Windsor & Hwang, 1999). In some research contexts this may be unnecessary. However, in auditory lexical decision studies claiming to assess the quality of lexical representations in long-term memory, it is essential to confirm that participants know the test words. Using normative data is common, though may not be appropriate in language-impaired samples unless carefully adjusted. Explicitly testing word knowledge prior to measuring auditory lexical decisions at delayed test may be preferable.

Third, designs showed variation in both the response required by children (e.g. verbal – ‘yes’/‘no’ – or non-verbal – ergonomic box or computer screen with green/red or smiley/sad-face buttons) and the method of auditory stimulus presentation (i.e. pre-recorded or spoken live by an experimenter; see Appendix A.1). While these dissimilarities appear trivial, they can introduce systematic bias. For instance, Maillart et al. (2004) describe a pilot study in which children with DLD responded differently to pre-recorded stimuli presented via computer and stimuli spoken live by an experimenter, arguably due to adopting a compensatory strategy involving visual cues (i.e. lip reading) to aid target word discrimination (see also Bishop, Brown, & Robson, 1990). Such accounts reaffirm that researchers must carefully consider and justify each methodological decision.

Fourth, the focus of the current meta-analysis has been behavioural studies. However, McArthur and Bishop (2005) note that one limitation of the use of behavioural paradigms with children with DLD is that results below criterion could reflect low attention or motivation rather than linguistic deficits. Considering this, future research using the auditory lexical decision task may benefit from integrating neuroimaging methods assumed less susceptible to disruption by fluctuations in attentiveness, e.g. electroencephalography.

4.6.4 Limitations of the current review

This meta-analysis attempted to reduce error and bias by following PRISMA guidelines, applying explicit eligibility and quality criteria, pre-registering a research protocol, consulting experts, using multiple coders, and making the associated R code publicly available. Notwithstanding this methodological thoroughness, the current analysis has a number of limitations. First, three known datasets were omitted due to reporting insufficient statistics (*n*=2) or no longer being available (*n*=1; the latter was identified through expert consultation). Importantly, two of these studies (Pizzioli & Schelstraete 2007, 2011), reported no significant difference in response accuracy between children with DLD and age-matched controls, and so contradict the strong effect size estimate reported in Table 4.4. Unfortunately, the degree to which the inclusion of these datasets would have affected the population estimates presented in the current meta-analysis is unclear. Second, heterogeneity in the population of children with DLD along with unsatisfactory attempts to accommodate posited subgroups in the primary literature restrict the extent to which we are currently able to generalise findings to the population. Relatedly, it is regrettable that the numbers of primary studies incorporating particular stimulus sub-classifications were not sufficient to warrant the inclusion of these variables as moderators.

4.7 Conclusion

Despite the apparent primacy of syntactic deficits, children with DLD often evidence lexical impairments. In particular, it has been argued that this population has difficulty forming lexical representations detailed enough to enable them to process spoken words efficiently. The current meta-analysis examined studies testing children with DLD in the auditory lexical decision task; a behavioural paradigm commonly used in both clinical and non-clinical research contexts to assess the quality of lexical representations. Effect size estimates suggest children with DLD were less accurate though not necessarily slower in this task than age-matched controls, with no significant difference with respect to either accuracy or response time between children with DLD and language-matched controls. The primary literature provides suggestive evidence that the observed accuracy deficit may not be attributable to a speed accuracy trade-off, difficulty identifying general forms of signal, formulating and articulating a verbal response, or the complexity of task demands. Future research using the auditory lexical decision task should address the issue of within-group heterogeneity by pre-registering experimental sub-groups of possible interest or using continuous rather than categorical predictors, and conducting prospective power analyses to determine adequate sample sizes. Better understanding of the specific lexical characteristics (e.g. the position of non-word manipulation) that make an auditory non-word difficult for certain children with DLD to reject is also required. Finally, reliability analysis constitutes an important part of the future research agenda given inconsistencies in the existing literature on lexical underspecification.

4.8 References

Ainsworth, S., Welbourne, S., & Hesketh, A. (2016). Lexical restructuring in preliterate children: Evidence from novel measures of phonological representation. *Applied Psycholinguistics*, *37*(4), 997–1023. https://doi.org/10.1017/S0142716415000338

Almodovar, D. (2014). *Effects of phonological neighborhood density on lexical access in adults and children with and without specific language impairment. City University of New York (CUNY) Academic Works*. The City University of New York. Retrieved from http://academicworks.cuny.edu/gc_etds/160/?utm_source=academicworks.cuny.edu%2Fgc_etds%2F160&utm_medium=PDF&utm_campaign=PDFCoverPages

Befi-Lopes, D. M., Pereira, A. C. S., & Bento, A. C. P. (2010). Phonological representation of children with specific language impairment (SLI). *Pró-Fono*, *22*(3), 305–310. https://doi.org/S0104-56872010000300025

Bishop, D. V. M. (1997). *Uncommon understanding: Development and disorders of language comprehension in children.* Hove, England, UK: Taylor & Francis.

Bishop, D. V. M. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science*, *15*(5), 217–221. https://doi.org/10.1111/j.1467-8721.2006.00439.x

Bishop, D. V. M., Brown, B. B., & Robson, J. (1990). The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals. *Journal of Speech and Hearing Research*, *33*, 210–219. https://doi.org/10.1044/jshr.3302.210

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLOS ONE*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Borovsky, A., Burns, E., Elman, J. L., & Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with history of specific language impairment. *Journal of Communication Disorders*, *46*, 413–427. https://doi.org/10.1016/j.jcomdis.2013.09.001

Claessen, M., & Leitão, S. (2012). Phonological representations in children with SLI. *Child Language Teaching and Therapy*, *28*(2), 211–223. https://doi.org/10.1177/0265659012436851

Crosbie, S. L., Howard, D., & Dodd, B. J. (2004). Auditory lexical decisions in children with specific language impairment. *British Journal of Developmental Psychology*, *22*(1), 103–121. https://doi.org/10.1348/026151004772901131

Dollaghan, C. (1998). Spoken word recognition in children with and without specific language impairment. *Applied Psycholinguistics*, *19*(2), 193–207. https://doi.org/10.1017/S0142716400010031

Edwards, J., & Lahey, M. (1996). Auditory lexical decision of children with specific language impairment. *Journal of Speech and Hearing Research*, *39*, 1263–1273.

Farquharson, K., Centanni, T. M., Franzluebbers, C. E., & Hogan, T. P. (2014).

Phonological and lexical influences on phonological awareness in children with specific language impairment and dyslexia. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00838

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis (2nd ed.)* (pp. 357–376). New York: Russell Sage Foundation.

Goodman, E., & Bates, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Haebig, E., Kaushanskaya, M., & Ellis Weismer, S. (2015). Lexical processing in school-age children with autism spectrum disorder and children with specific language impairment: The role of semantics. *Journal of Autism and Developmental Disorders*, *45*(12), 4109–4123. https://doi.org/10.1007/s10803-015-2534-2

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

James, D., Van Steenbrugge, W., & Chiveralls, K. (1994). Underlying deficits in language-disordered children with central auditory processing difficulties. *Applied Psycholinguistics*, *15*(3), 311–328. https://doi.org/10.1017/S0142716400065917

Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *53*(3), 739–756. https://doi.org/10.1044/1092-4388(2009/08-0248)

Maillart, C., Schelstraete, M.-A., & Hupet, M. (2004). Phonological representations in children with SLI: A study of French. *Journal of Speech, Language, and Hearing Research*, *47*(1), 187–198. https://doi.org/10.1044/1092-4388(2004/016)

McArthur, G. M., & Bishop, D. V. M. (2005). Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. *Brain and Language*, *94*(3), 260–273. https://doi.org/10.1016/j.bandl.2005.01.002

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. London: Taylor and Francis. https://doi.org/10.3102/1076998616659752

Montgomery, J. W. (1999). Recognition of gated words by children with specific language impairment: An examination of lexical mapping. *Journal of Speech, Language, and Hearing Research*, *43*(3), 735–743.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157–159. https://doi.org/10.2307/1164923

Pizzioli, F., & Schelstraete, M.-A. (2007). Auditory lexical decision in children with specific language impairment. *Proceedings of the 31st Boston University Conference on Language Development*. Retrieved from: http://www.bu.edu/bucld/files/2011/05/31-Pizzioli1.pdf

Pizzioli, F., & Schelstraete, M.-A. (2011). Lexico-semantic processing in children with specific language impairment: The overactivation hypothesis. *Journal of Communication Disorders*, *44*(1), 75–90. https://doi.org/10.1016/j.jcomdis.2010.07.004

Pizzioli, F., & Schelstraete, M.-A. (2013). Real-time sentence processing in children with specific language impairment: The contribution of lexicosemantic, syntactic, and world-knowledge information. *Applied Psycholinguistics*, *34*, 1–30. https://doi.org/10.1017/S014271641100066X

Quémart, P., & Maillart, C. (2016). The sensitivity of children with SLI to phonotactic probabilities during lexical access. *Journal of Communication Disorders*, *61*, 48–59. https://doi.org/10.1016/j.jcomdis.2016.03.005

Ramus, F., Marshall, C. R., Rosen, S., & van der Lely, H. K. J. (2013). Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain*, *136*(2), 630–645. https://doi.org/10.1093/brain/aws356

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research*, *55*(3), 683–694. https://doi.org/10.1044/1092-4388(2011/10-0263)

Robey, R. R., & Dalebout, S. D. (1998). A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research*, *41*, 1227–1241. https://doi.org/1092-4388/98/4106-1227

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468. https://doi.org/10.1111/j.1095-8649.2006.01157.x

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

van der Feest, S. V. H., & Fikkert, P. (2015). Building phonological lexical representations. *Phonology*, *32*(2), 207–239. https://doi.org/10.1017/S0952675715000135

Viechtbauer, W. (2010). Conducting meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Walley, A. C. (1993). The role of vocabulary development in children′s spoken word recognition and segmentation ability. *Developmental Review*, *13*(3), 286–350. https://doi.org/10.1006/drev.1993.1015

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21(2),* e12552. https://doi.org/10.1111/desc.12552

Windsor, J., & Hwang, M. (1999). Children's auditory lexical decisions: A limited processing capacity account of language impairment. *Journal of Speech,* *Language, and Hearing Research*, *42*(4), 990–1002. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10912254

Chapter 5 Do Children Really Acquire Dense Neighbourhoods?

Linking statement: Chapter four demonstrated group differences in the ability to represent spoken words. Chapter five presents the first of three studies that look at child and lexical factors (e.g. age, proficiency, word frequency, and neighbourhood density) that affect the quality of spoken word representations, as inferred from production data.

5.1 Abstract

Children learn high phonological neighbourhood density words more easily than low phonological neighbourhood density words (Storkel, 2004). However, the strength of this effect relative to alternative predictors of word acquisition is unclear. We addressed this issue using communicative inventory data from 300 British English-speaking children aged 12 to 25 months. Using Bayesian regression, we modelled word understanding and production as a function of: (i) phonological neighbourhood density, (ii) frequency, (iii) length, (iv) babiness, (v) concreteness, (vi) valence, (vii) arousal, and (viii) dominance. Phonological neighbourhood density predicted word production but not word comprehension, and this effect was stronger in younger children.

5.2 Introduction

A variable that has received considerable attention in studies of early vocabulary development is phonological neighbourhood density, commonly defined as the number of words in a given corpus that can be formed by the addition, substitution, or elimination of a single phoneme in a target word (e.g. *cat* neighbours *catch, mat,* and *at*; Luce & Pisoni, 1998; e.g. Storkel, 2004; Storkel & Lee, 2011; Stokes, 2010, 2014; Stokes, Kern, & Dos Santos, 2012; Takac, Knott, & Stokes, 2017). Work in this direction suggests that words with high phonological neighbourhood density – i.e. words that sound similar to many other words in the target language – may be learned developmentally earlier, and on fewer experimental exposures than words that are phonologically similar to few other words. Prominent causal accounts of this effect maintain that high neighbourhood density words contain regularly occurring sounds that are held in memory more accurately during short-term processing (e.g. the *at* in *cat*, *mat,* and *catch*; Gathercole, Frankish, Pickering, & Peaker, 1999), and that this supports the formation of highly detailed long-term word memory traces (Hoover, Storkel, & Hogan, 2010; Metsala & Walley, 1998; Sosa & Stoel-Gammon, 2012; Storkel, 2004; Walley, Metsala, & Garlock, 2003; see chapter two).

Previous studies reporting high neighbourhood density advantages in early word learning have, however, considered neighbourhood density alongside only a small number of alternative predictor variables, most notably word frequency, length, and phonotactic probability (i.e. the positional probabilities of adjacent phonemic segments) (e.g. Storkel, 2004; Stokes, 2014). This is unsatisfactory because properties that appear to facilitate word acquisition in relative isolation may prove to have only a limited impact when considered alongside a more representative range of explanatory variables. For instance, Braginsky, Yurovsky, Marchman, and Frank (2019) report that word valence and word arousal, semantic features identified by Moors et al. (2013) as important determinants of word acquisition, have a relatively limited effect when modelled as part of a more representative set of predictors.

The work of Braginsky and colleagues (Braginsky et al., 2019; Braginsky, Yurovsky, Marchman, & Frank, 2016) – an important impetus for the current study – predicted age of acquisition using word frequency, word length, and a range of semantic variables (including valence and arousal) that are fully defined below. In doing so, these authors have provided the most comprehensive survey to date of features previously linked to effects in early word learning. Braginsky et al. (2016; 2019) acknowledge, however, that their explanatory models of early word learning are incomplete, with a substantial proportion of variance left unexplained (estimated at R^2=71% in Braginsky et al. 2016). The purpose of the current study is to build on Braginsky and colleagues' work by asking: When adopting a similar multi-predictor methodology, how much does word sound matter in early word learning? The variable of primary interest in this study is phonological neighbourhood density, which, as outlined above, has been widely studied in child language research. Research Question 1 asks:

> *What is the strength of association between phonological neighbourhood density and word understanding and word production when neighbourhood density is modeled alongside a representative inventory of predictor variables?*

Following previous analyses by Braginsky et al. (2016; 2019), the current study also examines developmental changes in the importance of phonological neighbourhood density and control variables as predictors of word understanding and production. Research Question 2 asks:

> *Do phonological neighbourhood density and other predictors interact with age to affect word understanding and word production?*

5.3 Method

This study was pre-registered with the Open Science Framework on September 16th, 2018. A pre-registration protocol, R code, and all data required to re-run the analyses are available via the associated project page: https://osf.io/zfy2p/.

5.3.1 Dependent variables

We used communicative development inventory data to examine phonological neighbourhood density effects in early word learning. The common format of a communicative development inventory is a wordlist plus checkboxes with fixed response options. For instance, the word *cat* may be listed as one of many words, each with two response options: 'understands' and 'produces'. During administration, caregivers may check the first box if the target child is able to understand the word *cat*, and check the second box if the target child is able to produce the word *cat*. The dependent variables used in the current study were 'understands' and 'produces' responses to 418 words from the Oxford Communicative Development Inventory, accessed via the Stanford Wordbank project (Hamilton, Plunkett, & Schafer, 2000; Frank, Braginsky, Yurovsky, & Marchman, 2017). Following previous work by Braginsky and colleagues, we restricted our analysis to cross-sectional responses. This data, collected by Floccia (2017) over a five-year period at Plymouth University, contains responses from caregivers of 300 British English-speaking children (*n*=140 female) between the ages of 12 and 25 months (*M*=18.61 months).

Parental report data are subject to reasonable validity concerns, with respondents potentially over- or under-reporting the linguistic knowledge of target children and such biases potentially affecting modelling results (see Bennetts, Mensah, Westrupp, Hackworth, & Reilly, 2016, for review). One anonymous reviewer commented that parental report comprehension data may be particularly noisy. However, the cost of administering communicative inventories is low, meaning – as Braginsky et al. (2019) note – that sample sizes are often large enough to reduce the impact of noise at the individual respondent level. The advantages of parental report data are that they provide insight into the linguistic knowledge of the child as realised in a naturalistic setting during talk with familiar people; they assess a number of words way in excess of the typical stimulus count in an experimental design; and they provide an index of words both understood and produced, allowing researchers to assess how different lexical characteristics affect these different aspects of early word learning.

5.3.2 Independent variables

Braginsky et al. (2016; 2019) present an inventory of independent variables previously assessed with respect to their association with word acquisition. The authors' approach follows Goodman, Dale, and Li (2008) in appropriating predictor data from multiple sources. We broadly adopted Braginsky et al.'s (2016; 2019) inventory of predictor variables, although we made changes to certain data sources and excluded predictors related to sentence complexity, such as a word's mean length of utterance or utterance position frequency, in order to home in on lexical effects. We then built on Braginsky et al.'s inventory by incorporating ambient language phonological neighbourhood density. Predictors, associated data sources, and example words are shown in Table 5.1.

Table 5.1: Independent variables, data sources, and minimum and maximum value examples from the Oxford CDI data.

Variable	Source	Oxford CDI examples
Child-directed speech frequency, calculated from the Manchester corpus in CHILDES	Theakston, Lieven, Pine, and Rowland (2001); MacWhinney (2000)	Min: *broom* Max: *you*

Length, in phonemes	Balota et al. (2007)	Min: *eye* Max: *cockadoodledoo*
Adult babiness rating: [1] ‘not associated with babies’ to [10] ‘associated with babies’	Perry, Perlman, and Lupyan (2015)	Min: *donkey* Max: *baby*
Concreteness rating: [1] ‘abstract’ to [5] ‘concrete’	Brysbaert, Warriner, and Kuperman (2014)	Min: *how* Max: *apple*
Valence rating: [1] ‘unhappy’ to [9] ‘happy’	Warriner, Kuperman, and Brysbaert (2013)	Min: *sad* Max: *happy*
Arousal rating: [1] ‘calm’ to [9] ‘exciting’	Warriner, Kuperman, and Brysbaert (2013)	Min: *asleep* Max: *naughty*
Dominance rating: [1] ‘controlled’ to [9] ‘in control’	Warriner, Kuperman, and Brysbaert (2013)	Min: *cry* Max: *smile*
Phonological neighbourhood density, calculated using a +/-1 phoneme criterion from the English Lexicon Project data	Balota et al. (2007)	Min: *aeroplane* Max: *moo*

The log child-directed speech frequency of each word was calculated from caregiver utterances in the Manchester corpus, which is hosted within the CHILDES database (Theakston et al., 2001; MacWhinney, 2000). This corpus includes transcripts from 12 typically developing English-speaking children (age range 1;8.22–2;0.25 at study onset) and their caregivers, who were recorded in free play for one hour, twice every three weeks for one year. Collectively these transcripts comprised 1,454,060 child-directed word tokens and 12,734 child-directed word types. Phoneme counts for each

CDI word were retrieved from the English Lexicon Project (Balota et al., 2007), with dipthongs and affricates counted as single phonemes. The English Lexicon Project provides lexical characteristic data for 40,481 words, including behavioural measures (e.g. naming response times and accuracy) from 1200 subjects. Other commonly used measures of word length, including number of orthographic letters, syllables, or morphemes, are closely correlated, and may therefore provide similar results (e.g. as in Lewis & Frank, 2016). We selected the phoneme-based measure of word length given the central interest in the phoneme as a unit of representation in the current analysis (i.e. as the basis of similarity neighbourhoods). Multiple data sources were accessed to retrieve adult ratings for babiness, concreteness, valence, arousal, and dominance. Babiness refers to the relevance of a word to babies and infants; concreteness refers to word tangibility versus abstractness; valence refers to associations with happiness or sadness; arousal to degree of excitability; and dominance to whether the word invokes notions of being controlled or submissive, or being in control or strong. Note that this last variable, dominance, was not included in prior studies by Braginsky et al. (2016; 2019). We include this variable here because it has been associated with age-related interactions in previous studies, with early-learned words having relatively high dominance ratings (Brysbaert et al. 2014). Finally, plus-minus-one phoneme phonological neighbourhood densities for Oxford CDI words were retrieved from the English Lexicon Project (Balota et al., 2007). We should acknowledge that there are a number of alternative measures of word-level phonological similarity. For instance, similarity may be calculated across only word onsets, or by taking the average edit distance between the target word and that word's twenty nearest neighbours (i.e. PLD20; Suárez, Tan, Yap, & Goh, 2011). We selected the un-weighted measure of phonological neighbourhood density excluding homophones because this is the most commonly used criterion in the developmental literature, plausibly due to the long-term dominance of this measure in adult word recognition and production studies (e.g. Storkel, 2004; Storkel & Lee, 2011; Stokes, 2010, 2014; Stokes et al., 2012; Takac et al., 2017). Importantly, this consistency allows us to directly re-evaluate the existing developmental literature reporting high neighbourhood density word learning advantages in the context of a big data, multiple-predictor analysis. Given the strong correlation between different measures of word-level phonological similarity (Suárez et al., 2011), we would expect the results reported below to hold across alternative measures.

It is also important to acknowledge word sound variables other than phonological neighbourhood density. Given our central interest in neighbourhood density effects, we omitted alternative measures including phonological variability (i.e. the degree to which productions of a single word by a single speaker vary) and phonotactic probability, which was omitted because high correlation with neighbourhood density would have caused multicollinearity (Storkel & Lee, 2011; see *Missing data and multicollinearity* for further discussion of this issue). It is likely, however, that experimenting with alternative word sound variables within a similar multi-predictor framework will improve current understanding of the factors that facilitate early word learning. Readers are therefore invited to use our data to experiment with different configurations of predictor variables, for instance by including alternative measures of neighbourhood density (e.g. PLD20) or variables such as phonotactic probability (the data repository can be found at: https://osf.io/zfy2p/).

5.3.3 Missing data and multicollinearity

The percentage of missing data ranged from 0% to 22.73% across predictor variables (see Appendix B for rates of missing data, predictor correlations, and variance inflation factors). We imputed missing values using predictive mean matching via the mice (multivariate imputation by chained equations) package in R (Buuren & Groothuis-Oudshoorn, 2011; R Core Team, 2016). All predictors were then centred and scaled into comparable units (i.e. *M*=0, *SD*=1).

Figure B.1.1 shows substantial correlations between word length and phonological neighbourhood density (*r*=-0.66), as well as between word valence and dominance (*r*=0.61), and concreteness and frequency (*r*=-0.51). Multicollinearity risk was assessed by fitting a multivariate binomial multiple regression model and computing variance inflation factors (VIFs) using the lme4 and car packages in R

(Bates, Maechler, Bolker, & Walker, 2015; Fox & Weisberg, 2011). Estimates suggested multicollinearity risk was low across predictors, with a maximum value of VIF=1.93 for the word length variable. We also conducted a post-hoc sensitivity analysis, in which we removed the word length variable and refitted the Bayesian regression model introduced fully below (see *Model fitting*). Word length was selected for removal in this analysis because of its relatively high VIF and correlation with neighbourhood density, which was the primary independent variable of interest. We found no substantial difference in estimates from the model including word length and the model excluding word length, in terms of the direction or magnitude of the estimates, or the size of the estimate errors. This can be confirmed by recalling the model summaries using the R code associated with this project, available from: https://osf.io/zfy2p/.

5.3.4 Model fitting

We used the brms package (Bürkner, 2018) to fit a Bayesian multivariate multiple binomial regression model. The model specified two outcome variables; (i) understands and (ii) produces, as reported in the 418-item communicative inventory data from 300 children. Outcomes were configured as the proportion of children at each month of age (i.e. 12 to 25 months; a 14-month range) who were able to understand or produce each item. Therefore there were $14 \times 418 = 5852$ rows of data. Word understanding and production were predicted by the independent variables listed in Table 5.1 both as main effects and in interaction with the age of the target child at the time of communicative inventory completion. We specified a random slope for age for each word, a binomial family likelihood, and a weakly informative prior across beta parameters. This model fitted successfully, with a sufficient number of effective samples, stationery and well-mixing chains, no rhats above 1.1, and credible posterior predictive checks. These analytics can be confirmed by recalling the model summary in the R code associated with this project (https://osf.io/zfy2p/).

5.4 Results

Model summaries are shown in Appendix B.3. Main effects can be seen in Figure 5.1, where the estimated strength of association between each predictor and outcome variable is visualised as a probability distribution. A distribution with mass below zero indicates a negative association between variables; a distribution with mass above zero indicates a positive association between variables; and a probability distribution centred on zero suggests no relationship between variables.

Words that children both understood and produced typically occurred at high frequency in the corpus of child-directed speech (e.g. *you*, *it*, and *that*). While many children understood relatively long words (e.g. *cock-a-doodle-do*, *pushchair*, and *television*), they tended to produce words with relatively few phonemes (e.g. *no*, *yes*, *hi*, *bye*, and *ball*). Words children both understood and produced scored highly on adult ratings of babiness (e.g. *bottle*, *milk*, and *blanket*) and concreteness (e.g. *doll*, *ball*, and *fish*). The direction of effects for word valence, arousal, and dominance differed by outcome measure. Positive valence (e.g. *happy*, *hug*, and *love*) and positive arousal (e.g. *chase*, *naughty*, and *spider*) were negatively associated with understanding but positively associated with production. In contrast, high dominance (e.g. *smile*, *happy*, *help*) was positively associated with word understanding and negatively associated with production. Finally, and with central importance to the current study, the estimate probability mass for phonological neighbourhood density (PND) was centred on zero for understanding, but positive for word production. This suggests that when we have already taken into account a word's frequency, length, babiness, concreteness, valence, arousal, and dominance, additionally knowing that word's phonological neighbourhood density does little to improve the prediction of early word understanding, but does improve the prediction of early word production. The children assessed were more likely to produce words that were phonologically similar to many other words in the language to which they were exposed (e.g. *toe*, *show*, *shoe*, *bee*, and *key*).

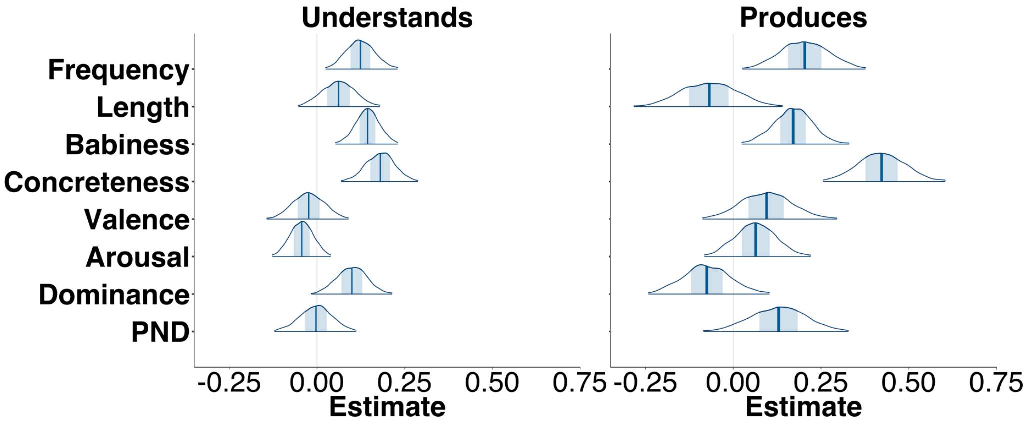

Figure 5.1 Estimate probability masses for each predictor variable in the inventory, split by understands and produces outcomes. The dark blue central line is the estimate mean, the light blue region is the 50% probability interval, and the distribution tails cover the 99% probability region. Positive values indicate that learned words were, on average, high in the associated variable. Negative values indicate that learned words were, on average, low in the associated variable. PND indicates phonological neighbourhood density.

Figure 5.2 shows interactions between each predictor and participant age, which ranged between 12 and 25 months. A positive interaction estimate indicates that the value of the predictor became more positive as age increased (e.g. a slope estimate increase from 0.01 to 0.03 between 12 and 25 months). A negative interaction estimate indicates that the value of the predictor became more negative as age increased (e.g. a slope estimate decrease from 0.01 to -0.01 between 12 and 25 months). An interaction estimate centered on zero suggests no change in the value of the predictor with age. Note that the interpretation of interaction effects depends on the direction (or sign) of the main effect. For instance, if the sign of the effect is positive, a positive interaction with age indicates a strengthening of this effect (e.g. an increase from 0.01 to 0.03). However, if the sign of the effect is negative, a positive interaction with age may indicate a weakening of this negative effect (i.e. an initially negative effect approaching zero as age increases; e.g. from -0.03 to 0).

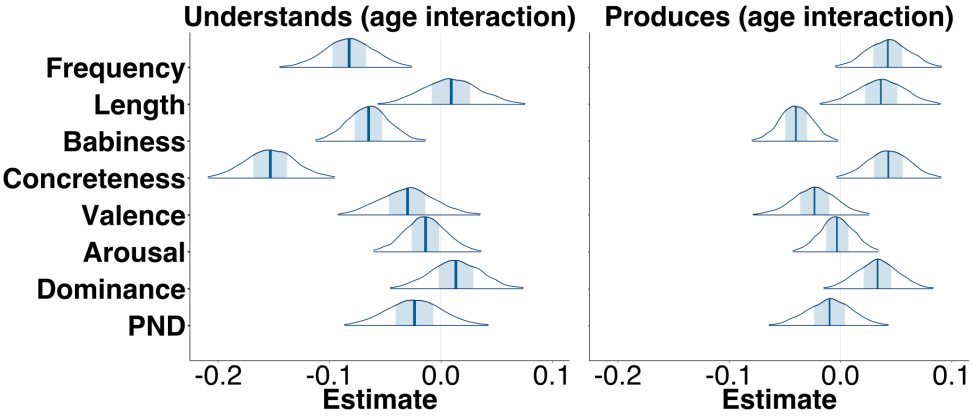

Figure 5.2: Predictor-age interaction effect probability masses by outcome. The dark blue central line is the estimate mean, the light blue region is the 50% probability interval, and the distribution tails cover the 99% probability region. Positive values indicate that the value of the predictor became more positive as age increased from 12 to 25 months. Negative values indicate that the value of the predictor became more negative between 12 and 25 months. PND indicates phonological neighbourhood density.

High input frequency became a less important determinant of word understanding across development. However, children became increasingly able to produce the words they were exposed to most frequently (e.g. *you*, *it*, and *that*). Older children were able to understand and produce words comprising more phonemes than younger children (e.g. *cock-a-doodle-do*, *pushchair*, and *television*). High relevance to the lives of babies and infants became a less important predictor of word understanding and production between 12 and 25 months, with older children acquiring low relevance words such as *broom*, *scissors,* and *write*. The association between concreteness and understanding weakened with age, as children learned abstract words such as *how*, *later* and *bad*. But the association between concreteness and production increased over development, with words such as *knee*, *bird*, and *comb* becoming part of the children's productive vocabularies. Negative trends were seen for both valence and (marginally) arousal across development, with older children more likely to understand and produce words such as *sad*, *sick*, and *hurt* (low valence), and *asleep*, *tea*, and *blanket* (low arousal). Dominance became more positively associated with understanding and less negatively associated with production (i.e. the production estimate approached zero). That is, older children were more likely to understand and produce words with associations of being in control (e.g. *smile*, *happy*, *help*, *eat*, and *say*).

For both understands and produces outcomes, the phonological neighbourhood density (PND) estimate was marginally negative, suggesting that phonological similarity to other words in the language to which children are exposed became a weaker determinant of word understanding and production across development. Estimates suggest that at around 12 months children are more likely to produce words that sound similar to other words they hear (e.g. *toe*, *show*, *shoe*, *bee*, and *key*), but that by 25 months they are able to both understand and produce words comprising less frequent sound sequences (e.g. *breakfast*, *telephone*, *toothbrush*, and *trousers*).

5.5 Discussion

In this study, we estimated the strength of the association between phonological neighbourhood density and word understanding and production when a wide range of other determining factors, including word frequency, length, valence, concreteness, babiness, arousal, and dominance, were taken into account. We also examined whether the importance of phonological neighbourhood density as a predictor of word understanding and production changed between the ages of 12 and 25 months. Results broadly comparable with prior research were observed where predictor inventories overlapped. Early-learned words were, for instance, high in child-directed speech frequency (for understanding and production), short in length (for production only), and high in babiness rating (for understanding and production) (Braginsky et al., 2016; 2019). Interaction effects also showed close parallels with prior work. A word's association with babies, for instance, was a more important predictor of understanding and production early in development than it was late in development (Braginsky et al., 2016; 2019).

Our estimates suggest that phonological neighbourhood density is an important predictor of early word production though not word understanding. In understanding a word, the balance of importance across the predictors assessed favoured high frequency of exposure, high concreteness, and high relevance to the lives of babies and infants. A word with such characteristics but complex phonology may be memorised imperfectly, which may be sufficient if the child is required to recognise and respond to though not necessary produce such a word (e.g. 'Eat your *breakfast!*' 'Do you want to rest in the *pushchair*?' 'Where's your *toothbrush*?'). However, accurate production is impossible with imperfect phonological memorisation. Therefore, with respect to word production there is an increase in the relative importance of high phonological neighbourhood density, and concurrently shorter word length (in phonemes). That is, words enter the productive lexicon more readily if their phonology is easy to remember, in terms of a low number of phonemes that occur frequently in the language to which children are exposed.

Estimates for the interaction between neighbourhood density and age suggest that phonological similarity to other words in the ambient language is a more important predictor of word understanding and production early in development rather than late in development. These results accord closely with those of prior studies reporting that the importance of phonological neighbourhood density as a predictor of word acquisition is greater in younger children and children with language delay, particularly with respect to word production (e.g. Storkel, 2004; Storkel & Lee, 2011; Stokes, 2010, 2014; Stokes et al., 2012; Takac et al., 2017). It is plausible that this effect signals increased competence in phonemic and word-level phonological representation. Accurately representing phonologically anomalous words may be difficult in early development given a relatively low frequency of exposure and limited production practice. As a result, young children may tend implicitly towards acquiring new words comprising familiar phonological patterns. Later in development, however, children are better able to represent a wider range of sounds, making phonological neighbourhood density a marginally less important predictor of whether or not a word is acquired.

A prominent explanatory account of the high neighbourhood density advantage is that cognitive demand is low during the initial processing of a novel spoken word comprising commonly occurring sounds, and that this enables the formation of detailed long-term phonological word memories that are relatively robust to forgetting and which provide detailed motor plans supporting accurate word production (Gathercole et al., 1999; Hoover et al., 2010; Metsala & Walley, 1998; Sosa & Stoel-Gammon, 2012; Storkel, 2004; Walley et al., 2003). A limitation of the current study is that it is impossible to provide evidence for any causal account on the basis of correlational data alone. In fact, it has proven difficult to test explanatory accounts of the high-density word learning advantage even in tightly controlled experiments, given multicollinearity between metrics such as neighbourhood density and phonotactic probability. The early high-density word learning advantage is, however, non-trivial, with a substantial literature documenting memorisation advantages for phonologically distinctive (i.e. as opposed to similar, or dense) stimuli (see Hunt & Worthen, 2006, for review), and further work is required to develop a causal account of this phenomenon. What the current study shows is that any explanatory model of early vocabulary development, particularly of early word production, must account for word sound features.

5.6 References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/17958156

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–48. http://www.jstatsoft.org/v67/i01/

Bennetts, S. K., Mensah, F. K., Westrupp, E. M., Hackworth, N. J., & Reilly, S. (2016). The Agreement between Parent-Reported and Directly Measured Child Language and Parenting Behaviors. *Frontiers in Psychology*, *7*, 1710. http://doi.org/10.3389/fpsyg.2016.01710

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

Society. Retrieved from: http://langcog.stanford.edu/papers_new/braginsky-2016-cogsci.pdf

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind, 3,* 52–67. https://doi.org/10.31234/osf.io/cg6ah

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904–911. Retrieved from: http://crr.ugent.be/archives/1330

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411. doi.org/10.32614/RJ-2018-017

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Floccia, C. (2017). Data collected with the Oxford CDI over a course of 5 years in Plymouth Babylab, UK. With the permission of Plunkett, K. and the Oxford CDI from Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language, 27,* 689–705. Retrieved from: http://centaur.reading.ac.uk/4542/1/Hamilton.Plunkett.Schafer.pdf

Fox, J. and Weisberg, S. (2011). *An {R} Companion to Applied Regression*, Second Edition. Thousand Oaks California: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, *44*(03), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic Influences on Short-Term Memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(1), 84–95. https://doi.org/10.1037/0278-7393.25.1.84

Gierut, J. A., & Dale, R. A. (2007). Comparability of lexical corpora: Word frequency in phonological generalization. *Clinical Linguistics and Phonetics*, *21*(6), 423–433. Retrieved from: https://doi.org/10.1080/02699200701299891

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language, 27,* 689–705. Retrieved from: http://centaur.reading.ac.uk/4542/1/Hamilton.Plunkett.Schafer.pdf

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Hunt, R. R., & Worthen, J. B. (Eds.). (2006). Distinctiveness and Memory. Oxford University Press. Retrived from: https://doi.org/10.1093/acprof:oso/9780195169669.001.0001

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. Retrieved from: https://doi.org/10.1097/00003446-199802000-00001

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from: http://talkbank.org/manuals/CLAN.pdf

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., . . . Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisuisition for 4,300 dutch words. *Behaviour Research Methods*, *45*(1), 169–177. Retrieved from: https://www.ncbi.nlm.nih.gov/pubmed/22956359

Perry, L. K., Perlman, M., Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE, 10*(9). Retrieved from: https://doi.org/10.1371/journal.pone.0137147

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from: http://www.R-project.org/

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596–608. https://doi.org/10.1044/1092-4388(2011/10-0113)

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

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

Stokes, S. F., Kern, S., & Dos Santos, C. (2012). Extended statistical learning as an account for slow vocabulary growth. *Journal of Child Language, 39*(1), 105–129. Retrieved from: https://doi.org/10.1017/S0305000911000031

Storkel, H. L. (2004). Do children acquire dense neighbourhoods? An investigation of similarity neighbourhoods in lexical acquisition. *Applied Psycholinguistics, 25*(2), 201–221. Retrieved from: https://doi.org/10.1017/S0142716404001109

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes, 26*(2), 191–211. Retrieved from: https://doi.org/10.1080/01690961003787609

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review, 18*(3), 605–611. https://doi.org/10.3758/s13423-011-0078-9

Takac, M., Knott, A., & Stokes, S. F. (2017). What can neighbourhood density effects tell us about word learning? Insights from a connectionist model of vocabulary development. *Journal of Child Language, 44*(2), 346–379. Retrieved from: https://doi.org/10.1017/S0305000916000052

Theakston, A. L., Lieven, E. V, Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*(1), 127–152. https://doi.org/10.1017/S0305000900004608

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176–1190. Retrived from: http://crr.ugent.be/papers/SUBTLEX-UK_ms.pdf

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing: An Interdisciplinary Journal*, *16*(1), 5–20. https://doi.org/10.1023/A:1021789804977

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. Retrieved from: https://doi.org/10.3758/s13428-012-0314-x

Chapter 6 Neighbourhood Density and Word Production in Delayed and Advanced Learners

Linking statement: Chapter five looked at the effect of neighbourhood density on word comprehension and production by 300 children aged 12 to 25 months. It was reported that the strength of the high-density bias declined during this period. In chapter six, I look at the association between neighbourhood density and word production in age-matched children with productive lexicons of different sizes.

6.1 Abstract

This study re-examines the claim that difficulty forming memories of words comprising uncommon sound sequences (i.e. low phonological neighbourhood density words) is a determinant of delayed expressive vocabulary development (e.g. Stokes, 2014). We modelled communicative development inventory data from *N*=442, 18-month old children, with expressive lexicon sizes between zero and 517 words (median=84). We fitted a Bayesian regression model in which the production of each communicative inventory word (*N*=680) by each child was predicted by interactions between that child's expressive lexicon size and the word's (i) phonological neighbourhood density, (ii) frequency in child-directed speech, (iii) length, (iv) babiness, and (v) concreteness. Children with larger expressive lexicons were more likely to produce words comprising uncommon sound sequences than age-matched children with smaller lexicons. However, the magnitude of the interaction between expressive lexicon size and phonological neighbourhood density was modest relative to interactions between expressive lexicon size and word frequency, length, babiness,

and concreteness. Emphasis on a difficulty with the memorisation of low neighbourhood density words as a determinant of slow vocabulary growth may be unwarranted, and the current evidence base in this direction is not robust enough to strongly support the development of possible interventions for late talkers (e.g. Stokes, 2014).

6.2 Introduction

Rates of spoken vocabulary development differ dramatically between children in the second year of life. By 18 months, children in the 95th centile (advanced learners) may produce an estimated 240 words, while same-age children below the 10th centile (so-called 'late-talkers') may produce fewer than five words (Alcock, Meints, & Rowland, 2017). Variance in expressive vocabulary size has been attributed to heritability, child gender, birth order, caregiver speech rate and quality, temperament, and attentional factors (Hammer et al., 2017; Rowe & Leech, 2017). Some studies into variance in expressive vocabulary size have also focussed on identifying the lexical characteristics that make a particular word easy or difficult for certain children to learn and produce. This work has addressed both semantic and phonological features, and suggests that the direction of discrepancy between delayed and advanced learners differs across these domains. In semantics, there is suggestive evidence that children in lower percentiles may be liberal learners (Beckage, Smith, & Hills, 2011). That is, the lexicons of late talkers may exhibit reduced semantic consistency. These children show a greater tendency than age-matched controls to acquire 'oddball' words, i.e. words that do not fit easily into existing semantic networks (though see Jimenez & Hills, 2017). With respect to phonology, however, there is evidence that children in lower expressive language percentiles are conservative learners. It has been argued that late talkers continue producing words that sound similar to many other words in the ambient language (i.e. high neighbourhood density words), when age-matched controls have started producing words comprising less common sounds (Stokes, 2010, 2014; Stokes, Kern, & Dos

Santos, 2012; Takac, Knott, & Stokes, 2017). This delay has been attributed to underlying working memory deficits impeding the accurate memorisation of words from sparse phonological neighbourhoods (e.g. Stokes, 2014). Having argued that processing phonologically uncommon words is a central determinant of delayed vocabulary growth, some of these studies have suggested interventions in which clinicians identify known words from dense phonological neighbourhoods and build vocabulary by transitioning outward from this knowledge base into increasingly sparse neighbourhoods (e.g. Stokes, 2010, 2014).

The purpose of the current study is to re-examine the claim that phonological neighbourhood density is more strongly associated with word production in children with small expressive vocabularies than in children with relatively large vocabularies. We analyse communicative development inventory data similar to that used in previous studies in this area (e.g. Stokes, 2014), but adopt a methodology that avoids some of the limitations of this earlier work. For instance, previous studies have dichotomised data into 'late talker' and 'typically developing' groups. This approach reduces both statistical power and the quality of inferences that can be drawn. Data dichotomisation may be justified when analysing populations with qualitatively different profiles, such as children with and without autistic spectrum disorder. However, it is unclear whether this approach is justifiable with respect to the study of individual differences in rates of expressive vocabulary development, including late talking, given that the majority of late talkers do not show later language difficulties (Hammer et al., 2017; Rowe & Leech, 2017).

In addition, evidence for a protracted density association in late talkers has previously involved the comparison of statements of statistical significance. For instance, Stokes (2014, p. 651) reports a statistically significant difference in the neighbourhood density of the expressive lexicons of typically developing children and late talkers, and a non-significant difference in the neighbourhood density of the receptive lexicons of typically developing and late talkers. It is argued that the expressive lexicons of late talkers, though not children in the normal range, are characterised by high neighbourhood density. This interpretation is, however, somewhat controversial because the difference between 'statistically significant' and 'non-significant' may not in itself be significant. This point is illustrated by Gelman and Stern (2006, p. 328), who imagine one analysis with a resulting effect estimate of 25 and a standard error of 10, and a second analysis with an effect estimate of 10 and a standard error of 10. Analysis one is significant at the 1% threshold, while analysis two is non-significant. Nevertheless, the difference between results is not itself significant, with a difference between estimates of 15 and a standard error of 14. Therefore while one result is significant and the other non-significant, the difference between outcomes may itself be of little practical importance.

To address these concerns in the current study, expressive lexicon size is modelled continuously. There is also an emphasis on estimate probability distributions rather than *p*-values. A probability distribution shows the relative plausibility of different parameter values, such as the beta (i.e. slope) coefficient in a linear regression model (McElreath, 2016). A probability distribution crossing zero would suggest that no linear relationship between variables was plausible (i.e. a horizontal regression line). A probability distribution with mass bound above zero would suggest a positive relationship between variables (i.e. a positive slope), and a probability mass bound below zero would suggest a negative relationship between variables (i.e. a negative slope). The decision to apply this methodology reflects our belief that probability distributions show uncertainty in the data better than point estimates such as *p*-values. The first research question we address is:

> *Is the importance of ambient language phonological neighbourhood density as a predictor of word production moderated by expressive vocabulary size in (N=442) children aged eighteen months?*

Throughout this study we are interested in whether variables such as phonological neighbourhood density are more important predictors of individual word production for children with relatively small or large expressive vocabularies. In statistical terms this means that there is an emphasis on interaction effects rather than main effects, most importantly the interaction between the child's expressive lexicon size and word phonological neighbourhood density. Evidence that children with small lexicons were more likely to produce words with high phonological neighbourhood density would be an interaction effect estimate probability distribution bound below zero. This would show that as vocabulary size increased, the strength of the positive association between high neighbourhood density and word production decreased (as reported by Stokes, 2010, 2014; Stokes et al., 2012; Takac et al., 2017). Our second research question is:

> *What is the strength of the interaction between expressive vocabulary size and phonological neighbourhood density as a predictor of word production relative to interactions between expressive vocabulary size and alternative variables associated with age of acquisition (i.e., word frequency, length, babiness, and concreteness)?*

As described above, previous studies have claimed that difficulty processing phonologically sparse words is a central determinant of limited expressive vocabulary size (Stokes 2010, 2014). These studies have also suggested interventions on the basis of evidence from parental report data similar to that used in the current study. However, because phonological neighbourhood density has to date been considered in isolation (i.e. commonly alongside only word length and frequency), we do not currently know whether the relative strength of the association between expressive lexicon size and phonological neighbourhood density is strong enough to constitute preliminary support for this position. Previous work by Braginsky, Yurovsky, Marchman, and Frank, (2019), for instance, has demonstrated that lexical features associated with significant variance in word understanding and production when modelled in isolation may show only limited relative effects when modelled as part of a larger, more representative inventory of predictors linked with age of acquisition. With this in mind, we model the interaction between expressive vocabulary size and neighbourhood density as a predictor of word production alongside interactions between expressive vocabulary size and a range of variables previously associated with age of acquisition effects; namely, word length (in phonemes), frequency (calculated from token counts in child-directed speech), babiness (i.e. adult ratings of the relevance of words for infants and babies), and concreteness. A substantial estimate for the interaction between expressive lexicon size and phonological neighbourhood density relative to estimates for the interactions between expressive lexicon size and word length, frequency, babiness, and concreteness, would constitute preliminary evidence that low phonological neighbourhood density may be a particular problem area for some children with language delay.

6.3 Method

This study was pre-registered with the Open Science Framework on 19th October 2018. A pre-registration protocol, R code, and all data required to re-run the analysis are available via the associated project page: https://osf.io/p8ax4/. The study was unfunded and undertaken as part of the first author's PhD. We declare no conflict of interest.

6.3.1 Database and sample

To answer the questions above, we used parental report data collected using the MacArthur-Bates communicative development inventory, words and sentences version (MCDI-WS; Fenson et al., 2007). The reason for using this data is that similar data were used in previous work which argued that a protracted neighbourhood density effect characterises the expressive lexicons of late talkers (i.e. children in low percentiles) (e.g. Stokes, 2010, 2014). We wanted to test whether this claim stands when using a different statistical approach and controlling for other variables (e.g. babiness and concreteness). The MCDI-WS comprises a checklist of words and phrases, but note that our analysis looked only at words. During administration, caregivers are asked to tick the boxes adjacent to items that their child is able to say. These responses (0 = does not produce; 1 = produces) for 680 words and 442 children form our dependent variable.

We accessed MCDI-WS data for 442 American English-learning children from the wordbank database using the wordbankr package in R (Braginsky, Yurovsky, Frank, & Kellier, 2018; Frank, Braginsky, Yurovsky, & Marchman, 2017; R Core Team, 2016). We selected the American English data because these were well sampled within wordbank. We selected the 18-month subset of the American English data because this was the best-sampled age group, and also because the existing work reporting protracted density effects has looked at a comparable age range (e.g. Stokes, 2010, 2014). Gender was not reported for 119 children, while 148 children were identified as female and 175 children were identified as male. Figure 6.1 shows the distribution of expressive lexicon sizes across children.

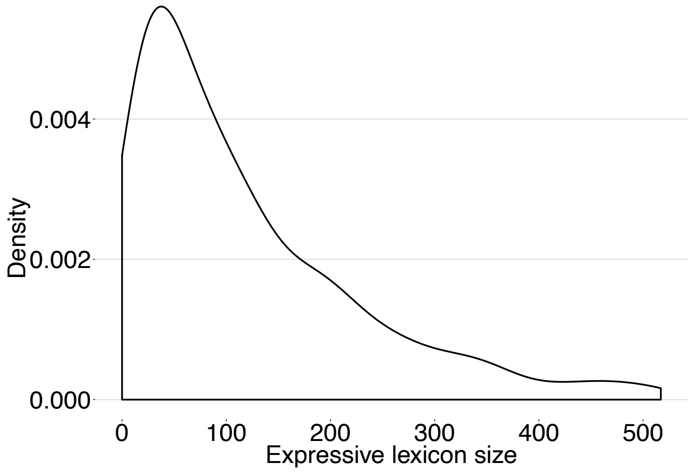

Figure 6.1: Density distribution of expressive lexicon sizes for 442 American English-learning children aged 18 months.

Figure 6.1 confirms that the sample showed the substantial individual differences in expressive lexicon size typical of their age (Alcock et al., 2017). The median lexicon size was 84 words (*M*=118 words), with a range of zero to 517 words. Ten children in the 442-participant sample had expressive lexicons of fewer than five words, and would be considered late talkers under a ≤10th centile criterion (e.g. Dale et al., 1998).

6.3.2 Predictor variables

We aimed to predict whether each child produced each MCDI-WS word using a range of lexical variables in interaction with the child's expressive vocabulary size.

The inventory of lexical variables we used was selected by reference to work by Braginsky et al. (2019), who found substantial effects for word length, frequency, babiness, and concreteness. We expanded this predictor inventory by adding phonological neighbourhood density, operationalised as the number of words in a given corpus that can be formed from a target word through one phoneme in addition, deletion, or substitution (Luce & Pisoni, 1998). Predictors, data sources, and minimum- and maximum-value example words from the MCDI-WS are shown in Table 6.1.

Table 6.1: Independent variables, data sources, and minimum and maximum value examples.

Variable	Source(s)	Examples
Child directed speech frequency	Fenson et al. (2007); Fernald, Marchman, and Weisleder (2013); Thal, Marchman, and Tomblin (2013)[2]	Min: *downtown* Max: *you*
Length, in phonemes	Balota et al. (2007)	Min: *a* Max: *cockadoodledoo*
Adult babiness rating: [1] 'not associated with babies' to [10] highly 'associated with babies'	Perry, Perlman, and Lupyan (2015)	Min: *donkey* Max: *baby*
Concreteness rating: [1] 'abstract' to [5] 'concrete'	Brysbaert, Warriner, and Kuperman (2014)	Min: *would* Max: *apple*
Phonological neighbourhood density	Balota et al. (2007)	Min: *aeroplane* Max: *boo*

[2] See http://wordbank.stanford.edu/contributors 'American English' for a full list of contributors.

Child-directed speech frequencies for each MCDI-WS word were calculated from American English transcripts in the wordbank database, before being transformed to log frequencies. We limited raw counts to transcripts in which speech from caregivers, siblings, or researchers was directed at children aged between 16 and 20 months of age. Word length was calculated in number of phonemes. Babiness and concreteness ratings from adults were retrieved from separate databases, each of which has been used in previous work by Braginsky et al. (2019). Finally, phonological neighbourhood density counts for each MCDI-WS word were retrieved from the English Lexicon project. We used the un-weighted measure of phonological neighbourhood density excluding homophones given the apparent preference for this criterion in the related literature (e.g. Stokes, 2014, Storkel, 2009).

We assessed multicolinearity risk (i.e. the possibility that high predictor correlation may distort estimates) by fitting a simple binomial regression model in which word production was predicted by each variable listed in Table 6.1 as a main effect and then calculating variance inflation factors (VIFs) using the car package in R (John et al., 2017). VIFs were low, with a maximum of 2.01 for the word length variable, suggesting that multicolinearity was not a significant issue.

The rates of missing data for each predictor variable were: 0% for expressive lexicon size, 0% for word length, 3.68% for child-directed speech frequency, 13.82% for babiness rating, 4.12% for concreteness rating, and 4.26% for phonological neighbourhood density. We imputed missing values using predictive mean matching via the mice (multivariate imputation by chained equations) package in R (Buuren & Groothuis-Oudshoorn, 2011). We then confirmed that the imputed values were plausible through strip plot visualisation, a process that can be repeated using the associated R code. All predictors were then scaled in order to make model fitting more efficient and to simplify the comparison of estimates.

6.3.3 Analysis

We used the brms package in R (Bürkner, 2018) to fit a Bayesian multiple logistic regression model in which MCDI-WS item production was predicted by each variable listed in Table 6.1 in interaction with each child's expressive vocabulary size. Child id was used as a grouping variable. We set a weakly informative prior across *β* parameters (a normal distribution centred on zero with a standard deviation of three), which we expected to be overwhelmed by the large number of observations (i.e. *N*=442 caregiver responses for 680 words=300,560 observations). This model fitted successfully, with an adequate number of effective samples, stationery and well-mixing chains, rhats uniformly at 1, and credible posterior predictive checks (see R code for analytics).

6.4 Results

A complete summary of model estimates (including main effects) is presented in Table C.1 of Appendix C. Figure 6.2 shows probability distributions for the interaction between each lexical predictor and expressive vocabulary size. A positive estimate, to the right of the grey line, indicates that as expressive vocabulary size increased, children were more likely to produce words with higher values of the associated variable. A negative estimate, to the left of the grey line, indicates that as expressive vocabulary size increased, children were more likely to produce words with lower values of the associated variable.

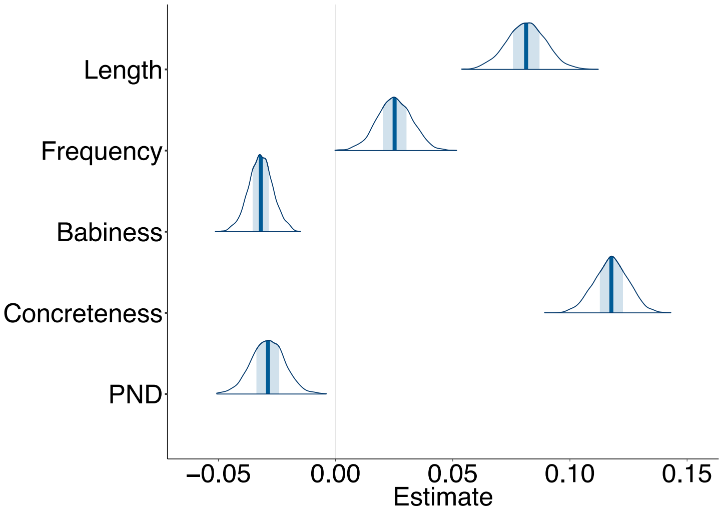

Figure 6.2: Predictor and expressive vocabulary size interaction effect probability distributions. The dark blue central line is the estimate mean, the light blue region is the 50% probability interval, and the distribution tails cover the 90% probability region.

From top to bottom (*y*-axis, Figure 6.2), children with larger expressive vocabularies were more likely to produce long words, as measured in phonemes (β =0.08; lower 95% credible interval=0.06; upper 95% credible interval=0.10). They were also more likely to produce words that occurred frequently in caregiver speech addressed to children between 16 and 20 months of age (β =0.03; lower 95% CI=0.01; upper 95% CI=0.04). Children with larger expressive lexicons were more likely than children with smaller lexicons to produce words with low babiness ratings (β =-0.03; lower 95% CI=-0.04; upper 95% CI=-0.02). They were also more likely to produce words with high concreteness ratings, with this being the most substantial effect (β =0.12; lower 95% CI=0.10; upper 95% CI=0.13). Finally, and with central importance to the current study, children with larger expressive lexicons were more likely than children with smaller expressive lexicons to produce words that were phonologically similar to few words in the ambient language (β =-0.03; lower 95% CI=-0.04; upper 95% CI=-0.01). Stated differently, high phonological neighbourhood density was more strongly associated with word production in children who could produce few words. Like all the observed estimates this interaction effect showed no probability distribution across zero. Importantly, however, the relative magnitude of the estimate for the interaction between expressive vocabulary size and phonological neighbourhood density was modest relative to interactions between expressive vocabulary size and other lexical predictors in the inventory. The neighbourhood density interaction effect was comparable to that of word frequency and word babiness, but substantially smaller in magnitude than the observed interactions with word length and concreteness. Thus, on the basis of the current or similar data, it is impossible to single out low neighbourhood density as a primary factor leading to delayed vocabulary development.

6.5 Discussion

The current study examined whether the association between phonological neighbourhood density and word production was stronger in children with small or large expressive lexicons. Research Question 1 was: *Is the importance of ambient language phonological neighbourhood density as a predictor of word production moderated by expressive vocabulary size in children aged eighteen months?* Results from parental report based on 442 children suggest that the association between phonological neighbourhood density and word production is moderated by expressive vocabulary size. The direction of the reported estimate accords with previous work on early density effects. Children with small productive lexicons were more likely to produce words with high phonological neighbourhood density (e.g. Stokes, 2014; Storkel, 2004). The interaction appears reliable, with a probability distribution bound below zero (*β*=-0.03; lower 95% CI=-0.04; upper 95% CI=-0.01).

We also considered the strength of the interaction between expressive vocabulary size and neighbourhood density relative to interactions between expressive vocabulary size and word length, frequency, babiness, and concreteness. These variables have shown substantial age of acquisition effects in previous work (e.g. Braginsky et al., 2019). Research Question 2 asked; *What is the strength of the interaction between expressive vocabulary size and phonological neighbourhood density as a predictor of word production relative to interactions between expressive vocabulary size and alternative variables associated with age of acquisition (i.e.,* *word frequency, length, babiness, and concreteness)?* None of the estimates for interactions between expressive vocabulary size and the selected lexical variables crossed zero, suggesting reliable effects for all predictors. Furthermore, the pattern of estimates for these predictors resembled those reported in work by Braginsky et al. (2019), who looked at interactions with age rather than interactions with lexicon size. For instance, our analysis showed that larger lexicons comprised more words with high CDS frequency (e.g. function words potentially omitted in early development such as *if, is,* and *that*), high concreteness (e.g. a substantial number of common nouns in addition to typically early-learned onomatopoeia and routine words such as *meow, moo, hello, bye, no*), and low babiness ratings (e.g. *glasses, stove, salt*). Recovering the reported age-related trajectories using age-matched participants serves as a reminder that the development of children in the lower percentiles we looked at was delayed though not deviant. That is, the composition of low-percentile children's lexicons in our analysis appears comparable to that of younger children in the normal range reported by Braginsky et al. (2019) (see also chapter five). Similarly, these results suggest that when discussing changes in the importance of a predictor variable, vocabulary size is a better indicator of development than age (e.g. Ainsworth, Welbourne, & Hesketh, 2016). High phonological neighbourhood density, for instance, becomes a less important predictor of word production when expressive vocabulary size rather than age per se increases.

Despite a probability distribution bound below zero signalling a reliable effect separable from other predictors, the strength of the estimate for the interaction between expressive vocabulary size and phonological neighbourhood density was modest relative to interactions between expressive vocabulary size and the other lexical variables we considered in our model. The magnitude of the phonological neighbourhood density interaction was comparable to interactions between expressive vocabulary size and word babiness and frequency. Much stronger estimates were seen for interactions between expressive vocabulary size and word length (larger lexicons comprised longer words, in number of phonemes) and concreteness. In short, the neighbourhood density interaction estimate did not stand out, with other lexical characterises similarly or more strongly associated with variance in expressive lexicon size.

A large number of experimental, naturalistic, and computational studies have demonstrated that phonology matters in word learning (e.g. Hogan, Bowles, Catts, & Storkel, 2011; Hoover, Storkel, & Hogan, 2010; Schwartz & Leonard, 1982; Stokes, 2010; Storkel, 2002, 2004, 2006, 2009; Storkel & Lee, 2011). For instance, children are more likely to recall words from dense phonological neighbourhoods at delayed test. They are also more likely to memorise and accurately produce non-words that contain sounds already in their expressive lexicons. Such experimental results suggest that the reported high phonological similarity advantage in early word learning is not an epiphenomenon but a substantive and separable effect. However, when it comes to identifying lexical characteristics that help explain the variance observed in expressive vocabulary development, previous studies reliant on parental report data may have overestimated the importance of the high neighbourhood density association in smaller lexicons by excluding important alternative predictor variables (e.g. Stokes, 2010, 2014; Stokes et al., 2012). Other factors leading to an overestimation of the importance of phonological neighbourhood density may be data dichotomisation and an emphasis on statements of statistical significance. A large number of environmental (Hammer et al., 2017; Rowe & Leech, 2017) and lexical variables (e.g. word frequency, length, babiness, and concreteness) are associated with variance in rates of expressive vocabulary growth. Placing central emphasis on a difficulty processing uncommon word phonology on the basis of the current or similar data may therefore be unwarranted.

6.5.1 Limitations

Following previous studies in this area, we analysed data from the MacArthur-Bates communicative development inventory, words and sentences version (MCDI-WS). One general limitation with repeating this correlational approach is that we cannot discuss causality. In other words, we cannot say *why* high phonological similarity appears to continue to be a more important predictor of word production for children in low language percentiles. Prior work has linked an early high neighbourhood density advantage to undeveloped phonemic representation capacity (e.g. Storkel, 2002, 2004). This work has also linked a reported protracted density association in late talkers to memory deficits such as those sometimes identified in language-impaired children (e.g. Gathercole & Baddeley, 1990; as in Stokes, 2014). In each case, it is argued that children may find it more difficult to form detailed memories of words containing sounds that occur infrequently in the ambient language. While word comprehension is possible despite underspecified lexical representations, accurate word production is not, leading to a heightened density effect in the expressive lexicons of young and language-delayed children. While prior correlational studies in this area have argued that findings similar to our own corroborate this causal account (e.g. Stokes, 2010, 2014; Stokes et al., 2012; Takac et al., 2017), the validity of any causal account can only be determined on the basis of experimental data.

A second limitation of the current analysis is that the MCDI-WS data tells us nothing about production variability. A disclaimer on the front page of the inventory addressed to caregivers reads: “If your child uses a different pronunciation of a word (for example, ‘raffe’ instead of ‘girraffe’ or ‘sketti’ instead of ‘spaghetti’), mark the word anyway” (Fenson et al., 2007, p. 1). Prior work in this area has, however, argued that production accuracy stabilises over time, and that words from dense neighbourhoods are first produced most accurately (e.g. McLeod & Hewett, 2008; Sosa & Stoel-Gammon, 2012). It is therefore probable that children in lower percentiles not only produce fewer words than age-matched peers, but also that they are less accurate and more variable in their productions, particularly with respect to phonologically uncommon words (i.e. words from sparse neighbourhoods). Given the binary outcome variable used (i.e. ‘produces’ or ‘does not produce’), we were unable to examine associations between the selected predictors and accuracy and variability in word production. However, it would be informative to repeat the current analysis using a similar inventory with graded response options (e.g. 0=‘does not produce’; 1=‘produces poorly’, 2=‘produces adequately’, 3=‘produces well’), or by calculating the accuracy and variability of transcribed phonological words (e.g. McLeod & Hewett, 2008; Sosa & Stoel-Gammon, 2012).

Future research should examine whether the results reported here generalise to different age ranges and populations, including clinical populations and children learning different languages. Such studies would improve our current understanding of individual differences in the importance of high phonological neighbourhood density as a cue to early word production.

6.6 Conclusion

A number of studies have used correlational data to argue that difficulty processing phonologically uncommon words is a central determinant of delayed expressive vocabulary development (e.g. Stokes, 2010, 2014; Stokes et al., 2012). Applying a revised methodology to comparable data we found that high phonological neighbourhood density was a reliable predictor of early word production and that this effect appears necessarily protracted in language-delayed children. However, the magnitude of this estimate relative to other known predictors of word acquisition was modest. Therefore, the claim that a difficulty acquiring low phonological neighbourhood density words is a central determinant of delayed expressive vocabulary growth may be unwarranted. The existing parental report evidence of a protracted density association in late talkers is not robust enough to support the development of possible interventions (e.g. Stokes, 2010, 2014). Experimental data is required to explore this line of inquiry further, and to determine the validity of any associated causal account.

6.7 References

Ainsworth, S., Welbourne, S., & Hesketh, A. (2016). Lexical restructuring in preliterate children: Evidence from novel measures of phonological representation. *Applied Psycholinguistics*, *37*(4), 997–1023. https://doi.org/10.1017/S0142716415000338

Alcock, K. J., Meints, K., & Rowland, C. F. (2017). *UK-CDI Words and Gestures - Preliminary norms and manual*. Retrieved from http://lucid.ac.uk/ukcdi

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–59. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958156

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5), e19348. https://doi.org/10.1371/journal.pone.0019348

Braginsky, M., Yurovsky, D., Frank, M., & Kellier, D. (2018). wordbankr: Tools for connecting to wordbank, an open repository for developmental vocabulary data. Retrieved from https://github.com/langcog/wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind, 3,* 52–67. https://doi.org/10.31234/osf.io/cg6ah

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bürkner, P.-C. (2018). Advanced bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411. doi.org/10.32614/RJ-2018-017

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Dale, P., Simonoff, E., Bishop, D., Eley, T., Oliver, B., Price, T., … Plomin, R. (1998). Genetic influence on language delay in two-year-old children. *Nature Neuroscience*, *1*(4), 324–328. https://doi.org/10.1038/1142

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(03), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*, 336–360. https://doi.org/10.1016/0749-596X(90)90004-J

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not Itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Hammer, C. S., Morgan, P., Farkas, G., Hillemeier, M., Bitetti, D., & Maczuga, S. (2017). Late Talkers: A population-based study of risk factors and school readiness consequences. *Journal of Speech, Language, and Hearing Research*, *60*(3), 607–626. https://doi.org/10.1044/2016_JSLHR-L-15-0417

Hogan, T. P., Bowles, R. P., Catts, H. W., & Storkel, H. L. (2011). The influence of neighborhood density and word frequency on phoneme awareness in 2nd and 4th grades. *Journal of Communication Disorders*, *44*(1), 49–58. https://doi.org/10.1016/j.jcomdis.2010.07.002

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Jimenez, E., & Hills, T. (2017). Network analysis of a large sample of typical and late talkers. In In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2302–2307). Austin, TX: Cognitive Science Society. Retrieved from https://mindmodeling.org/cogsci2017/papers/0438/paper0438.pdf

John, F., Weisberg, S., Adler, D., Bates, D., Baud-bovy, G., Ellison, S., … Venables, W. (2017). Package 'car.' *CRAN Repository.*

https://doi.org/10.1177/0049124105277200

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195. https://doi.org/10.1016/j.cognition.2016.04.003

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. London: Taylor and Francis. https://doi.org/10.3102/1076998616659752

McLeod, S., & Hewett, S. R. (2008). Variability in the production of words containing consonant clusters by typical 2- and 3-year-old children. *Folia Phoniatrica et Logopaedica*, *60*(4), 163–172. https://doi.org/10.1159/000127835

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLOS ONE*, *10*(9), e0137147. https://doi.org/10.1371/journal.pone.0137147

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from: http://www.R-project.org/

Rowe, M. L., & Leech, K. A. (2017). Individual differences in early word learning. In I. G. Westermann & N. Mani (Eds.), *Early word learning*. Abingdon, Oxon: CRC Press - Taylor & Francis.

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language*, *9*(2), 319–336. https://doi.org/10.1017/S0305000900004748

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596–608. https://doi.org/10.1044/1092-4388(2011/10-0113)

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech Language and Hearing Research*, *53*(3), 670–

683. https://doi.org/10.1044/1092-4388(2009/08-0254)

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

Stokes, S. F., Kern, S., & Dos Santos, C. (2012). Extended statistical learning as an account for slow vocabulary growth. *Journal of Child Language*, *39*(1), 105–129. https://doi.org/10.1017/S0305000911000031

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language*, *29*(2), 251–274. https://doi.org/10.1017/S0305000902005032

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221. https://doi.org/10.1017/S0142716404001109

Storkel, H. L. (2006). Do children still pick and choose? The relationship between phonological knowledge and lexical acquisition beyond 50 words. *Clinical Linguistics and Phonetics*, *20*(7–8), 523–529. https://doi.org/10.1080/02699200500266349

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

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609

Takac, M., Knott, A., & Stokes, S. F. (2017). What can neighbourhood density effects tell us about word learning? Insights from a connectionist model of vocabulary development. *Journal of Child Language*, *44*(2), 346–379. https://doi.org/10.1017/S0305000916000052

Thal, D. J., Marchman, V. A., & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued delay. In L. Rescorla & P. Dale

(Eds.), *Late Talkers: Language Development, Interventions, and Outcomes*. Baltimore, MD: Brookes.

Chapter 7 Accuracy and Variability in Early Spontaneous Word Production

Linking statement: Chapters five and six looked at binary outcomes, i.e. 'produces' or 'does not produce'. In chapter seven, I go beyond binary outcomes and analyse the associations between age, frequency, and neighbourhood density, and word production accuracy and variability.

7.1 Abstract

High rates of error and variability in early word production may signal speech sound disorder. However, there is little consensus regarding the degree of error and variability that may be expected in the typical range. Relatedly, while variables including child age, word frequency, and word phonological neighbourhood density are associated with variance in word production accuracy and variability, such effects remain under-examined in spontaneous speech. This study measured the accuracy and variability of 234,551 spontaneous word productions from five typically developing children in the Providence corpus (0:11-4;0). Using Bayesian regression, accuracy and variability rates were predicted by age, input frequency, phonological neighbourhood density, and interactions between these variables. Between 61% and 72% of word productions were both inaccurate and variable according to strict criteria. However loosening these criteria to accommodate production inconsistencies unlikely to be considered erroneous (e.g. the target /æləgeɪtəɹ/ pronounced /ælɪgeɪtəɹ/) reduced this figure to between 10% and 17%, with the majority of word productions then classed as accurate and stable (48% to 58%). In addition, accuracy was higher and variability

was lower in later months of sampling, and for high-frequency words and high-density words. I discuss the implications of these results for future research and the differential diagnosis of speech sound disorder, and present an explanatory account of findings emphasising the development of oral-motor skills and increasingly detailed phonological word representations.

7.2 Introduction

Word learning is often construed as a binary phenomenon: Either a child is able to understand or produce a word or not. This is reflected, for instance, in studies using Communicative Development Inventory (CDI) data, in which variables such as word frequency and neighbourhood density may be modelled as predictors of caregiver estimates of age of acquisition (e.g. Jones & Brandt, 2019a; Braginsky, Yurovsky, Marchman, & Frank, 2019; Storkel, 2004). It is clear, however, that early word learning is far from black-and-white. Children's early word productions – the focus of the current study – are often recognisable though inaccurate, and different productions of the same word can vary considerably. For instance, in a landmark study, Ferguson and Farwell (1975) describe a child aged 1;3 (one year; three months) producing ten variants of the word *pen* in a thirty minute elicitation. Such observations serve as a reminder that word learning is a dynamic process, in which oral-motor, lexical, and phonological development closely interact, and in which word productions typically become more accurate and less variable over time (Macrae, 2013).

The purpose of this study is to examine word accuracy and variability in spontaneous speech from five children recorded between 11 months and four years of age. The novelty of the current analysis is that it provides a more representative account of early word production accuracy and variability than previous work, which has been limited to a small number of target words, utterances, or consonant clusters. In the experimental literature, for instance, Sosa (2015) assessed the production of 25 words; Sosa and Stoel-Gammon (2012) assessed 30 words; Macrae (2013) assessed 20 words; and Betz and Stoel-Gammon (2005) assessed just five words all elicited repeatedly in controlled fashion. Meanwhile, with respect to prior naturalistic analyses, McLeod and Hewett (2008) assessed spontaneous speech samples collected over a six-month period but limited their analysis to words with initial or final consonant clusters that occurred in a subset of 100 utterances. Similarly, Ota and Green (2013) analysed accuracy and variability rates among three children recorded in the Providence corpus (Demuth & McCullough, 2009) – the corpus used in the current study – but limited their analysis to six classes of consonant cluster. In contrast, this study involves the analysis of 234,551 word tokens (4360 types) spoken over a three-year period.

The selection of a small number of target words, utterances, or consonant clusters may be seen as an advantage rather than as a limitation. Restricting the test inventory to a handful of items makes experimentation and analysis more practical and establishes a procedural framework that may be applied in clinical settings, where rates of accuracy and variability are of diagnostic interest and in which time or the child's attention may be limited. For this reason, measures such as the Word Inconsistency Assessment (developed to identify inconsistent speech disorder; Dodd, Hua, Crosbie, Holm, & Ozanne, 2002) test the accuracy and variability of just 25 words. However, it is also likely that an analysis unrestricted by phoneme cluster, word class, syllable, word, or by utterance count, can provide additional insight into early word production accuracy and variability rates, which may in turn improve understanding of early language, memory, and oral-motor development. This is the broad aim of the current study.

The current manuscript presents two analyses. The first is a descriptive summary of word production accuracy and variability rates for each child in the Providence corpus (Demuth & McCullough, 2009). Following Grunwell (1992), prior experimental studies have categorised word productions into four classes: (i) accurate and stable (i.e. correct across multiple productions); (ii) accurate but variable (i.e. different across productions but including correct forms); (iii) inaccurate but stable; and (iv) inaccurate and variable (e.g. Holm et al., 2007; McLeod & Hewett, 2008; Sosa, 2015). The current study is novel in applying this taxonomy to early spontaneous speech. In the prior literature using this approach there has been special interest in the rate of words produced variably in the absence of accurate forms; this class is termed variable without hits (e.g. Holm et al., 2007; Grunwell's, 1992, class (iv) listed above). This is because high rates of variability without hits has been proposed as a marker of early speech sound disorder, which is an umbrella term describing a general difficulty acquiring accurate and intelligible speech in line with peers (Sosa, 2015, p. 24). The production of an excessive number of words categorised as variable without hits has been termed inconsistent speech disorder in order to differentiate this profile from the spoken word error and variability expected within the normal range (e.g. Holm et al., 2007).

A problem with using inconsistency rates to identify speech sound disorder, however, is that substantial discrepancies in accuracy and variability estimates from studies involving typically developing children make it difficult to determine what constitutes the normal range. For instance, Holm et al. (2007) report on an elicitation task involving 409 typically developing children in which only 13% of words were produced variably at age 3;0–3;5, with this figure dropping to 2.5% by age six. Notably, these authors report that the majority of variable forms produced were variable with hits, and conclude that "inconsistency [i.e. variability without hits] is not a feature of normal development at any age" (Holm et al., 2007, p. 483). In contrast, a number of studies have reported much higher rates of error and variability in typically developing children. McLeod and Hewett (2008), for instance, report a variability rate of 53.7% among children aged 2;0-3;4; Macrae (2013) reports a variability rate of 77.7% among children aged 1;9-3;1; and, in a direct replication of Holm et al. (2007), Sosa (2015) reports a variability rate of 77% among children aged 2;6-2;11, and 57% among children aged 3;6-3;11. Importantly, the most frequent response type reported by Sosa (2015) was variable without hits, which comprised 45% of all responses across age groups for 25 words (range = 4% to 76%). Sosa (2015) attributes the discrepancy in estimates with the original Holm et al. (2007) study in part to heightened transcription validity. Sosa (2015) made offline transcriptions of recordings and adopted a so-called consensus procedure in which two or more listeners transcribed each spoken word. In contrast, Holm et al. (2007) used online transcription (i.e. transcriptions were made as the child was speaking) and no consensus procedure, with reliability checks for up to only 10% of the data. Sosa (2015) notes, however, that transcription validity alone cannot fully account for the discrepancy observed. Re-coding the replication study data and ignoring vowel quality differences – which it is argued may be vulnerable to online transcription error – Sosa (2015) reports an overall variability rate of 56%. Although lower than the rate of 68% variability from the initial coding, this revised figure remains considerably higher than the 12% reported by Holm et al. (2007). Ultimately, it is concluded that the reason for the discrepancy in estimates remains difficult to establish, and this may mean that transcription-based assessment is too unreliable for use in a clinical setting. Sosa (2015) maintains, however, that the prevalence of variable without hits responses in the typically developing population (observed under both methods of transcription) calls into question the validity of using production inconsistency as a marker for the differential diagnosis of early speech sound disorder. An aim of the current study – specifically the first analysis – is to contribute to on-going debate regarding the degree of accuracy and variability that may be expected in the normal range.

The second analysis of the current study looks at factors explaining early accuracy and variability rates. Despite between-study discrepancies in estimates, there is general agreement that older children show higher word production accuracy and lower word production variability than younger children (Macrae, 2013; Sosa 2015; Holm et al., 2007). For this reason, age at word production is included as a predictor in the statistical model of accuracy and variability rates later presented. It is important to note, however, that age is serving here as a proxy variable in lieu of more fine-grained measurements, and that two dominant overlapping mechanisms have been suggested to explain this developmental trend. In one account, early production inaccuracy and variability is attributed to immature oral-motor control (e.g. Kent, 1992). This position is apparently supported by evidence of heightened spatial and temporal variation in the movement of articulators (i.e. jaw, tongue, lips) during childhood (Goffman & Smith, 1999). One limitation of this account, noted by Sosa (2015, p. 33), is that the mapping between motor control and segment production accuracy and variability is imperfect. Goffman, Gerken, and Lucchesi (2007), for instance, note that word production can be accurate despite spatial and temporal variation in motor control, while conversely segment production inaccuracy and variability may occur despite apparently mature motor control. There is, nevertheless, good evidence that children's oral-motor skills develop substantially during the early years, and furthermore that this development correlates with language skills independently of general cognitive ability (e.g. Alcock, 2006). Thus while production error or instability may not always indicate immature or disordered motor control, it would appear reasonable to assume that oral-motor development to some degree underpins children's developing word production accuracy and stability.

A second and compatible account attributes early word production error and variability to underspecified phonological word representations. In learning a new word, the child must remember that word's phonological features alongside semantic and pragmatic information. A large number of studies have argued that phonological word representation follows a trajectory from holistic to segmental (Metsala & Walley, 1998; Ventura, Kolinsky, Fernandes, Querido, & Morais, 2007; Walley, 1993; Ferguson & Farwell, 1975). For instance, older and linguistically advanced children often identify mispronunciations in known words more rapidly and accurately than younger or less advanced peers (e.g. Ainsworth, Welbourne, & Hesketh, 2016; see also Edwards, Beckman, & Munson, 2004; Munson, Edwards, & Beckman, 2005, for related evidence with respect to non-word repetition accuracy). This work remains somewhat controversial, with apparently conflicting studies reporting early sensitivity to sub-lexical phonemic detail and mispronunciations (e.g. Swingley & Aslin, 2002). Nevertheless, a broad view is that phonological word representations become increasingly detailed as the lexicon grows, and subsequently with the onset of literacy. One possibility, then, is that holistic phonological word representations provide an insufficient basis for accurate and stable motor planning and output, which evidences in production error and variability in early typical development and protracted speech inconsistency in atypical development (Holm et al., 2007; Sosa, 2015). In line with this position, Macrae and Sosa (2015) report no effect of child age on word production variability when controlling for expressive vocabulary size.

In addition to these child-based factors – i.e., oral-motor and memory/representational development – it is important to acknowledge lexical influences on early word production accuracy and variability. Young children produce certain words more or less accurately or stably than other words (Sosa & Stoel-Gammon, 2006), suggesting that child-based factors interact with specific features of the target word. In order to understand this observation, a number of studies have modelled accuracy and variability rates as a function of lexical variables of specific interest, such as phonological complexity (Macrae, 2013). In the same way, the current study examines how child-directed speech frequency and phonological neighbourhood density affect spoken word accuracy and variability. Frequency effects occur at all levels of linguistic representation (e.g. phoneme, word, and syntax), and it is therefore argued that such effects must be accommodated under any credible account of first language acquisition (Ambridge, Kidd, Rowland, & Theakston, 2015). Prior work using a range of paradigms (e.g. elicitation, naming) shows a negative association between word frequency and error and variability rates (e.g. Sosa & Stoel-Gammon, 2012). This pattern has been attributed to repeated exposure to a target word strengthening the corresponding phonological word representation and therefore providing a fine-grained motor plan.

High phonological neighbourhood density – i.e. phonological similarity between a target word and other words in a given lexicon – is also associated with higher accuracy and more stable word production (e.g. Sosa & Stoel-Gammon, 2012), as well as with lower age of acquisition and better target retention in experimental paradigms (Storkel, 2009; Storkel & Lee, 2011). Such effects are separable from those of word frequency, despite a high correlation between these variables (i.e. high frequency words are usually high density; Storkel, 2004). A dominant explanatory account of this effect is that high neighbourhood density words contain regular sound patterns that are held in short-term memory more precisely during initial processing (e.g. the *at* in *mat, cat*, and *catch*; Gathercole, Frankish, Pickering, & Peaker, 1999). This supports the subsequent formation of detailed phonological word representations in long-term memory, which may in turn provide fine-grained motor plans (Hoover, Storkel, & Hogan, 2010; Metsala & Walley, 1998).

Word frequency and phonological neighbourhood density are also reported to interact in early word learning. For instance, Hollich, Jusczyk, and Luce (2002) and Storkel (2004) report that high neighbourhood density predicted successful acquisition and production for low though not high frequency words. This suggests that high neighbourhood density is important when word frequency is low but that high frequency nullifies the high neighbourhood density advantage. Both neighbourhood density and frequency are also considered to interact with age (Jones & Brandt, 2019a; Braginsky et al., 2019). For instance, in a study of 300 British English-speaking children, Jones and Brandt (2019a) found that high neighbourhood density and high frequency were more strongly associated with caregiver reports of word production at 12 months than at 25 months. Whether or not similar interactions are associated with degrees of early spontaneous word production accuracy and variability remains unknown.

7.2.1 The current study

This study estimates spontaneous word production accuracy and variability rates in longitudinal data from five American English-speaking children. I present a classification of spontaneous word productions in terms of: (i) accurate and stable; (ii) accurate but variable; (iii) inaccurate but stable; and (iv) inaccurate and variable. The purpose of this analysis is to contribute to on-going discussion regarding the rates of accuracy and variability that can be expected in the normal range. Given widespread disagreement in the prior literature in this area (e.g. Holm et al., 2007; Sosa, 2015), I made no predictions regarding the results of this analysis. I also present an analysis of accuracy and variability rates modelled as a function of child age, input frequency, and ambient language phonological neighbourhood density, both as main effects and in interaction. Based on the literature reviewed, my predictions were that word production accuracy would increase with age, while production variability would decrease with age. I predicted that high frequency and high neighbourhood density would be associated with greater production accuracy and stability, and that these associations would be stronger in earlier periods of sampling. Finally, high neighbourhood density was expected to be more strongly associated with accurate and stable production for low frequency words.

7.3 Method

7.3.1 Corpus

This study examined accuracy and variability rates in spontaneous speech recorded in the Providence corpus (Demuth & McCullough, 2009). The Providence corpus contains transcripts of 364 hours of audio and video recordings from six monolingual children (three girls, three boys) aged 0:11-4;0. Data from one child, Ethan, were excluded from the current analysis given this child's diagnosis of Asperger's Syndrome at age five. From the onset of first words, children were recorded for a minimum of one hour every two weeks during interaction with their caregivers, ordinarily their mothers. Details of each child's data are shown in Table 7.1.

Table 7.1: Corpus summary. Showing total recorded utterances and glosses, mean length of utterance (MLU) in morphemes, and usable token and type counts. Glosses identifies transcribed strings, whether or not these are words, e.g. 'mum', 'cat', 'hmm', 'haha', 'achoo'. Tokens and Types identify lexical items (e.g. 'mum', 'cat') for which independent variable data was available; see *Independent variables: Age, frequency, and neighbourhood density*.

Speaker	Age (months)	Utterances	Glosses	MLU	Tokens	Types
Alex	16-41	29,251	63,727	2.31	31,150	1434
Lily	13-48	40,027	105,003	3.07	58,088	2011
Naima	11-46	43,499	145,783	4.03	72,280	2765
Violet	14-47	17,296	41,924	2.92	20,750	1533
William	16-40	21,291	46,508	2.38	26,361	1314

Transcript format was the major motivation for using the Providence corpus. Recordings are narrowly transcribed in the International Phonetic Alphabet (IPA), and produced word forms are listed alongside target forms. This makes it straightforward to calculate production accuracy and variability scores. A second motivation for using the Providence corpus was that transcription reliability for the corpus is high. After initial transcription, a second trained coder transcribed a sample of 10% of each recording, with inter-rater reliability reported between 80-98%. Given its high suitability to early word accuracy and variability research, it is perhaps unsurprising that the Providence corpus has been used in related previous work. Notably, Ota and Green (2013) analysed the effect of input frequency on the production of consonant clusters by three children in this corpus.

7.3.2 Data preparation

Providence corpus data files in Phon software format (Hedlund & Rose, 2019) were accessed via the project website (https://phonbank.talkbank.org/access/Eng-NA/Providence.html) and converted to .csv files in Phon to enable further pre-processing and modelling in R (R Core Team, 2016). Raw .csv files are hosted on the associated project repository alongside an R script allowing readers to re-create all analyses reported in the current study (https://osf.io/w9y27/). These files contain the following columns for each word token: (i) participant name; (ii) participant age; (iii) orthographic word; (iv) IPA target word; and (v) IPA produced form. Analysis in R began with the removal of non-lexical items including conversational sounds such as ‘hmm’, ‘haha’, and ‘achoo’.

7.3.2.1 Independent variables: Age, frequency, and neighbourhood density

Independent variable preparation then proceeded with the transformation of participant age into an appropriate format for statistical modelling, e.g. from ‘P1Y10M24D’ to ‘16’ (months). Using the childesr package in R (Sanchez et al., 2019), frequencies for each word produced by each child were then calculated from all American English caregiver transcripts in the CHILDES database (MacWhinney, 2000) in which the children addressed were aged between 22 and 36 months. This included 2,194,651 word tokens and 21,981 word types. Raw counts from this corpus were then log-plus-one transformed. Finally, I retrieved phonological neighbourhood density values for each produced word. In many developmental studies, phonological neighbourhood density is operationalised as the number of words in a given corpus that can be formed by the addition, substitution, or elimination of a single phoneme in a target word, e.g. *cat* neighbours *hat*, *cot, can,* and *catch* (e.g. Stokes, 2014; Storkel, 2004; following Luce & Pisoni, 1998). A general limitation of this operational definition, however, is that it may result in a substantial proportion of words in a given corpus being categorised as lexical hermits with zero neighbourhood density (Suárez, Tan, Yap, & Goh, 2011). Accordingly, the current study adopted a metric of word-level phonological similarity called phonological Levenshtein distance, or PLD20, defined as the mean number of additions, substitutions, or eliminations of phonemes required to change a particular word into its nearest twenty phonological neighbours (Suárez et al., 2011, p. 606). PLD20 values for each word produced were calculated across words in the English Lexicon Project, which provides lexical characteristic data for 40,481 words and which may be considered representative of the ambient language (Balota et al., 2007; retrieved from: http://www.talyarkoni.com/downloads/pld20.txt). The PLD20 metric is operationalised continuously in order to maximise statistical power. In contrast, the common approach of splitting tokens into high- and low-density groups has the effect of reducing statistical power, and limiting the quality of inferences that can be drawn. In contrast to plus/minus one-phoneme metrics of word-level phonological similarity (e.g. Luce & Pisoni, 1998), where a high value equals greater density, a high PLD20 indicates greater phonological distance between a target and its nearest neighbours, or low neighbourhood density. Different criteria of word-level phonological similarity such as the plus-minus-one phoneme criterion and PLD20 are highly correlated, and have been shown to confer analogous effects (Suárez et al., 2011). I therefore strongly expect the results reported below to hold across alternative measures of neighbourhood density.

7.3.2.1.1 Predictor correlation and multicolinearity

One limitation of the use of observational data without restriction to a particular target cluster, word, or utterance count is that it is difficult to mitigate the detrimental impact of high predictor correlation. High predictor correlation is an issue because it may cause multicolinearity, which manifests as a distortion of regression model results such as a substantial increase in the size of the estimate or the estimate error, or a shift in estimate direction (e.g. from a positive to a negative value). For this reason, researchers are commonly required to select only variables of personal theoretical interest for testing a specific hypothesis and to omit highly correlated variables that may be of general theoretical interest. In the current study, for instance, high rates of correlation motivated the omission of a word length variable and alternative word-sound variables including phonotactic probability (note that Storkel, 2004, among others takes the same approach; though see Storkel & Lee, 2011). Correlations between the three predictors included in this study are shown in Table 7.2.

Table 7.2: Pearson correlation matrix for independent variables.

	Age	Frequency	PLD20
Age	1	0.12	-0.05
Frequency	0.12	1	-0.33
PLD20	-0.05	-0.33	1

Notably, a moderate correlation was observed between PLD20 and word frequency (*r*=-0.33), with high frequency words commonly being high density (i.e. low PLD20). Multicollinearity risk was therefore tested by computing variance inflation factors (VIFs) using the lme4 and car packages in R (Bates, Maechler, Bolker, & Walker, 2015; Fox & Weisberg, 2011). These estimates suggested multicollinearity risk was low across predictors, with a maximum value of VIF = 2.61 for the frequency and neighbourhood density interaction term. Recommended maximum VIFs range from four to ten in the literature (e.g. Hair, Anderson, Tatham, & Black, 1995; Pan & Jackson, 2008). In a second assessment of multicolinearity risk I conducted a sensitivity analysis. This involved removing each predictor and re-fitting the main regression model (introduced fully below) to test for changes in the resulting coefficients. No substantial difference in estimates was found during this analysis, in terms of the direction or magnitude of estimates or the size of the estimate errors.

7.3.2.2 Dependent variables: Accuracy and variability

The dependent variables were word production accuracy and word production variability. Numerous operational definitions of each of these variables have previously been used (see Ingram, 2002, for review). In the current study, the Levenshtein distance between target and actual transcriptions was used as a measure of word accuracy. A word production identical to the listed adult form scored zero and lower accuracy was coded in terms of the number of phonetic insertions, substitutions, or deletions required to turn the produced form into the listed adult form. For instance, if the target word alligator listed /æləgeɪtəɹ/ was produced /ælɪgeɪɾə/, this production was scored a Levenshtein distance of three: One change from /ɪ/ to /ə/; one change form /ɾ/ to /t/; and the addition of /ɹ/. Levenshtein distance provides an accuracy metric that is not only intuitive but also computationally efficient. The measure also provides a graded picture of target and produced form distance, in contrast to the binary scoring of accurate and inaccurate forms using zeros and ones sometimes used (e.g. Macrae, 2013).

The second dependent variable of interest was word production variability. For this measure, I followed Ingram (2002, see p. 719 for examples) and used the proportion of whole-word variability (PWV) defined as the number of distinct productions of a word divided by the total number of productions. Where only one distinct form was produced, this form was attributed a variability score of zero. Using tidyverse package functions in R (Wickham, 2019), the data were grouped by child and age in months before calculating the degree of variability for each word, produced by each child, within each month of sampling.

The master dataset lists 234,551 word tokens (4360 types) with columns for: (i) speaker name; (ii) speaker age at word production; (iii) orthographic form of word produced; (iv) child-directed speech frequency; (v) phonological neighbourhood density, PLD20; (vi) IPA target form; (vii) IPA produced form; (viii) accuracy (Levenshtein distance); and (ix) variability (PWV) for the produced word in that month of age. This file is available from the project repository (https://osf.io/w9y27/).

7.3.3 Accuracy and variability profiles

One aim of this study was to adopt Grunwell's (1992) conventions to provide accuracy and variability profiles based on spontaneous speech from five children, without restriction to a particular lexical subset. To do this, conditional statements were used in R to divide all tokens from the master dataset into four classes, before calculating the proportion of produced words within each class for each child. For each of the 234,551 tokens produced, classification worked as follows:

1. If target / actual distance = 0 and variability = 0, then class = "Hit / stable"
2. If target / actual distance = 0 and variability > 0, then class = "Hit / variable"
3. If target / actual distance > 0 and variability = 0, then class = "Miss / stable"
4. If target / actual distance > 0 and variability > 0, then class = "Miss / variable"

As discussed in the introduction, there has been specific interest in the rate of words produced variably without hits in the typically developing population (i.e. statement 4 above; "Miss / variable"). Estimating this rate in spontaneous speech from typically developing children may improve our understanding of expected rates of accuracy and variability, and in turn help determine whether a high rate of variability without hits constitutes a useful clinical marker. Note that in this analysis accuracy is calculated for each word production, while variability is calculated across all productions of each word during each month.

During peer review two anonymous reviewers raised concerns that the coding method presented above may be too stringent. It was noted, for instance, that the production of /æləgeɪtəɹ/ as /ælɪgeɪɾə/ may be considered accurate as the vowel change from /ə/ to /ɪ/, the use of /ɾ/ instead of /t/, and the dropping of the word-final /ɹ/ do not constitute errors per se and may be attributable to dialectical variation. Along similar lines, it was suggested that requiring zero variability might be an unrealistic standard given that tokens are being collapsed across a month of sampling. These points are well taken, and in line with the reviewer suggestions I present a second accuracy and variability taxonomy with the modified standards listed below. Note that <= indicates ‘smaller than or equal to’, while >= indicates ‘greater than or equal to’.

1. If target / actual distance <= 1 and variability <= 0.1, then class = “Hit / stable”
2. If target / actual distance <= 1 and variability >= 0.1, then class = “Hit / variable”
3. If target / actual distance >= 1 and variability <= 0.1, then class = “Miss / stable”
4. If target / actual distance >= 1 and variability >= 0.1, then class = “Miss / variable”

These modified standards allow for minimal deviation from the listed adult form: A Levenshtein distance of zero or one phoneme, and variability of 10% across productions. I encourage readers to experiment further by modifying these threshold values (i.e. 1, 0.1) in the Boolean statements listed in the R code associated with this paper (https://osf.io/w9y27/).

7.3.4 Statistical modelling

The second analysis looks at child and lexical influences on word production accuracy and variability. To do this, the brms package (Bürkner, 2018) was used to fit two simple Bayesian regression models in R. In model one, accuracy (Levenshtein distance) was predicted by (i) the child’s age at production; (ii) the word’s child-directed speech frequency; and (iii) the word’s phonological neighbourhood density (PLD20) in the ambient language. In model two, variability (for each target word during each month of age) was predicted by (i) the child’s age at production; (ii) the word’s child-directed speech frequency; and (iii) the word’s neighbourhood density (PLD20) in the ambient language. Given known interactions between these variables (e.g. Storkel, 2004), interaction terms were also included for each combination of predictors, i.e. age:frequency, age:PLD20, and frequency:PLD20. All predictors were centred (i.e. $M = 0$) prior to model fitting. This explains the presence of zeros and negative values (e.g. negative frequencies) on the x-axes of the figures that follow (see *Modelling results*). In both models, brms package default priors were used (see R code), which I expected to be overwhelmed by the large number of observations (i.e. 234,551 cases). These models fitted well, with a large number of effective samples, stationery and well-mixing chains, rhats uniformly at 1, and good posterior predictive checks (see R code for detailed diagnostics, and the brms package documentation for a detailed description of diagnostic terminology; Bürkner, 2018).

The goal of modelling is to estimate parameters (e.g. β, the beta coefficient) that define the relationship between variables of interest – in this case the relationship between age, frequency, and neighbourhood denisty (main effects and interactions), and rates of spontaneous word production accuracy and variability. In Bayesian statistics the outcome of modelling is a probability distribution that describes the plausibility of different values of the parameter of interest (e.g. β). One motivation for this approach is that it communicates uncertainty in the data better than an emphasis on point estimates such as p values. Of particular interest in this study is the beta parameter estimate, β; i.e. the slope for the regression line for each predictor and response. A distribution for β bound above zero (e.g. 0.2 to 0.5) suggests a positive association between variables. That is, as the predictor value increases so does the response value; there is an upward-sloping regression line. A distribution for beta bound below zero (e.g. -0.5 to -0.2) suggests a negative association between the predictor and outcome, i.e. as the predictor value increases the response value decreases; there is a downward-sloping regression line. And a distribution for β spanning zero (e.g. -0.2 to 0.2) suggests no linear relationship between predictors, i.e. a flat regression line, is plausible.

7.4 Results

7.4.1 Accuracy and variability profiles

Table 7.3 shows accuracy and variability profiles for each child, based on the conventions developed by Grunwell (1992) and shown in the *Method, Accuracy and variability profiles* section as conditional statements (i.e. under the initial zero distance / zero variability criterion). Percentages of forms within each class do not differ substantially between children, despite differences in the rates of usable forms (see Table 7.1). Importantly, miss / variable rates were high across children, with inaccurate productions of variable words comprising between 61% and 72% of all productions. Hit / stable, i.e. consistently accurate, was the least common production type, ranging from 4% to 7%.

Table 7.3: Proportions of words produced within each accuracy and variability class, under the initial zero distance / zero variability criterion.

Speaker	Hit / stable	Hit / variable	Miss / stable	Miss / variable
Alex	0.04	0.16	0.10	0.71
Lily	0.06	0.18	0.14	0.61
Naima	0.04	0.17	0.09	0.69
Violet	0.07	0.18	0.15	0.61
William	0.04	0.14	0.11	0.72

In Table 7.4, I report the results of the categorisation using less strict criteria in which discrepancies of one phoneme and variability of up to 10% were allowed. To re-cap, this approach was prompted by a concern that the criteria generating the results shown in Table 7.3 were too stringent, and that these criteria would qualify fair deviations from the corpus target listing as errors. Table 7.4 shows that the proportion of words in each category again ranks similarly across participants. However, there is a substantial difference in the proportions across categories relative to those reported in Table 7.3. Under the revised criteria, hit / stable is the most common production type, followed by miss / stable, miss / variable, and hit / variable. Criteria selection therefore has a substantial impact on the shape of the taxonomy.

Table 7.4: Proportions of words produced within each accuracy and variability class, under revised criteria tolerating one-phoneme of distance and 10% variability.

Speaker	Hit / stable	Hit / variable	Miss / stable	Miss / variable
Alex	0.48	0.08	0.28	0.16
Lily	0.58	0.06	0.26	0.10
Naima	0.52	0.07	0.28	0.13
Violet	0.49	0.11	0.23	0.17
William	0.49	0.08	0.28	0.15

7.4.2 Modelling results

Regression model summaries are presented in the Appendix. Figure 7.1 shows marginal effects from model one, in which production accuracy (Levenshtein distance) was predicted by child age, child-directed speech frequency, and ambient language neighbourhood density (PLD20). Production accuracy improved with age, with a reduction in target / actual distance in later months of sampling (*β* = -0.03; lower 95% credible interval = -0.03; upper 95% credible interval = -0.03). Words that occurred at relatively high frequency in child-directed speech were produced more accurately than low frequency words (*β* = -0.12; lower 95% CI = -0.12; upper 95% CI = -0.12). Finally, words with many neighbours in the ambient language were produced more accurately than words with few neighbours in the ambient language (*β* = 0.10; lower 95% CI = 0.10; upper 95% CI = 0.10).

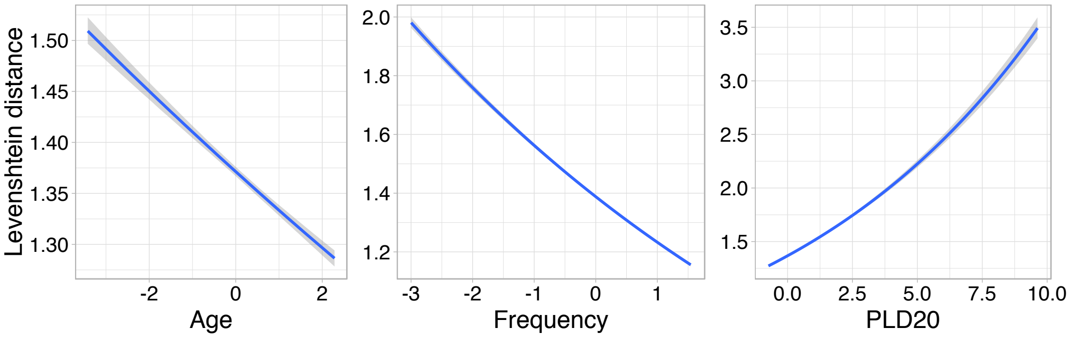

Figure 7.1: Associations between child age (0:11-4;0), word frequency, and phonological neighbourhood density (PLD20), and production accuracy (Levenshtein distance). Shading represents the 95% credible intervals, i.e. the range in which the parameter value falls with 95% probability. Note the scale differences on the *y*-axes.

I also tested for interactions between age, frequency, and PLD20 as predictors of word production accuracy. The results of this analysis are shown in Figure 7.2. No interaction was found between age and word frequency (β = 0.00; lower 95% CI = 0.00; upper 95% CI = 0.00), indicating that the strength of association between frequency and production accuracy did not change during the sampling period. The interaction between age and PLD20 was marginally negative (β = -0.01; lower 95% CI = -0.01; upper 95% CI = 0.00), indicating that low phonological distance (i.e. high density) is a more important predictor of accurate word production in early rather than late development. Finally, there was a negative interaction between word frequency and PLD20 (β = -0.02; lower 95% CI = -0.03; upper 95% CI = -0.02), indicating that as word frequency increased the association between high-density and word production accuracy weakened.

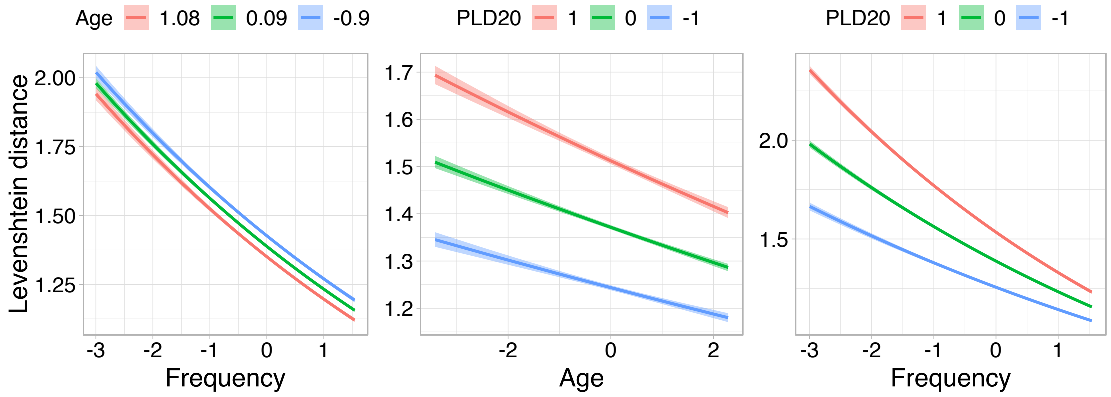

Figure 7.2: Interactions between: (i) child age (0:11-4;0) and frequency; (ii) child age and neighbourhood density (PLD20), and; (iii) frequency and neighbourhood density (PLD20), with respect to the accuracy response (Levenshtein distance). To ease the interpretation of interactions, age and PLD20 are binned into three levels by default by the brms package (i.e. high, mid, low). Shading represents the 95% credible intervals. Note the scale differences on the *y*-axes.

Figure 7.3 shows posterior probability distributions from model two, in which the proportion of whole-word variability (PWV) was predicted by child age, child-directed speech frequency, and neighbourhood density (PLD20). Production variability declined with age, with lower PWV scores in later months of sampling (β = -0.02; lower 95% credible interval = -0.03; upper 95% credible interval = -0.02). Words that occurred at relatively high frequency in child-directed speech were produced more stably than low frequency words (β = -0.48; lower 95% CI = -0.48; upper 95% CI = -0.47). Finally, there was a positive association between PLD20 and word variability (β = 0.10; lower 95% CI = 0.10; upper 95% CI = 0.11), indicating that words that sounded similar to many other words in the ambient language were produced more stably than words that sounded similar to few other words in the ambient language.

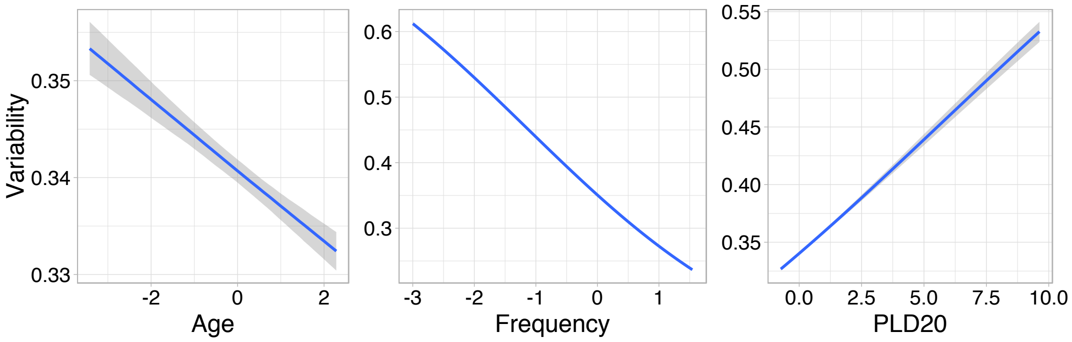

Figure 7.3: The associations between child age (0:11-4;0), word frequency, and phonological neighbourhood density (PLD20), and production variability. Shading represents the 95% credible intervals. Note the scale differences on the *y*-axes.

I also tested for interactions between age, frequency, and PLD20, as predictors of word production variability. The results of this analysis are shown in Figure 7.4. The interaction between age and frequency was marginally positive (β = 0.02; lower 95% CI = 0.01; upper 95% CI = 0.02), indicating that high frequency words tended to be produced stably across the sampling period, while low frequency words tended to be produced more stably in later months of sampling. The interaction between age and PLD20 was also positive (β = 0.02; lower 95% CI = 0.02; upper 95% CI = 0.03), indicating that in earlier months of sampling word productions were often variable regardless of a word's neighbourhood density, but that in later months variability was particularly high for low-density words. Finally, there was a positive interaction between word frequency and PLD20 (β = 0.01; lower 95% CI = 0.01; upper 95% CI = 0.02). This suggests that variability for a low frequency word was often high regardless of that word's PLD20. However, productions of high frequency words were marginally more stable for words that sounded similar to many other words in the ambient language.

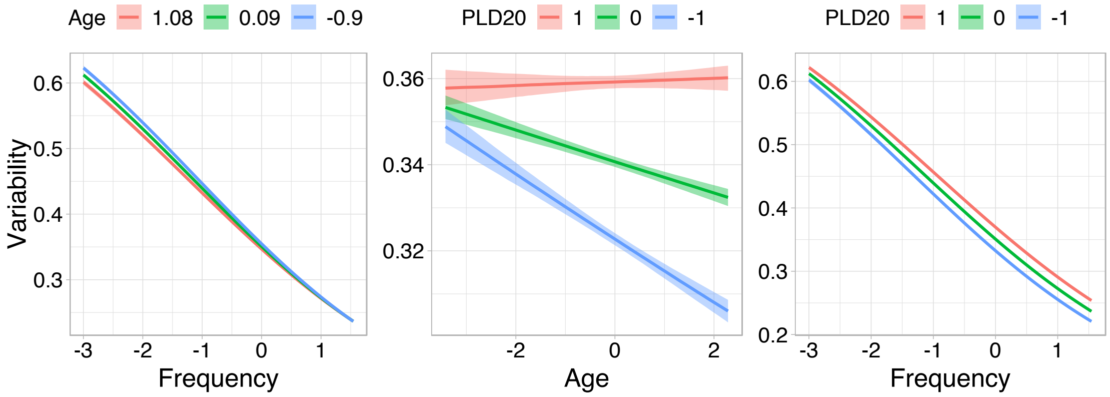

Figure 7.4: Interactions between: (i) child age (0:11-4;0) and frequency; (ii) child age and neighbourhood density (PLD20), and; (iii) frequency and neighbourhood density (PLD20), with respect to the variability response. To ease the interpretation of interactions, age and PLD20 are binned into three levels by default by the brms package (i.e. high, mid, low). Shading represents the 95% credible intervals. Note the scale differences on the *y*-axes.

7.5 Discussion

This study presented two analyses. First, I estimated overall rates of word production accuracy and variability in the spontaneous speech of five typically developing children recorded between the ages of 0;11 and 4;0. The aim here was to contribute to on-going debate regarding the rates of error and variability that may be expected in the typical range, and relatedly to debate regarding whether a high rate of word production inaccuracy and variability can provide a useful marker of speech sound disorder. Second, the study used Bayesian regression to model word production accuracy and variability as a function of age, frequency, neighbourhood density, and interactions between these variables. While these variables have previously been linked with word production accuracy and variability effects in experimental studies (e.g. Macrae, 2013), such effects remained poorly understood with respect to early spontaneous speech. The results from each analysis are discussed in the following sections.

7.5.1 Accuracy and variability profiles

Following Grunwell (1992) and others (e.g. Holm, Crosbie, & Dodd, 2007; McLeod & Hewett, 2008; Sosa, 2015), spontaneously produced words were categorised into four classes: (i) accurate and stable; (ii) accurate but variable; (iii) inaccurate but stable; and (iv) inaccurate and variable. These classes were populated according to two criteria. Under the first criterion, spoken word productions were classified as accurate and stable only if they did not differ from the listed adult form. This approach broadly replicates the experimental method of Sosa (2015, p. 28). The results of this analysis (Table 7.3) indicated high rates of error and variability broadly in line with Macrae (2013), McLeod and Hewett (2008), and Sosa (2015), and in contrast to the relatively low estimates presented by Holm et al. (2007). Under this criterion, up to three quarters of the words produced by children in the Providence corpus were variable without hits, which, as in some prior work (e.g. Sosa, 2015), was the most frequent production type. Apparently in direct contrast to Holm et al.'s (2007) claim that "inconsistency [i.e. variability without hits] is not a feature of normal development at any age" (p. 483), the results shown in Table 7.3 of the current study imply that young typically developing children are highly inconsistent in their early word productions. This in turn makes it reasonable to suggest, following Sosa (2015, p. 33), that overall variability rates – particularly rates of variability without hits – may not provide a useful index to aid the differential diagnosis of children with speech sound disorder.

Under a second criterion, however, spoken word tokens were classified as accurate despite differing from the listed adult target form by a single phoneme, and classified as stable across multiple productions up to a 10% variability threshold. These modifications to the initial criteria were prompted by concerns raised during peer review that the original approach may unfairly qualify reasonable deviations from the adult target listing as errors. This point is well taken, and closely examining the data I found a number of examples supporting the anonymous reviewers' claims: /mɑm/, for example, pronounced /mʌm/. Though unlikely to be considered erroneous by many listeners, such productions were classed as erroneous under the criterion first used.

As might be expected, the change in criterion had a dramatic impact on the accuracy and variability profiles derived (Table 7.4). Under the revised criterion, hit / stable was the most common production type (48%–58%), while miss / variable comprised between just 10% and 16% of productions. These figures appear broadly continuous with Holm et al.'s (2007) estimate of 13% variability in 409 typically developing children aged 3;0–3;5 (Holm et al., 2007) as well as with the previously-cited conclusion that production inconsistency is not a feature of typical language development. The revised results (Table 7.4) suggest that although young children do deviate minimally from the listed adult form used as an experimental standard, they remain generally accurate and stable in their spontaneous word productions, and this in turn constitutes support for the claim that a high overall inaccuracy and variability rate may be considered a valid marker of speech sound disorder (Holm et al., 2007).

Increasing the thresholds used during classification to a Levinshtein distance of one and variability of 10% successfully accommodates productions that deviate from the adult listed form but which are unlikely to be considered erroneous. However, this approach comes with a significant cost, as loosening the thresholds permits the classification of minimal errors as accurate forms. This may appear particularly damaging with respect to short words, which dominate the early productive lexicon. The mode length of words in the Providence corpus is three phonemes (*M* = 3.74), and 155,088 words – or 66% of the corpus – comprise three phonemes or fewer. For such words one discrepant phoneme may represent a substantial erroneous deviation, which is ignored under the relaxed thresholds. For example, instances classed as accurate productions under the revised criteria include the production of /bæg/ as /bæk/, the production of /bæθ/ as /bæ/, and the production of /bæt/ as /bɛt/. Inspecting the data, such instances appear far from exceptional. Setting a hard and fast decision boundary for the quantification of accurate and variable spoken word forms therefore involves a difficult trade-off: (i) categorise a level of apparently tolerable production deviance as erroneous or (ii) categorise minimal production errors as accurate. When comparing child productions to adult listed forms, this trade-off would apparently exist whether considering large-scale spontaneous speech data such as that of the current study or small-scale elicited speech data, such as that of prior experimental studies (e.g. Sosa, 2015; Holm et al., 2007).

Unfortunately it is impossible to select between the contrasting taxonomies presented in this study on the basis of the current or existing data. Each underlying criterion is well justified, and each generates a taxonomy with proportions of accuracy and variability broadly continuous with those previously reported (e.g. Sosa, 2015; Holm et al., 2007). Given the apparent sensitivity of spoken word classification to minimal changes in the underlying criterion, and given the extensive discrepancies in rates of accuracy and variability reported in the existing literature, it remains unclear whether accuracy and variability profiles can provide a useful method of identifying speech sound disorder. It may well be, as Sosa (2015, p. 32) writes, that:

> The use of phonetic transcription to quantify [accuracy and] variability is too unreliable to be used for differential diagnosis of speech sound disorder; more refined acoustic and/or kinematic analysis methods may be needed.

Establishing a robust method of quantifying the degree of spoken word accuracy and variability that occurs in early naturalistic and elicited speech constitutes an important part of the future research agenda, both for our understanding of typical and atypical language development and for the purposes of assessment and intervention. One conceivable direction for future research would be to collect large samples of child, spontaneous and elicited speech data, and then to record accuracy judgements for specific spoken word tokens within that data from a large group of impartial, adult listeners. Listeners would hear spoken word instances and identify (for instance via button pressing) whether they considered each token to be accurate or inaccurate (e.g. /æləgeɪtəɹ/ produced /ælɪgeɪtəɹ/ and /ælɪgeɪɾə/). These accuracy judgments could then form a basis for the classification of spoken word tokens into accuracy and variability taxonomies. This study would be highly resource intensive and the method clearly could not be applied directly in clinical contexts. However such an approach may provide the baseline data needed to break the apparent deadlock and help resolve widespread disagreement in the existing literature on early spoken word accuracy and variability rates.

7.5.2 Age and lexical effects on accuracy and variability

The second part of this study looked at child and lexical influences on early rates of spoken word accuracy and variability. The child-related predictor of interest was age, which has previously been reported to have a positive association with word production accuracy and a negative association with word production variability (Holm et al., 2007; Macrae, 2013). The current analysis is the first to confirm that these findings hold in longitudinal spontaneous speech across a sampling age range considerably larger than that of any prior study (i.e. 0:11-4;0). That is, I reported that both error and variability were lower in the later months sampled. The first lexical predictor of interest was child-directed speech frequency, a variable central to the study of early language acquisition which has previously been positively linked to high spoken word accuracy and stability (e.g. Sosa & Stoel-Gammon, 2012). The current study also replicated this finding, reporting a robust association between high frequency, and better accuracy and reduced variability. The second lexical predictor of interest was phonological neighbourhood density (PLD20), which has been positively linked to memorisation and production advantages (e.g. Storkel, 2004, 2009; Storkel & Lee, 2011). As in Sosa and Stoel-Gammon, (2012), I reported heightened accuracy and reduced variability for high neighbourhood density words, and in doing so demonstrated that previous findings scale-up when assessing children's spontaneous speech without restriction to a spoken word token limit (e.g. Sosa & Stoel-Gammon, 2012, assessed 30 elicited words).

Modelling also suggested interactions between age, word frequency, and phonological neighbourhood density as predictors of spoken word accuracy and variability. For instance, I reported that in early months of sampling productions were highly variable regardless of the words' neighbourhood density but that in later months of sampling high-density (i.e. low PLD20) words were substantially less variable (Figure 7.4, centre panel). The current study also corroborated previous work by Hollich et al. (2002), and Storkel (2004), who found that high neighbourhood density predicted word acquisition for low- but not high-frequency words. Such findings contribute to a developing picture of high word exposure frequency as a primary force driving growth of the productive lexicon, and alternative word characteristics such as neighbourhood density ‘stepping in’ and supporting learning when exposure frequency is relatively low. The results of this study show for the first time that this effect extends to the accuracy of children’s early spontaneous word productions (Figure 7.2, right panel).

With the exception of Ethan, who was excluded from the current analyses on the basis of a later diagnosis of Asperger’s Syndrome, children in the Providence corpus are typically developing. It is therefore important to understand the associations between predictors and accuracy and variability rates reported here as part of a typical trajectory, i.e. not within the framework of speech sound disorders. Explanatory accounts compatible with the results reported in this study emphasise oral-motor maturity and a shift from holistic to segmental word representations. There is evidence that children’s oral-motor skills are associated with their language skills independently of their general cognitive abilities (Alcock, 2006). Thus, although further experimentation is required, it may be reasonable to assume that increases in production accuracy and stability with age are to some degree attributable to improved control of the articulators. In addition, early word phonology may be memorised only approximately as a result of working memory limitations or initial focus on relatively holistic word features (Metsala & Walley, 1998), and accurate and stable production will be compromised in the absence of a mental representation detailed enough to provide a solid motor plan. High-frequency, high-density words hold an advantage in this early trajectory of oral-motor and cognitive development because they are repeatedly encountered and encoded in memory both explicitly and implicitly. High input frequency, high-density words are also likely to be produced by the child more regularly, and contain familiar sound patterns that may require minimal articulatory and cognitive recourses. Assessing this emergent explanatory account remains an important on-going research line, both for its contribution to our general understanding of the interaction between early oral-motor and cognitive development, and for our understanding of how to improve phonological word representation and production in children with developmental language disorder. The current study constitutes an important addition to our understanding of how developmental stage, exposure frequency, and phonological neighbourhood density influence the dynamics of early word production accuracy and variability.

7.5.3 Limitations

Despite its contributions to the literature the current study has a number of limitations. I have already discussed the issue of multicolinearity, which prevents against the inclusion of alternative variables of theoretical interest such as phonotactic probability. This is regrettably unavoidable, and I can only encourage readers to use the published code to experiment with different configurations of predictor variables that may be of personal theoretical interest (https://osf.io/w9y27/). A second issue is that at a number of points in this manuscript I draw parallels between the estimates I derived under different criteria of accuracy and variability and the estimates reported in prior studies (e.g. Sosa 2015; Holm et al., 2007). However, given differences in the sampling methods of the current study and previous work presenting accuracy and variability profiles (e.g. Holm et al., 2007, evaluated variability across only three repetitions), such comparisons are imperfect. It is acknowledged, for instance, that there is more opportunity for word production error and variability in spontaneous speech than there is during elicitation tasks, and that the possibility of spoken word error or variability grows with rates of production, which are not uniform across words in the corpus (McLeod & Hewett, 2008). Furthermore, it is noted that the Providence corpus uses a narrow transcription, while some prior studies (e.g. Sosa, 2015) have made broad transcriptions. Finally, in the introduction to the current study it was argued that the analysis of spontaneous speech data unrestricted by a target cluster, word type, or utterance count can provide insight beyond the analysis of a small number of target words elicited in an experimental setting, for instance by supporting or challenging the validity of such experimental data. I believe the current manuscript to have delivered on this claim not only by raising important questions regarding methods of quantifying early accuracy and variability rates, but also by showing for the first time that a range of findings from the early word acquisition literature (e.g. age, frequency, neighbourhood density effects, and interactions) can be found in accuracy and variability rates derived from large-scale, longitudinal, naturalistic word production data. That said, an important trade-off of the use of longitudinal data with a relatively high sampling rate such as the Providence corpus is that because the collection and transcription of such data is both challenging and highly time-consuming participant numbers are often low. A further limitation of the current study is, therefore, that it includes data from just five children, making it difficult to extrapolate findings to the broader population.

7.6 Conclusion

This study examined rates of spontaneous word production accuracy and variability with respect to three predictor variables: Child age, word frequency, and word phonological neighbourhood density. Increases in accuracy and decreases in variability between the ages of 11 months and four years were interpreted within a framework of early memory and oral-motor development – a trajectory within which high exposure frequency and high neighbourhood density confer acquisition and production advantages. I also presented two taxonomies of early accuracy and variability rates that highlighted the difficulty of setting hard and fast error discrimination thresholds. I proposed an accuracy judgement study that may address this issue and help resolve widespread disagreement regarding the rates of accuracy and variability expected within the typical range. Without such normative data it may be difficult to determine the validity of measures of spoken word accuracy and variability used in research and clinical settings.

7.7 References

Ainsworth, S., Welbourne, S., & Hesketh, A. (2016). Lexical restructuring in preliterate children: Evidence from novel measures of phonological representation. *Applied Psycholinguistics*, *37*(4), 997–1023. https://doi.org/10.1017/S0142716415000338

Alcock, K. (2006). The development of oral motor control and language. *Down Syndrome Research and Practice*, *11*(1), 1–8. https://doi.org/10.3104/reports.310

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/17958156

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. http://www.jstatsoft.org/v67/i01/

Betz, S. K., & Stoel-Gammon, C. (2005). Measuring articulatory error consistency in children with developmental apraxia of speech. *Clinical Linguistics and Phonetics, 19*(1), 53–66. https://doi.org/10.1080/02699200512331325791

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind, 3,* 52–67. https://doi.org/10.31234/osf.io/cg6ah

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411. doi.org/10.32614/RJ-2018-017

Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language*, *36*(01), 173–200. https://doi.org/10.1017/S0305000908008921

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2002). Diagnostic Evaluation of Articulation and Phonology–U.S. Edition (DEAP) Technical Report. San Antonio, TX: Pearson.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The Interaction Between Vocabulary Size and Phonotactic Probability Effects on Children's Production Accuracy and Fluency in Nonword Repetition. *Journal of Speech, Language, and Hearing Research, 47*(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language*, *51*, 419–439. https://doi.org/10.1017/CBO9780511980503.007

Fox, J. & Weisberg, S. (2011). *An {R} Companion to Applied Regression*, Second Edition. Thousand Oaks California: Sage. http://socserv.socsci.mcmaster.ca/ jfox/Books/Companion

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *25*(1), 84–95. https://doi.org/10.1037/0278-7393.25.1.84

Goffman, L., Gerken, L., & Lucchesi, J. (2007). Relations Between Segmental and Motor Variability in Prosodically Complex Nonword Sequences. *Journal of Speech, Language, and Hearing Research*, *50*(2), 444–58. https://doi.org/10.1044/1092-4388(2007/031)

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 649–660. https://doi.org/10.1037/0096-1523.25.3.649

Grunwell, P. (1992). Assessment of child phonology in the clinical context. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 457 – 483). Timonium, MD: York.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate Data Analysis* (3rd edition). New York: Macmillan.

Hedlund, G., & Rose, Y. (2019). Phon 3.0. [Computer software]. Available from https://www.phon.ca/phon-manual/misc/Welcome.html

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In B, Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 314–323). Boston, MA: Cascadilla Press.

Holm, A., Crosbie, S., & Dodd, B. (2007). Differentiating normal variability from inconsistency in children's speech: Normative data. International Journal of Language and Communication Disorders, *42*(4), 467–86. https://doi.org/10.1080/13682820600988967

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language*, *29*(04), 713–733. https://doi.org/10.1017/S0305000902005275

Jones, S. D. &, Brandt, S. (2019a). Do children really acquire dense neighbourhoods? *Journal of Child Language, 46*(6), 1260–1273. https://doi.org/10.1017/S0305000919000473

Kent, R. D. (1992). The biology of phonological developement. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 65–90). Timonium, MD: York Press.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Macrae, T. (2013). Lexical and child-related factors in word variability and accuracy in infants. In *Clinical Linguistics and Phonetics, 27*(6–7), 497–507. https://doi.org/10.3109/02699206.2012.752867

Macrae, T., & Sosa, A. V. (2015). Predictors of token-to-token inconsistency in preschool children with typical speech-language development. *Clinical Linguistics & Phonetics*, *29*(12), 922–937. https://doi.org/10.3109/02699206.2015.1063085

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. Retrieved from http://talkbank.org/manuals/CLAN.pdf

McLeod, S., & Hewett, S. R. (2008). Variability in the production of words containing consonant clusters by typical 2- and 3-year-old children. *Folia Phoniatrica et Logopaedica*, *60*(4), 163–172. https://doi.org/10.1159/000127835

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum.

Munson, B., Edwards, J., & Beckman, M. E. (2005). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research*, *48*(1), 61–78. https://doi.org/10.1044/1092-4388(2005/006)

Ota, M., & Green, S. J. (2013). Input frequency and lexical variability in phonological development: A survival analysis of word-initial cluster production. *Journal of Child Language*, *40*(03), 539–566. https://doi.org/10.1017/S0305000912000074

Pan, Y, & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology and Infection, 136*, 421–431.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(2), 1928–1941. Retrieved from https://psyarxiv.com/93mwx

Sosa, A. V., & Stoel-Gammon, C. (2006). Patterns of intra-word phonological variability during the second year of life. *Journal of Child Language*, *33*(1), 31–50. https://doi.org/10.1017/S0305000905007166

Sosa, A. V. (2015). Intraword variability in typical speech development. American Journal of Speech-Language Pathology, *24*(1), 24–35. https://doi.org/10.1044/2014_AJSLP-13-0148

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596–608. https://doi.org/10.1044/1092-4388(2011/10-0113)

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

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221. https://doi.org/10.1017/S0142716404001109

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

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review*, *18*(3), 605–11. https://doi.org/10.3758/s13423-011-0078-9

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Ventura, P., Kolinsky, R., Fernandes, S., Querido, L., & Morais, J. (2007). Lexical restructuring in the absence of literacy. *Cognition*, *105*(2), 334–361. https://doi.org/10.1016/j.cognition.2006.10.002

Walley, A. C. (1993). The role of vocabulary development in children′s spoken word recognition and segmentation ability. *Developmental Review*, *13*(3), 286–350. https://doi.org/10.1006/drev.1993.1015

Wickham et al., (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Chapter 8 Density and Distinctiveness in Early Word Learning: Evidence from Neural Network Simulations

Linking statement: Chapters four to Six have identified individual differences in the ability to represent spoken words that are attributable to age, clinical profile, and lexical influences. In particular, these studies have shown that high-density words are usually better represented and therefore more accurately produced than low-density words. In this chapter I simulate the high-density bias in a neural network and provide an interpretation of network performance that accommodates conflicting behavioural evidence of high distinctiveness word learning advantages. This chapter strongly informs the development of the explanatory account of the high-density bias linking each empirical study of this thesis.

8.1 Abstract

High phonological neighbourhood density has been associated with both advantages and disadvantages in early word learning. High density may support the formation and fine-tuning of new word sound memories; a process termed lexical configuration (e.g. Storkel, 2004). However, new high-density words are also more likely to be misunderstood as instances of known words, and may therefore fail to trigger the learning process (e.g. Swingley & Aslin, 2007). To examine these apparently contradictory effects, we trained an autoencoder neural network on 587,954 word tokens (5497 types; including mono- and multi-syllabic words of all grammatical classes) spoken by 279 caregivers to English-speaking children aged 18 to 24 months. We then simulated a communicative development inventory administration and compared network performance to that of 2292 children aged 18 to 24 months. We argue that autoencoder performance illustrates concurrent density advantages and disadvantages, in contrast to prior behavioural and computational literature treating such effects independently. Low network error rates signal a configuration advantage for high-density words, while high network error rates signal a triggering advantage for low-density words. This interpretation is consistent with the application of autoencoders in academic research and industry, for simultaneous feature extraction (i.e. configuration) and anomaly detection (i.e. triggering). Autoencoder simulation therefore illustrates how apparently contradictory density and distinctiveness effects can emerge from a common learning mechanism.

8.2 Introduction

Words with high phonological neighbourhood density (i.e. words that sound similar to many other words in the language to which children are exposed) are learned developmentally earlier and remembered and produced more accurately than words with low phonological neighbourhood density (Fourtassi, Bian, & Frank, 2018; Hollich, Jusczyk, & Luce, 2002; Stokes, 2014; Storkel, 2004). One way to understand this effect is in terms of long-term auditory priming (e.g. Church & Fisher, 1998). In this account, phonological representations of words heard in child-directed and overheard speech are formed in the child's long-term memory (Port, 2007). These representations may be perceptual, meaning that they are stored without semantic details, or they may be conceptual, meaning that they are stored with semantic details. High neighbourhood density words are memorised more easily than low neighbourhood density words because high-density words contain sound features that are well represented in existing perceptual and conceptual word memories. The novel high-density word *coal*, for instance, may be acquired through analogy to existing memories including *coat*, *pole*, *cone*, *hole, code*, and *mole* (Church & Fisher, 1998).

One challenge for research in early word learning has been to reconcile evidence of a high-density word learning advantage with contrasting evidence of a high-density word learning *dis*advantage in specific contexts (e.g. Stager & Werker, 1997; Swingley & Aslin, 2007). Swingley and Aslin (2007), for instance, found that children aged 1;6 (one year, six months) struggled to associate phonologically similar labels (e.g. *tog*, neighbouring the known word *dog*) to novel objects and reported a learning advantage for distinctive stimuli with no or very few phonological neighbours (e.g. *meb*). One interpretation of this finding is that children may misidentify a novel high-density word as an instance of a known neighbour, particularly in the absence of additional cues to support word leaning, such as a sentence frame or speaker gaze. This behavior is generally adaptive because stored word sound memories and related perceptual mechanisms must be flexible enough to support cross-contextual comprehension on the fly, for instance when a learner encounters a known word in an unfamiliar dialect (Church & Fisher, 1998). Furthermore, the number of minimally different words that young children know and hear regularly in the speech directed to them is limited (Guevara-Rukoz et al., 2018), and this makes it reasonable to classify a novel sound sequence that is very similar to a known word as an instance of that known word instead of as an instance of an unknown word (Swingley & Aslin, 2007).

Overall, then, the evidence suggests that phonological density and phonological distinctiveness support different aspects of word learning. Phonological distinctiveness supports the *triggering* of word learning, in which potential targets of acquisition are identified. Phonological density meanwhile supports lexical *configuration*, or the formation and ongoing fine-tuning of sound memories for these words. These effects have commonly been treated separately, as in the aforementioned studies by Storkel (2004) and Swingley and Aslin (2007), and in related work by Hoover, Storkel, and Hogan (2010) and McKean, Letts, and Howard (2014). Furthermore, there has been a tendency to frame evidence of either a high-density or high-distinctiveness learning advantage as evidence against the opposing effect (e.g. as in Vitevitch & Storkel's, 2013, p. 520, reference to Swingley & Aslin, 2007). The purpose of the current study is to provide a unified framework for understanding apparently contradictory density and distinctiveness effects in early word learning. We use a simple autoencoder neural network to illustrate how these effects can emerge from a common underlying mechanism.

The current study was motivated by Vitevitch and Storkel (2013), who examined neighbourhood effects in early word learning by training and testing an autoencoder on a small number of monosyllabic non-words (*N*=60), which were dichotomised into high-density and low-density groups. One novel contribution of the current study is to determine how the high-density advantage reported by Vitevitch and Storkel (2013) scales when using sizeable naturalistic data. In order to make the training data representative of young children's input, we trained an autoencoder on 587,954 word tokens (5497 word types) spoken by 279 caregivers to English-speaking children aged 18 to 24 months. This age range was selected to reflect participants in the aforementioned literature on density and distinctiveness effects (e.g. Storkel, 2004; Swingley & Aslin, 2007). The training data included mono- and multi-syllabic words from all grammatical classes, for instance nouns, verbs, adjectives, and prepositions. To test the trained network, we simulated a MacArthur-Bates communicative development inventory administration (Fenson et al., 2007). Then, to validate network performance, we compared the results of this simulated administration to those from 2292 real administrations involving children aged 18 to 24 months. Note that this validation phase was not possible in prior work using non-words (Vitevitch & Storkel, 2013). In addition to testing the network's ability to represent and output trained words, we also tested the network's ability to generalise and process new, previously untrained words. In all phases, neighbourhood density was modeled continuously, avoiding dichotomisation that can reduce statistical power and limit the quality of inferences drawn.

Our interpretation of network performance is informed by our understanding of the application of autoencoders in academic research and industry. Autoencoders are a class of neural networks in which – in three-layer instantiations – input is received in the first layer, compressed in a second 'hidden' layer, and then reconstructed in a third output layer. Autoencoders learn through back propagation, updating between-layer connection weights in order to reduce input-output error.

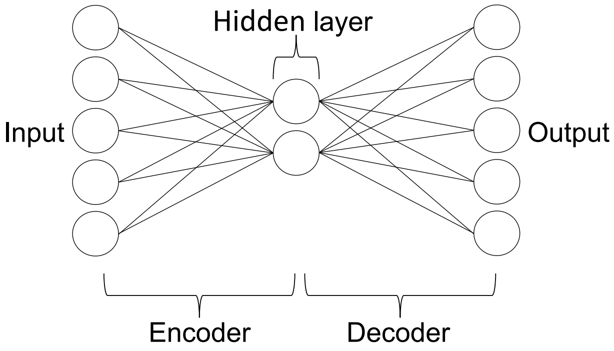

Figure 8.1: A simplified autoencoder architecture.

Autoencoders show large error when there is a big difference between the input data representation and the output data representation. Importantly, whether or not high network error is undesirable depends on the task at hand. Low error indicates that a given data point has features consistent with the well-represented properties of the previous network input, such as the dominant features in a set of images or the semantic or phonological features common across a set of words. In the context of neighbourhood density effects, the low error rate reported by Vitevitch and Storkel (2013) represents a configuration advantage for high-density words. However, high error may be considered advantageous when the purpose of the autoencoder is to detect anomalies. For example, in credit card fraud detection, an autoencoder may be trained on non-fraudulent transactions only, with both non-fraudulent and fraudulent transactions subsequently presented and the latter prompting an increase in error rate. Similarly, in a categorisation task simulation, the network may habituate to a set of similar stimuli and de-habituate on presentation of an anomalous stimulus. In each case, high error rates indicate that a novel data point (i.e. a transaction or stimulus) is unlikely to be a member of any trained class. In the context of simulating neighbourhood density effects in early word learning, a spike in error rate indicates that a novel string is unlikely to be an instance of any previously trained word. And in this sense, high autoencoder error provides a strong analogy to the triggering advantage for distinctive stimuli observed in human participants (e.g. Swingley & Aslin, 2007).

A broad similarity may be seen between the computational approach used in this study and behavioural paradigms such as the naming task, in which participants must accurately read a word or verbally label a stimulus, or the non-word repetition task, in which participants must accurately repeat a nonsense auditory word stimulus. In each case, lower error rates are taken as evidence of better-memorised properties of the input. However, we want to emphasise that the focus of this report is a simple model of word sound memory configuration and associated triggering effects, rather than an explicit model of word comprehension or production. In addition, we remain agnostic regarding the nature of actual word sound representations, for instance prototypes, exemplars, or hybrids (see Ambridge, 2019, for discussion).

8.3 Method

8.3.1 Network specification

A full network specification can be retrieved via the R code hosted on the Open Science Framework repository associated with this project (https://osf.io/2qk5j/). We used the h2o machine learning platform (H2O.ai, 2016) to build an autoencoder with rectified linear unit activation functions, a learning rate of .1, one thousand training epochs, and randomised initial weights. These parameters make our network broadly comparable to that of Vitevitch and Storkel (2013). Our autoencoder had 114 input nodes and 114 output nodes; a number determined through the numerical encoding of words from the training corpus (see *Training*). In a basic sensitivity analysis, we compared networks with 10, 20, and 30 hidden-layer nodes, i.e. with smaller or larger processing resources. Having observed equivalent main effects we settled on a hidden-layer size of 20 nodes.

8.3.2 Training

The autoencoder was trained on 587,954 tokens (5497 mono- and multi-syllabic unique word types, including all grammatical classes) from child-directed speech from 279 caregivers, directed at American English-speaking children aged 18 to 24 months. These data were retrieved from the Child Language Data Exchange System (CHILDES) using the childesr package in R (MacWhinney, 2000; Sanchez et al., 2019). For each word type we extracted a machine-readable phonological encoding (i.e. a string of 0s and 1s; an example follows) from the pre-embedded Medical Research Council (MRC) dictionary hosted as part of the PyPatPho package (Coltheart, 1981; Grimm & Tulkens, 2015; see https://github.com/RobGrimm/PyPatPho). Only words listed in this database were included in the training inventory. These numerical encodings were generated using PatPho via PyPatPho in Python (Grimm & Tulkens, 2015; Li & MacWhinney, 2002). PatPho converts words into 114-unit binary value vectors on the basis of a range of articulatory features (e.g., voiced, voiceless, front, back, labial, high, lateral, etc.) adopting a syllabic template scheme that accommodates input of varying length and therefore enabling us to model mono- and multi-syllabic words within a parallel architecture. Truncated example PatPho encodings for the words *cat* and *hat* are shown below. Note that encodings were fronted, meaning that word-initial features start at the beginning of the 114-digit vector.

/kæt/ = [0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0...0_{144}]
/hæt/ = [0 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0...0_{144}]

Shading identifies the portion of the vector containing the differences in 0s and 1s that map to the difference in the first phonemes of *cat* and *hat* (i.e. /k/ versus /h/). The subsequent string identity – continuous up to 114 digits – reflects the shared phonemes /æt/ and placeholders supporting the encoding of longer, multi-syllabic words. During training, the encoded child-directed speech corpus was passed to the network defined in *Network specification*.

8.3.3 Test

After training, we tested the network on a 586-item subset of the trained data that appear in the MacArthur-Bates communicative development inventory, words and sentences version (MCDI-WS; Fenson et al., 2007). The MCDI-WS contains a list of words and phrases and accompanying checkboxes under the response option 'produces'[3]. During real-world administration, caregivers are asked to tick the boxes next to the words that their child is able to say. We accessed the MCDI-WS data using the wordbankr package in R (Braginsky, Yurovsky, Frank, & Kellier, 2018; Frank, Braginsky, Yurovsky, & Marchman, 2017). The test word list was encoded using the process described in *Training*.

For each test word we calculated three independent variables: Phonological neighbourhood density, frequency, and length. Following Luce and Pisoni (1998), developmental researchers commonly define phonological neighbourhood density as the number of words in a given corpus that can be formed by the addition, substitution, or elimination of a single phoneme in a target word, e.g. *cat* neighbours *hat*, *cot, can,* and *catch*. A limitation of this approach, however, is that many of the words to which young children are exposed are 'lexical hermits' with zero plus/minus one-phoneme neighbourhood density. Accordingly, we used a continuous metric of similarity called phonological Levenshtein distance, or PLD20, defined as the mean number of additions, substitutions, or eliminations of phonemes required to change a particular word into its nearest twenty phonological neighbours (Suárez, Tan, Yap, & Goh, 2011, p. 606). PLD20 values for each test word were calculated using all words in the training corpus. A smaller PLD20 indicates greater phonological similarity (i.e. high density).

Frequency and length variables were also included in our statistical model because close association with neighbourhood density (i.e. high-density words are typically high frequency and short) makes it important to control statistically for these

[3] Note that we only tested MCDI-WS words and that MCDI-WS phrases were excluded from our analysis.

effects. Previous studies have also reported interactions between these variables. For instance, Storkel (2004) found a significant association between high phonological neighbourhood density and early age-of-acquisition for low- but not high-frequency words. In the current study, we used log token frequencies for each test word in the training inventory, and length was measured in number of phonemes. Alternative measures of word length, including number of letters, syllables, or morphemes, are highly correlated and may therefore provide similar results (Lewis & Frank, 2016). We selected the phoneme-based measure given the central interest in this unit of representation in the current study (i.e. as the basis of the PLD20 calculation).

The statistical analysis of test phase error rates was conducted in R (R Core Team, 2016) using the brms package (Bayesian regression models using Stan) (Bürkner, 2018). For all models, likelihood functions were selected on the basis of response variable distribution. In the test phase analysis, we fitted a multiple regression model with a lognormal likelihood, in which autoencoder mean squared error was predicted by word frequency, word length, phonological distance (PLD20), and interactions between PLD20 and word frequency and length (i.e. PLD20*frequency, PLD20*length). We used brms default priors, with each predictor centered and scaled prior to model fitting. This model fitted successfully, with a good number of effective samples, stationery and well-mixing chains, rhats uniformly at 1, and credible posterior predictive checks (see R code for full diagnostics, and the brms package documentation for further description of diagnostic terminology; Bürkner, 2018).

8.3.4 Validation

Using real words during training and test made it straightforward to compare network performance to data from children. We used the network's test-phase error rates to predict rates of word production among 2292 American English-speaking children aged 18 to 24 months, i.e. matched in age to the training inventory. That is, we compared the results of our simulated MacArthur-Bates communicative development inventory administration to a large database of completed, real-world administrations. This data was retrieved from the wordbank database using the wordbankr package in R (Braginsky et al., 2019; Frank et al., 2017; R Core Team, 2016). We calculated the proportion of children that were able to produce each test word and used this as the dependent variable in a Bayesian regression model in which the by-word mean squared error rate from our autoencoder was the independent variable. We used a gamma family likelihood and brms default priors, and the predictor was centered and scaled for model fitting (see R code for diagnostics).

8.3.5 Generalisation

In this phase, we exposed the trained network to 500 words it had not been trained on and measured the error rates for these items. Generalisation-phase words were randomly sampled from the Massive Auditory Lexical Decision (MALD) database (Tucker et al., 2018), and the degree of phonological similarity between each generalisation word and words in the training inventory was calculated using the PLD20 metric. The question addressed in this analysis was whether error rates were higher or lower for generalisation words that sounded relatively similar or dissimilar to words that the autoencoder had been trained on. We addressed this question using a Bayesian regression model in which generalisation word mean squared error rate was predicted by PLD20 and word length in phonemes. We used a skew-normal family likelihood and brms default priors, with predictors again centered and scaled for model fitting (see R code for diagnostics).

8.4 Results

We begin with the results from the test phase, in which we simulated a MacArthur-Bates communicative development inventory (MCDI-WS) administration on an autoencoder trained on a large corpus of authentic child-directed speech (see Appendix E for model summaries). We found main effects for each predictor, which are visualised as posterior probability distributions in Figure 8.2. High reconstruction error rates were associated with: (i) Long word length in phonemes (β =0.04; error=0.02; lower 95% credible interval=-0.00; upper 95% credible interval=0.08); (ii) low child-directed speech frequency (β =-0.02; error=0.01; lower 95% credible interval=-0.04; upper 95% credible interval=0.00); and (iii) high phonological Levenshtein distance (PLD20), i.e. low phonological neighbourhood density (β =0.18; error=0.02; lower 95% credible interval=0.13; upper 95% credible interval=0.22).

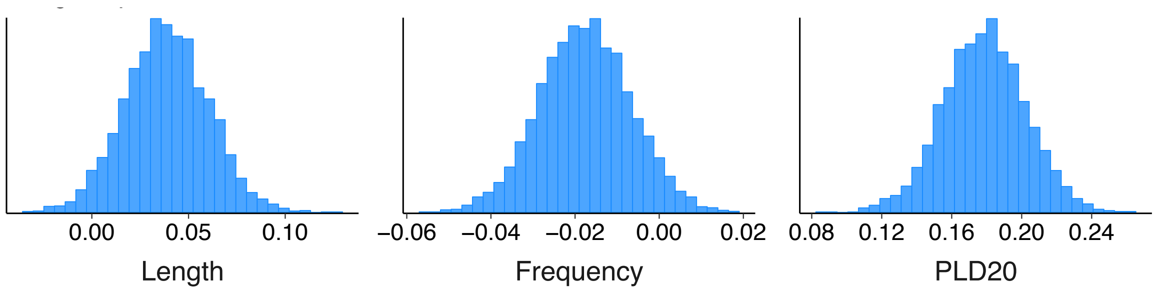

Figure 8.2: Posterior probability distributions for the beta (β) coefficients representing the association between autoencoder mean squared error and; (i) word length (in phonemes), (ii) log child-directed speech frequency, and (iii) phonological Levenshtein distance (PLD20).

We also found evidence of a higher-order interaction between PLD20 and word frequency (β =-0.04; error=0.01; lower 95% credible interval=-0.07; upper 95% credible interval=-0.02). This indicates that the association between high neighbourhood density and low error rate was particularly strong for low frequency words, with high frequency nullifying the PLD20 effect. No higher-order interaction was observed between word length and PLD20 (β =-0.01; error=0.01; lower 95% credible interval=-0.02; upper 95% credible interval=0.01).

During the subsequent validation phase, we used the error rates from our simulated MCDI-WS administration to predict proportions of MCDI-WS word production among 2292 American English-speaking children matched in age to the training inventory (i.e. 18-24 months). We found a negative trend, with words with higher autoencoder error rates produced by a smaller proportion of children (β =-0.03; error=0.03; lower 95% credible interval=-0.09; upper 95% credible interval=0.02).

Finally, during the generalisation phase, we exposed the trained autoencoder to a randomly sampled inventory of 500 previously unseen words that varied in phonological similarity to words in the training inventory. Higher error rates were observed for high-PLD20 (i.e. low-density) words when controlling for the effect of word length (β =0.02; error=0.00; lower 95% credible interval=0.01; upper 95% credible interval=0.02).

8.5 Discussion

This study used an autoencoder neural network to simulate phonological neighbourhood density and distinctiveness effects observed in early word learning. One contribution of this study was to determine how the results of Vitevitch and Storkel (2013) scaled when using sizeable naturalistic training and test data, avoiding data dichotomisation, and incorporating validation against real world data. We trained a three-layer autoencoder using a large corpus of child-directed speech before simulating a communicative development inventory administration at test and then comparing network performance to that of children who were age-matched to the training data (i.e. 18-24 months). Lower reconstruction error rates were observed for words that sounded similar to many other words in the child-directed speech on which the autoencoder was trained. This effect was separable from the effects of word frequency and word length, which also tended in the expected directions given the existing behavioral data. That is, lower error rates were observed for high frequency words and for short words (Braginsky, Yurovsky, Marchman, & Frank, 2019). Despite the extreme simplicity of our network, we were therefore able to simulate the high phonological neighbourhood density configuration advantage reported behaviorally (e.g. Fourtassi et al., 2018; Hollich, Jusczyk, & Luce, 2002; Stokes, 2014; Storkel, 2004). We also reported a higher-order interaction between word frequency and phonological distance. As demonstrated behaviorally by Hollich et al. (2002) and Storkel (2004), we found that high frequency nullified the high phonological neighbourhood density advantage, with amplified error rates for low-frequency, low-density words.

In network validation, we used test-phase error to predict word production rates among 2292 children. Despite a credible interval including zero – indicating that zero may be the true value of the effect – we observed a negative trend in which fewer children produced words that the autoencoder had difficulty representing and reconstructing at test (β =-0.03). Finally, we examined the network's ability to generalise to previously unseen data and found an advantage for words with low PLD20 (i.e. high density) relative to the training corpus. That is, the autoencoder was better able to represent and reconstruct novel words that sounded similar to trained words than novel, phonologically anomalous words. Broadly similar results have been reported behaviorally by Schwartz and colleagues, who found that children were more likely to learn to successfully produce a novel word if that word contained IN-sounds – i.e. sounds that the child had previously produced – than if it contained previously unattested OUT-sounds (Schwartz & Leonard, 1982; Schwartz, Leonard, Frome Loeb, Swanson, & Loeb, 1987; see also Storkel, 2006).

High neighbourhood density is associated with low network error because the encodings of phonologically similar words exhibit similar patterns (i.e. comparable series of 0s and 1s; see the *cat* and *hat* example in *Training*) that are better represented across the network during dimensionality reduction, a process sometimes termed a *conspiracy effect* in machine learning research (Rumelhart, McClelland, and the PDP Research Group, 1986). This makes it possible to reconstruct high phonological neighbourhood density words more accurately, as reflected in low error rates during training, test and generalisation. For instance, exposure to the words *coat*, *pole*, *cone*, *hole, code*, and *mole* prompts changes in the connection weights that support the reconstruction of the novel neighbour *coal*. As the autoencoder is forced through the hidden layer bottleneck (see Figure 8.1) to extract dominant input properties, generalisation to a novel word exhibiting features orthogonal to those previously experienced is inhibited, as reflected by high reconstruction error rates for phonologically distinctive, high PLD20 words.

In our view, a real world parallel to the computational mechanism described above is the cognitive process of long-term auditory priming (e.g. Church & Fisher, 1998). In this account, representations of direct and indirect spoken word exposures are stored in long-term memory (Port, 2007). These representations are initially perceptual rather than conceptual in nature and may be formed implicitly in the absence of semantic information, much like the representations formed by our network. Children are sensitive to the degree of similarity between stored perceptual representations and are able to use this sensitivity to identify (e.g. in the head-turn preference procedure) word sounds that occur at high-probability in their native language (Fourtassi et al., 2018; Jusczyk, Luce, & Charles-Luce, 1994). Novel high-density target words comprising phonological features consistent with existing perceptual memory traces may be held in memory more easily during initial processing (Gathercole, Frankish, Pickering, & Peaker, 1999; Hoover et al., 2010), and this supports the formation of long-term, perceptual and conceptual memory traces that are well detailed and robust to forgetting (Metsala & Walley, 1998; Sosa & Stoel-Gammon, 2012; Storkel, 2004; Walley, Metsala, & Garlock, 2003). Learners may increasingly use their awareness of high-probability word sounds, as well as their related aptitude in producing such sounds, to generalise readily to novel though phonologically familiar words, as in the aforementioned IN-sound/OUT-sound studies of Schwartz and colleagues (Schwartz & Leonard, 1982; Schwartz et al., 1987; see also Storkel, 2006). Low-density words are in general difficult for young children to acquire because there exist few similar stored word representations – whether perceptual or conceptual – from which to generalise.

In the introduction we noted a tendency in the prior literature to treat density and distinctiveness effects separately, and to frame evidence of either a high-density or high- distinctiveness learning advantage as evidence against the opposing effect (e.g. Storkel, 2004; Swingley & Aslin, 2007; Hoover et al., 2010; McKean et al., 2014; Vitevitch & Storkel, 2013). In contrast to this approach, the second contribution of this study is to provide a unified framework for understanding density and distinctiveness effects in early word learning. To do this, we want to emphasise that autoencoder neural networks perform both feature extraction and anomaly detection in parallel. In this sense, it would be inaccurate to suggest that high autoencoder error rates for low-density words provide an analogy to learning deficits in children

(Vitevitch & Storkel, 2013). Whereas low network error rates may indeed be understood as exposure to high-density words prompting a conspiracy effect supporting lexical configuration, high autoencoder error signals the detection of an anomalous target word comprising phonological features inconsistent with those previously trained. This latter effect – i.e. computational anomaly detection – parallels the triggering advantage observed for low-density words in children (e.g. Stager & Werker, 1997; Swingley & Aslin, 2007), which itself may be decomposed into attention- or curiosity-based learning advantages (Twomey & Westermann, 2017; we note that additional learning mechanisms conceivably dependent on the fundamental triggering mechanism simulated form no part of our model). Autoencoder neural networks therefore provide a neat computational analogy to both the density advantages and the distinctiveness advantages observed in behavioral studies of early word learning. Triggering effects may be seen as the advantageous by-product of long-term auditory priming (or a conspiracy effect), which itself supports lexical configuration. These effects can be simulated in parallel within a single autoencoder employing common algorithms and parameter values. In this way, autoencoder simulation illustrates how apparently contradictory density and distinctiveness advantages emerge from a common cognitive mechanism.

The current study demonstrates neighbourhood density and distinctiveness effects similar to those observed in young children in the absence of semantic and pragmatic information. This illustrates the crucial role that raw auditory word similarity plays in the formation of the early lexicon. It is important to emphasise, however, that high phonological neighbourhood density is just one of many factors supporting early word learning, including high exposure frequency, high concreteness, high relevance to babies and infants, and alternative sound variables including phonotactic probability, i.e. the probability of phoneme co-occurrence (Braginsky et al., 2019; Jones & Brandt, 2019a; see *Limitations*, for discussion of phonotactic probability). The current study, for instance, accorded with prior behavioral work in reporting that the high neighbourhood density effect was nullified by high exposure frequency (e.g. Hollich et al., 2002; Storkel, 2004); a finding that suggests an apparent primacy of word-level frequency effects relative to word sound characteristics. It is therefore expected that if a child hears a target word frequently enough, or if that target word is, for instance, highly concrete or highly relevant to the child, then the implicit generalisation preference for words with familiar phonological properties may be nullified.

8.5.1 Limitations

Computational cognitive modelling requires researchers to make numerous decisions, from the overall model type used (e.g. a neural network or Bayesian network) to fine-grained details regarding parameters (e.g. priors, network learning algorithm and learning rate, number of training epochs, etc.). Inevitably, then, some readers may question particular choices we made. One particular point of concern may be our decision to use an autoencoder rather than a recurrent neural network or long short-term memory network, given that recurrent architectures are so commonly used in natural language processing research. The rationale for our choice of architecture was twofold. First, an autoencoder was used in the work by Vitevitch and Storkel (2013) that inspired this study, and replication with naturalistic data necessitated the use of the same architecture. Second, autoencoders are a somewhat distinctive branch of architecture in the sense of performing parallel feature extraction and anomaly detection. This choice of architecture was therefore essential to our aim of illustrating how apparently contradictory behavioural evidence of both density and distinctiveness advantages can be explained in terms of a common mechanism. We have made all of our data and code fully available online, and researchers are welcome to access this material to test alternative configurations of network or stimulus encoding approaches.

Another potential limitation of this report is the exclusion of alternative predictor variables, perhaps most importantly phonotactic probability. High positive correlation between neighbourhood density and phonotactic probability may cause multicolinearity (Storkel, 2004; Storkel & Lee, 2011), which distorts results by changing the magnitude or the direction of estimates, or by inflating estimate errors. While it is possible to tease apart the effects of neighbourhood density and phonotactic probability in controlled experimental settings (e.g. Storkel & Lee, 2011), this is usually not possible when working with naturalistic data or communicative development inventory data (see Storkel, 2004, with respect to MacArthur-Bates data). In this case, the safest way to address multicolinearity risk is to exclude the variable of least interest from the regression model. For us, given our central interest in neighbourhood density effects, this meant omitting phonotactic probability. However, as one anonymous reviewer commented, this makes it impossible to determine the potential contribution of phonotactic probability to the results presented. We would like to re-emphasise that all our code and data can be accessed via the project repository accompanying this paper, and that researchers with a primary interest in sub-lexical phonotactic probability effects rather than the word-level neighbourhood density effects covered in this study are welcome to modify these materials.

8.6 Conclusion

High phonological neighbourhood density has been associated with both advantages (Storkel, 2004) and disadvantages (Swingley & Aslin, 2007) in behavioral studies of early word learning. We explored these effects using an autoencoder neural network in conjunction with corpus and communicative development inventory data. We suggested that the widely reported high-density advantage is explicable in terms of exposure to a phonological neighbourhood prompting a natural conspiracy effect; a process termed long-term auditory priming in the behavioural literature (e.g. Church & Fisher, 1998). We then noted that high phonological distinctiveness supports word learning by reducing the risk of mis-processing novel words as known words in competitive learning environments. Autoencoder modelling encourages us to think of these apparently contradictory effects as emerging from a common mechanism.

8.7 References

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

Braginsky, M., Yurovsky, D., Frank, M., & Kellier, D. (2018). wordbankr: Tools for connecting to wordbank, an open repository for developmental vocabulary data. Retrieved from https://github.com/langcog/wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind, 3,* 52–67. https://doi.org/10.31234/osf.io/cg6ah

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411. doi.org/10.32614/RJ-2018-017

Church, B. A., & Fisher, C. (1998). Long-term auditory word priming in preschoolers: Implicit memory support for language acquisition. *Journal of Memory and Language, 39*(4), 523–542. https://doi.org/10.1006/jmla.1998.2601

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

Fourtassi, A., Bian, Y., & Frank, M. C. (2018). *Word learning as network growth: A cross-linguistic analysis*. Unpublished manuscript, Language and Cognition Lab, Stanford University, Stanford, California. Retrieved from http://langcog.stanford.edu/papers_new/fourtassi-2018-cogsci.pdf

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(03), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(1), 84–95. https://doi.org/10.1037/0278-7393.25.1.84

Grimm, R., & Tulkens, S. (2015). PyPatPho: A phonological pattern generator. GitHub repository. Retrieved from https://github.com/RobGrimm/PyPatPho

Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin, A., Mazuka, R., & Dupoux, E. (2018). Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science, 42*(5), 1586–1617. https://doi.org/10.1111/cogs.12616

H2O.ai. (2016). R Interface for H2O, R package. GitHub repository.

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In B, Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 314–323). Boston, MA: Cascadilla Press.

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Jones, S. D. &, Brandt, S. (2019a). Do children really acquire dense neighbourhoods? *Journal of Child Language, 46*(6), 1260–1273. https://doi.org/10.1017/S0305000919000473

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

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195. https://doi.org/10.1016/j.cognition.2016.04.003

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 408–415. https://doi.org/10.3758/BF03195469

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

McKean, C., Letts, C., & Howard, D. (2014). Triggering word learning in children with Language Impairment: The effect of phonotactic probability and neighbourhood density. *Journal of Child Language*, *41*(6), 1224–1248. https://doi.org/10.1017/S0305000913000445

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs. The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs (3rd ed.).* (3rd ed.). New York: Psychology Press (Taylor and Francis group).

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology 25*(2), 143–170. https://doi.org/10.1016/j.newideapsych.2007.02.001

Python Software Foundation. (2013). Python language reference. *Python Software Foundation*. https://doi.org/https://www.python.org/

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from: http://www.R-project.org/

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986). Parallel distributed Processing: Explorations in the Microstructure of Cognition (Volume 1: Foundations). Cambridge, MA: MIT Press.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(2), 1928–1941. Retrieved from https://psyarxiv.com/93mwx

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition. *Journal of Child Language, 9*(2), 319–336. https://doi.org/10.1017/S0305000900004748

Schwartz, R. G., Leonard, L. B., Frome Loeb, D., Swanson, L. A., & Loeb, D. M. (1987). Attempted sounds are sometimes not: An expanded view of phonological selection and avoidance. *Journal of Child Language, 14(3), 411*–418. https://doi.org/10.1017/S0305000900010205

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596. https://doi.org/10.1044/1092-4388(2011/10-0113)

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102

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

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(2), 201–221. https://doi.org/10.1017/S0142716404001109

Storkel, H. L. (2006). Do children still pick and choose? The relationship between phonological knowledge and lexical acquisition beyond 50 words. *Clinical Linguistics and Phonetics*, *20*(7–8), 523–529. https://doi.org/10.1080/02699200500266349

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review, 18*(3), 605–11. https://doi.org/10.3758/s13423-011-0078-9

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132. https://doi.org/10.1016/j.cogpsych.2006.05.001

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods, 51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Twomey, K. E., & Westermann, G. (2017). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, *21*(4), e12629. https://doi.org/10.1111/desc.12629

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the acquisition of phonological word forms with computational experiments. *Language and Speech*, *56*(4), 493–527. https://doi.org/10.1177/0023830912460513

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing: An Interdisciplinary Journal*, *16*(1), 5–20. https://doi.org/10.1023/A:1021789804977

Chapter 9 Summary and Conclusions

I began this thesis by describing the following contradiction. Memory advantages for distinctive stimuli are well established (Hunt & Worthen, 2006). You are, for instance, more likely to form an accurate memory of the standout object in an array or the standout word in a list. Despite this, young children learn high-density words more readily than they learn phonologically distinctive, low-density words (e.g. Storkel, 2004). This high-density bias is the defining characteristic of the emerging auditory lexicon, which itself underpins language comprehension and production, and provides the foundation of grammatical development and literacy (Claessen & Leitão, 2012; Goodman & Bates, 1997). The question is, then:

Why do young children learn dense rather than distinctive words?

This question guided the five empirical studies presented in this thesis, as well as the review that preceded them. As an integrated whole this thesis provides the most comprehensive account to date of the high-density bias that characterises early auditory word learning. Evidence of a high-density bias of course precedes the empirical studies included in this thesis. However, by adopting pre-registration, large-samples, open data and code, and modelling principles such as the avoidance of unwarranted predictor and group dichotomisation, I believe that this thesis has contributed substantially to putting such evidence on a firmer foundation. The development of an explanatory account of the high-density bias grounded in the principle of analogous generalisation and spelled out by recourse to pre-existing implicit, short-term, and long-term memory research is also a major contribution. The purpose of this final chapter is to summarise the major contributions of this work, and to outline directions for future research.

9.1 Summary of empirical findings

Chapter four, *Auditory lexical decisions in developmental language disorder*, presented a meta-analysis of studies using the auditory lexical decision task to measure the quality of word sound representations in children with and without developmental language disorder (DLD). The auditory lexical decision task was selected for this study because it minimises or even removes the requirement of a verbal response (e.g. requiring a simple yes/no response or button press), meaning that performance deficits cannot be attributed to retrieval issues or motor delay. Electronic database searches and emails sent out to researchers in the field initially identified 2372 studies, with this reduced to nine studies as a result of duplicate removal and the application of stringent eligibility and quality criteria. The final collection of studies included 499 children aged between 3;8 and 11;4. Analysis indicated that children with DLD were significantly less accurate in the auditory lexical decision task than age-matched controls, but that there was no substantial difference between these groups in terms of response time. There was also no reliable difference between children with DLD and language-matched controls in either accuracy or response time. This pattern of results is in line with the general view that the linguistic profiles of children with DLD are delayed though not deviant. The results of this study broadly support the hypothesis that some children with DLD have difficulty forming detailed lexical representations relative to age- though not language-matched peers. However, further work is required to determine the performance profiles of potential subgroups, as well as the impact of manipulating different lexical characteristics, such as the position and degree of non-word error, phonotactic probability, and semantic network size. Due to the small cohort size (i.e. nine studies), I was unfortunately unable to include such factors in moderator analyses. This first study set the stage for those that followed by identifying group differences in the ability to represent word sounds.

In the subsequent studies of this thesis, the quality of word sound representations was inferred from comprehension and production data (e.g. parental report data, accuracy and variability rates), and there was a specific focus on the effect of phonological neighbourhood density. Chapter five, *Do children really acquire* *dense neighbourhoods?*, used communicative development inventory data from 300 British English-speaking children aged between 12 and 25 months. In this study, Bayesian regression was used to model word understanding and production as a function of: (i) phonological neighbourhood density, (ii) frequency, and (iii) length, as well as adult ratings of; (iv) babiness, (v) concreteness, (vi) valence, (vii) arousal, and (viii) dominance. Results showed a separable, positive association between phonological neighbourhood density and word production, particularly among younger children, though no reliable association between phonological neighbourhood density and spoken word comprehension. That is, young children were more likely to produce words that sounded like many other words, but they could apparently understand words with relatively uncommon phonology as long as they were, for instance, concrete, frequent, and highly relevant to their lives (e.g. *pushchair*). It was argued that cognitive demand may be low during the initial processing of spoken words comprising commonly occurring sounds (i.e. high-density words), and that this may support the formation of detailed long-term phonological word memories that provide motor plans facilitating accurate word production. Comprehension, in contrast, remains possible even when the corresponding word sound memory lacks the detail required for accurate production (Bishop, 2014). The observed age-related decline in the importance of high phonological neighbourhood density as a predictor of word production indicates that the ability to remember words comprising uncommon phonological sequences improves across early development.

Chapter six, *Neighbourhood density and word production in delayed and advanced learners*, then pursued this line of inquiry further by looking at the association between phonological neighbourhood density and word production in 442 18-month old children with expressive lexicon sizes between zero and 517 words. The emphasis here, then, was on individual differences in the importance of phonological neighbourhood density as a predictor of word production (i.e. variance associated with vocabulary size), in contrast to the age-related variance examined in chapter five. In particular, this study aimed to re-examine the claim that a difficulty forming memories of low phonological neighbourhood density words may be a central determinant of delayed expressive vocabulary growth (e.g. Stokes, 2014). To do this, I fitted a

Bayesian regression model in which the production of each communicative inventory word by each child was predicted by interactions between that child's expressive lexicon size and the word's (i) phonological neighbourhood density, (ii) frequency in child-directed speech, and (iii) length, as well as adult ratings of; (iv) babiness, and (v) concreteness. I found that children with larger expressive lexicons were more likely to produce words containing uncommon sound sequences than age-matched children with smaller lexicons. This indicates that the decline in the importance of high phonological neighbourhood density as a predictor of word production reported in chapter five could be a function of vocabulary size rather than of age per se. The magnitude of the interaction between expressive lexicon size and phonological neighbourhood density remained modest, however, relative to interactions between expressive lexicon size and word frequency, length, and adult ratings of babiness and concreteness. This makes it impossible to single out the acquisition of low-density words as a specific problem area for language-delayed children – including late talkers – on the basis of this or similar data (cf. Stokes, 2010, 2014). On the basis of such correlational data alone it would, for instance, be equally valid to argue that late talking is the result of a difficulty learning words with low relevance to the child's life (i.e. words with low babiness ratings). This prompted the conclusion that prior emphasis on a difficulty with the memorisation of low neighbourhood density words as a determinant of slow vocabulary growth may be unwarranted, and that the current evidence base in this direction is not robust enough to strongly support the development of possible interventions for late talkers (cf. Stokes, 2014).

Chapter seven, *Accuracy and variability in early spontaneous word production*, then went beyond the binary 'produces' / 'does not produce' outcomes analysed in chapters five and six to assess the effects of age, frequency, and neighbourhood density on rates of spoken word accuracy and stability. These effects were studied across 234,551 spontaneous word productions from five typically developing children (0:11-4;0) in the Providence corpus. This was the first study I am aware of to look at these effects in large-scale naturalistic data without restriction to a particular phoneme cluster, word class, or utterance count. In keeping with the results of chapters five and six, Bayesian regression indicated positive associations between age, input frequency, and phonological neighbourhood density, and spoken word accuracy and stability rates. Accordingly, an explanatory account of findings emphasising the quality of phonological word memories was again presented.

Finally, in chapter eight, *Density and distinctiveness in early word learning*, I presented a computational simulation of the neighbourhood effects reported in chapters five, six, and seven. I trained a vanilla autoencoder neural network on a large corpus of child-directed speech and simulated a communicative development inventory administration to test the accuracy of the word sound representations that the network had formed. I then validated the results of this simulation using communicative development inventory data from over two thousand children. Like the children recorded in the validation data, and similarly those assessed in chapters five, six, and seven, the network represented high-density words more accurately than it represented low-density words, and this bias was separable from the effects of exposure frequency and word length. In an additional generalisation phase simulation, the network was also shown to represent novel words that sounded like previously trained words more accurately than it represented novel phonologically unfamiliar words. I presented an account of network performance accommodating conflicting evidence of distinctiveness advantages (e.g. Swingley & Aslin, 2007). The high-density bias was interpreted in terms of a conspiracy effect, which supported generalisation to novel words with sound features similar to those dominant in the training data. It was then argued that distinctiveness advantages emerge as a by-product of this fundamental process, with a spike in error rate observed for words containing sound features orthogonal to those previously experienced. An analogy was made between computational anomaly detection and the trigger stage learning advantage reported for low-density words among young children (e.g. Storkel & Lee, 2011). This study showed, then, that the apparently contradictory density and distinctiveness advantages reported in behavioral studies of early word learning can emerge within a single computational architecture.

9.2 Summary of methodological contributions

9.2.1 Identifying limitations of existing applications of the auditory lexical decision task and related paradigms

The primary aim of chapter four, *Auditory lexical decisions in developmental language disorder*, was to produce aggregated data summaries that could provide a useful benchmark for future analyses. However, when reviewing the studies included in this analysis, I also identified a number of issues with existing applications of the auditory lexical decision task that may similarly affect related paradigms including mispronunciation identification and gating. For instance, some studies included no formal assessment of whether the words used at test were known to the participant (e.g. Windsor & Hwang, 1999). Additionally, while cross-references were often made between studies, the experimental designs used across studies showed considerable variation in both the response required by children (e.g. verbal or non-verbal) and the method of stimulus presentation (e.g. pre-recorded or spoken live by an experimenter; see Appendix A.1). It was noted that although such factors may appear trivial, they can in fact introduce systematic bias. Maillart et al. (2004), for instance, describe a study in which children with DLD responded differently to stimuli presented via computer and stimuli spoken live by an experimenter, perhaps due to adopting a lip-reading strategy during target word discrimination. A review of the methods used to measure the quality of word sound memories is required given such disparities, and given the importance of these paradigms in translational research relating to the identification and assessment of children with DLD. As described below (section 9.4.2), determining the validity and reliability of the various measures held to converge on the quality of lexical representations will form a key part of this process.

9.2.2 Encouraging a shift away from unwarranted data dichotomisation in early language research

The decision to use a particular method of data preparation or statistical analysis is ultimately research question dependent, and any number of methods may be adequately justified with respect to a given study. However, there are two approaches involving data dichotomisation commonly adopted in language development research that I believe to be often unwarranted, and which I hope the empirical studies of this thesis have helped to discourage.

The first is the application of arbitrary cut-offs to standardised language test scores in order to distinguish experimental and control groups for use as predictors in analyses such as t-tests and ANOVAs. As written in chapter six, this approach not only reduces statistical power but also appears unjustifiable with respect to the study of early developmental delay (e.g. late talking, possible DLD) given that the majority of children with early language delay do not show later language difficulties (Hammer et al., 2017; Rowe & Leech, 2017). It is not the case, therefore, that children with and without early language difficulties comprise qualitatively different groups. Instead there is a continuous range of ability that modelling should in many cases aim to capture, for instance by using standardised language test scores as a continuous (rather than categorical) predictor in a regression model.

The second common area of unwarranted data dichotomisation involves the neighbourhood density variable central to this thesis. Many studies into neighbourhood effects in early language development – including two studies of this thesis – adopt a plus-minus-one-phoneme criterion of neighbourhood density, under which, for instance, *cat* neighbours *hat*, *cot, can,* and *catch* (e.g. Jones & Brandt, 2019a, 2019b; Stokes, 2014; Storkel, 2004). However, as reviewed at length in section 2.2.2, this approach involves throwing out a great deal of information regarding degrees of phonological distance. In section 2.2.2.1, for instance, I reported results from a preliminary analysis showing that approximately half (48.31%) of words within a representative child lexicon may be attributed zero neighbourhood density under the plus-minus-one phoneme approach. To make this point clear, consider that both *bag* and *supercalifragilisticexpialidocious* are non-neighbours of *cat* under a categorical, plus-minus-one-phoneme criterion. The solution to this limitation is to adopt a continuous measure of word-level phonological similarity such as PLD20, defined as the mean number of additions, substitutions, or eliminations of phonemes required to change a given word into its nearest twenty phonological neighbours (Suárez et al., 2011, p. 606). This is the approach taken in chapter seven of this thesis and currently the metric I consider most appropriate to the study of neighbourhood density effects. It is worth noting, however, that the use of the categorical, plus-minus-one-phoneme criterion in chapters five and six is unlikely to compromise the results of these studies. The plus-minus-one phoneme criterion and PLD20 are highly correlated, and have been shown to confer analogous effects (Suárez et al., 2011). However, the value of continuous measures of phonological word similarity remains that they avoid the issue of data loss and the 'lexical hermit' problem.

9.2.3 Highlighting the fragility of hard-and-fast spoken word accuracy and variability thresholds

A major contribution of chapter seven was to illustrate the extent to which accuracy and variability rates fluctuate given apparently minor changes in the underlying categorisation criteria used (see section 7.3.3). This factor may contribute to the current lack of consensus regarding the accuracy and variability rates expected in the typically developing population (e.g. Holm, et al., 2007; Sosa, 2015). Without such consensus it is somewhat difficult to interpret data from accuracy and variability assessments of children suspected of having language disorder. On the basis of the data modelled in chapter seven, it was impossible to select between the taxonomies presented (i.e. Table 7.3 and Table 7.4). Each underlying categorisation criterion was well justified, and each generated a taxonomy with proportions of accuracy and variability broadly continuous with those previously reported (e.g. Sosa, 2015; Holm et al., 2007). Furthermore, this element of the study highlighted the trade-off inherent in setting hard-and-fast decision boundaries (section 7.3.3). Setting strict criteria meant some apparently acceptable word productions were classified as erroneous (e.g. the target /æl**ə**geɪtəɹ/ pronounced /æl**ɪ**geɪtəɹ/). In contrast, loosening the error and variability thresholds permitted the classification of some clear errors as accurate forms, for instance the production of /bæg/ as /bæk/, the production of /bæθ/ as /bæ/, and the production of /bæt/ as /bɛt/. It was noted that this may be particularly damaging with respect to short words, which dominate the early productive lexicon. (The mode word length among children in the Providence corpus is three phonemes.) For such words one discrepant phoneme may represent a substantial error, which was ignored under the minimally relaxed thresholds. Further research that aims to address this issue is vital given the potential importance of baseline accuracy and variability data to translational research and clinical practice. In section 9.4.3, I re-iterate my proposal of an accuracy judgement task to break the apparent deadlock in the literature on early spoken word accuracy and variability rates.

9.3 Summary of theoretical contributions

9.3.1 Identifying the high-density word learning bias as a separable production effect associated with vocabulary size

Chapter five, *Do children really acquire dense neighbourhoods?*, was the first empirical study that I am aware of to show that the neighbourhood density bias is separable from the effects of a large range of alternative predictors commonly linked to variance in age of acquisition, such as frequency, concreteness, and relevance to infants. A similar approach was also taken in chapter six, *Neighbourhood density and word production in delayed and advanced learners*. Together, these analyses constitute an important contribution to the literature because one plausible answer to the question guiding the studies of this thesis – that is, *why do young children learn dense rather than distinctive words?* – is that high-density words also happen to be, for instance, highly concrete or highly relevant to the child's life, and that it is in fact these factors that underpin the high-density bias. In other words, the high-density bias may be an epiphenomenon. The major contribution of chapters five and six was to rule out this possibility by demonstrating that the high-density bias holds over and above a large range of alternative explanatory factors.

Chapter five, *Do children really acquire dense neighbourhoods?*, was also novel in using Bayesian multivariate regression to separate the effects of neighbourhood density on word comprehension and on word production. In doing so, this study identified a strong association between high neighbourhood density and early word production. On this basis, it was argued that young children are able to understand words that are, for instance, highly frequent, highly concrete, or of high relevance to their lives, but that accurate word production also depends on a high-quality word sound representation, and this is more likely to be available for high-density words early in development. This is an important contribution because the high neighbourhood density bias is often framed as a learning advantage, broadly defined. However, these results suggest instead that it is more accurate to talk about an early high-density word *production* advantage. Chapter six, *Neighbourhood density and word production in delayed and advanced learners*, then built on these findings by showing that the age-related decline in the importance of high phonological neighbourhood density as a predictor of word production in fact emerges as a function of vocabulary size. In combination, then, the contribution of chapters five and six has been to situate the high-density bias as a separable production effect, characteristic among young children and children with limited expressive vocabularies.

9.3.2 Detailing an explanatory account of the high-density word production bias in terms of analogous generalisation

Chapter eight, *Density and distinctiveness in early word learning*, contributed significantly to the development of the broader theoretical account of the high-density production bias detailed in chapter two of this thesis and drawn on throughout. Building this account involved trying to spell out the fundamental processes underpinning the high-density bias in the autoencoder neural network used in this chapter in terms of analogous cognitive processes identified in the developmental literature. The account I developed has three essential elements. First, direct and indirect spoken word exposure results in the formation of perceptual and conceptual memory traces that represent the sound structure of the ambient language in the mind of the child (Jusczyk et al., 1994). In computational terms this is the conspiracy effect; the fine-tuning of connection weights to accommodate dominant sound patterns in the language to which the network is exposed. Second, short-term memorisation advantages are observed for words with sound features consistent with those represented perceptually and conceptually in long-term memory. In computational terms this means that connection weights are required to update less on exposure to a novel stimulus with familiar auditory features. Third, the short-term memory trace is passed to long-term memory in high detail, supporting accurate processing (e.g. mispronunciation identification) and production for this high-density item (e.g. Gathercole et al., 1999). Computationally, in chapter eight, this is analogous to the reconstruction error being lower for high-density stimuli. In addition to addressing a series of open questions related to the high-density bias, then, this thesis went beyond the contribution of each individual study to offer a novel answer to the question: *Why do young children learn dense rather than distinctive words?* Namely: *Because the auditory lexicon is built through analogous generalisation*. The explicit and implicit memorisation of direct and indirect exposure to the neighbourhood *catch*, *hat*, *mat*, *can*, *sat*, *match*, and *bat*, for instance, supports analogous generalisation to the word *cat* by way of short- and associated long-term memory advantages. To re-iterate the qualifying points made at the beginning of this thesis, my claim is not that analogous generalisation is the only factor driving growth of the auditory lexicon. The empirical studies of this thesis make this clear. For instance, in chapter five I report that children are also more likely to produce words that are concrete and relevant to their lives (e.g. *ball*, *pushchair*) than they are to produce abstract low-relevance words (e.g. *how*, *later*). Instead, my claim is that, all else being broadly equal, a production bias for high-density words will be observed because these words are easier to generalise to from existing explicit and implicit long-term word sound memories. I also recognise that each component of the account outlined here (i.e. implicit, short-term, and long-term memory processes) has previously been associated with neighbourhood effects in early word learning. However, this thesis is to my knowledge the first attempt to unify these existing fragmentary accounts of the high-density production bias under the label of analogous generalisation.

9.4 Directions for future research

9.4.1 Determining the significance of exemplar and prototype frameworks in translational research and clinical practice

Exemplar theory has a long history in cognitive science, but has only recently begun to gain traction in the study of child language, where prototype theories remain dominant (Ambridge, 2019). As noted throughout this thesis, there is a large literature purporting to test the quality of 'underlying representations', where underlying or similar terminology denotes being abstracted from the variation inherent in natural speech. However, evidence of speaker effects and the difficulty of specifying the form of the underlying abstract representation appear to make this position untenable (see chapter two). Much of the research on auditory word representation quality, like chapter four of the current thesis, relates to theories of early language impairment and clinical practice, for instance the identification of disorders and the development of programs of intervention (see Claessen & Leitão, 2012, for review). It will be interesting to see how this area of research responds to growing interest in exemplar-based theories of early auditory word representation. The fundamental question for translational researchers and clinicians is whether conceiving of word sound representations as exemplars rather than as abstract prototypes may inform approaches to experimental design and the interpretation of results, and also to the identification and treatment of language disorder.

9.4.2 An assessment of the accuracy and reliability of tasks measuring word sound representation quality

The studies of this thesis were linked by an interest in factors affecting the quality of word sound representations, whether child-based factors such as age or clinical profile, or lexical factors such as neighbourhood density. Many of the studies of this thesis follow a long trend, starting perhaps with Ferguson and Farwell (1975), in assuming that the quality of word sound representations can be inferred from comprehension and production data. For instance, if a child without pronounced motor impairment is able to comprehend but not accurately and stably produce a word, it is reasonable to assume that the mental representation of that word's sound lacks detail. As described in chapter four, a number of experimental paradigms have been developed to assess the quality of underlying word sound representations. This includes gating, naming, non-word repetition, mispronunciation identification, and the lexical decision task. Results from these tasks are, however, sometimes mixed. For instance, in the gating task auditory words are cut and presented in chunks of increasing length – for instance [e], [ele], [elepha], [elephant] – with the aim being to identify the target word as quickly as possible. In one of the earliest studies to adopt this paradigm, Dollaghan (1998) reports that children with developmental language disorder require considerably longer gates to recognise newly taught though not familiar words than their age-matched peers. Subsequent studies, however, have failed to replicate this effect (Mainela-Arnold, Evans, & Coady, 2008; Montgomery, 1999). Such discrepancies make the experimental literature on word sound representation quality reminiscent of that on procedural learning, recently critiqued by West, Vadillo, Shanks, and Hulme (2017). These authors identified a prevalence of small sample sizes and extreme group designs that may exaggerate effects (i.e. language disordered versus language typical), as well as low reliability across tasks thought to measure procedural learning. Given that the quality of early word sound representations remains an area of considerable debate which feeds directly into clinical practice, and given that existing results are often inconsistent and based on small-sample, extreme group designs, a large-scale study to determine the reliability of the various tasks held to converge on the quality of word sound representations is much needed.

9.4.3 A comprehensibility judgement task to determine baseline spoken word accuracy and variability rates

Chapter seven, *Accuracy and variability in early spontaneous word production*, highlighted a sharp division in the existing literature on spoken word accuracy and variability rates (e.g. Sosa, 2015; Holm et al., 2007), as well as the difficulty of setting hard-and-fast decision boundaries. These are important issues because baseline accuracy and variability data may prove useful in aiding the identification of language disorder. One way to address these issues may be to conduct a comprehensibility judgement task. This would involve collecting large samples of spontaneous and elicited child speech data, and then having a group of adult listeners make accuracy judgements for word tokens within that data. Listeners would hear isolated spoken word tokens and identify whether they considered each to be accurate or inaccurate, with these judgments then forming the basis for the classification of tokens into accuracy and variability taxonomies. It would also be possible to stratify this data by child age, to provide a detailed assessment of changes in accuracy and variability rates throughout development. As written in chapter seven, such a study may deliver the baseline data needed to break the apparent deadlock and resolve existing disagreement regarding the spoken word accuracy and variability rates that can be expected within the typical range.

9.4.4 Testing the account of analogous generalisation in early auditory word learning

Future behavioural research should test the explanatory account of the high-density production bias presented in this thesis. For instance, it may be informative to expose two groups of children to two artificial-language lexicons; one in which high-density neighbourhoods were present and one in which they were not. It would then be possible to expose children to a novel high-density target word and test their delayed recall or production of that word in order to determine whether the sound structure of the ambient language during training had a causal influence on successful word learning. It would also be possible to incorporate short-term memory measures in order to probe the influence of short-term memory on analogous generalisation. This type of investigation would be valuable because much of the existing data in this area is correlational.

The theoretical account presented would also benefit from further instantiation as a computational model. While the vanilla autoencoder neural network presented in chapter eight significantly helped to develop this account, it would be possible to improve on this work by creating an architecture with distinct short-term and long-term memory modules that could be differentially manipulated alongside the neighbourhood structure of the training data in order to assess the contributions of each module to successful analogous generalisation. Note also that despite my general support for an exemplar-based framework (see chapter two), the account of analogous generalisation developed throughout this thesis makes no direct predictions regarding the underlying nature of word sound representations (i.e. analogy from both prototypes and exemplars is possible), and similarly the artificial neural network presented in chapter eight is not a formal instantiation of either an exemplar- or prototype-based approach. It would also be interesting, therefore, to build features consistent with both exemplar- and prototype-based frameworks into a number of architectures, and then to compare each network's performance on simulated analogous generalisation tasks.

Chapter 10 Bibliography

10.1 Consolidated bibliography

Ainsworth, S., Welbourne, S., & Hesketh, A. (2016). Lexical restructuring in preliterate children: Evidence from novel measures of phonological representation. *Applied Psycholinguistics*, *37*(4), 997–1023. https://doi.org/10.1017/S0142716415000338

Alcock, K. (2006). The development of oral motor control and language. *Down Syndrome Research and Practice*, *11*(1), 1–8. https://doi.org/10.3104/reports.310

Alcock, K. J., Meints, K., & Rowland, C. F. (2017). *UK-CDI Words and Gestures - Preliminary norms and manual*. Retrieved from http://lucid.ac.uk/ukcdi

Almodovar, D. (2014). *Effects of phonological neighborhood density on lexical access in adults and children with and without specific language impairment. City University of New York (CUNY) Academic Works*. The City University of New York. Retrieved from http://academicworks.cuny.edu/gc_etds/160/?utm_source=academicworks.cuny.edu%2Fgc_etds%2F160&utm_medium=PDF&utm_campaign=PDFCoverPages

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

Ambridge, B. & Lieven, E. V. M. (2011). *Child language acquisition: Contrasting theoretical approaches.* Cambridge: Cambridge University Press.

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

Aslin, R., Jusczyk, P. W., & Pisoni, D. (1998). Speech and auditory processing during infancy: Constraints on and precursors to language. In W. Damon (Ed.), *Handbook of childpsychology: Volume 2: Cognition, perception, and language* (pp. 147–198). Hoboken, NJ: John Wiley.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory (Vol. 8)* (pp. 47–89). New York: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–59. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958156

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. http://www.jstatsoft.org/v67/i01/

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5), e19348. https://doi.org/10.1371/journal.pone.0019348

Befi-Lopes, D. M., Pereira, A. C. S., & Bento, A. C. P. (2010). Phonological representation of children with specific language impairment (SLI). *Pró-Fono*, *22*(3), 305–310. https://doi.org/S0104-56872010000300025

Bennetts, S. K., Mensah, F. K., Westrupp, E. M., Hackworth, N. J., & Reilly, S. (2016). The Agreement between Parent-Reported and Directly Measured Child Language and Parenting Behaviors. *Frontiers in Psychology*, *7*, 1710. http://doi.org/10.3389/fpsyg.2016.01710

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of*

Sciences, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Betz, S. K., & Stoel-Gammon, C. (2005). Measuring articulatory error consistency in children with developmental apraxia of speech. *Clinical Linguistics and Phonetics, 19*(1), 53–66. https://doi.org/10.1080/02699200512331325791

Bishop, D V M, Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLOS ONE*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Bishop, D. V. M. (1997). *Uncommon understanding: Development and disorders of language comprehension in children.* Hove, England, UK: Taylor & Francis.

Bishop, D. V. M. (2006). What causes specific language impairment in children? *Current Directions in Psychological Science*, *15*(5), 217–221. https://doi.org/10.1111/j.1467-8721.2006.00439.x

Bishop, D. V. M. (2014). *Uncommon understanding: Development and disorders of language comprehension in children, classic edition.* Hove, England, UK: Psychology Press (imprint of Taylor & Francis).

Bishop, D. V. M., Brown, B. B., & Robson, J. (1990). The relationship between phoneme discrimination, speech production, and language comprehension in cerebral-palsied individuals. *Journal of Speech and Hearing Research*, *33*, 210–219. https://doi.org/10.1044/jshr.3302.210

Bishop, D. V. M., Hardiman, M. J., & Barry, J. G. (2012). Auditory deficit as a consequence rather than endophenotype of specific language impairment: Electrophysiological evidence. *PLoS ONE*, *7*(5), e35851. https://doi.org/10.1371/journal.pone.0035851

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., & Greenhalgh, T. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLOS ONE*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Bishop, D. V.M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*(4), 391–403. https://doi.org/10.1111/j.1469-7610.1996.tb01420.x

Borovsky, A., Burns, E., Elman, J. L., & Evans, J. L. (2013). Lexical activation during sentence comprehension in adolescents with history of specific language impairment. *Journal of Communication Disorders*, *46*, 413–427. https://doi.org/10.1016/j.jcomdis.2013.09.001

Braginsky, M., Yurovsky, D., Frank, M., & Kellier, D. (2018). wordbankr: Tools for connecting to wordbank, an open repository for developmental vocabulary data. Retrieved from https://github.com/langcog/wordbankr

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind, 3,* 52–67. https://doi.org/10.31234/osf.io/cg6ah

Braginsky, M., Yurovsky, D., Marchman, V. A., Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Retrieved from: http://langcog.stanford.edu/papers_new/braginsky-2016-cogsci.pdf

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395-411. doi.org/10.32614/RJ-2018-017

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Čeponienė, R., & Keren-Portnoy, T. (2011). The role of production practice in lexical and phonological development – a commentary on Stoel-Gammon's 'Relationships between lexical and phonological development in young children.' *Journal of Child Language*, *38*(1), 41–45. https://doi.org/10.1017/S0305000910000504

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper-Row.

Church, B. A., & Fisher, C. (1998). Long-term auditory word priming in preschoolers: Implicit memory support for language acquisition. *Journal of Memory and Language, 39*(4), 523–542. https://doi.org/10.1006/jmla.1998.2601

Claessen, M., & Leitão, S. (2012). Phonological representations in children with SLI. *Child Language Teaching and Therapy, 28*(2), 211–223. https://doi.org/10.1177/0265659012436851

Claessen, M., Heath, S., Fletcher, J., Hogben, J., & Leitão, S. (2009). Quality of phonological representations: A window into the lexicon? *International Journal of Language and Communication Disorders*, *44*(2), 121–144. https://doi.org/10.1080/13682820801966317

Coady, J. A., & Evans, J. L. (2008). Uses and interpretations of non‐word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language and Communication Disorders, 43*(1), 1–40. https://doi.org/10.1080/13682820601116485

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Conti‐Ramsden, G., Durkin, K., N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders, 53*(2), 237–255. https://doi.org/10.1111/1460-6984.12338

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development, 61*(5), 1584–1595. http://www.ncbi.nlm.nih.gov/pubmed/2245748

Crosbie, S. L., Howard, D., & Dodd, B. J. (2004). Auditory lexical decisions in children with specific language impairment. *British Journal of Developmental Psychology*, *22*(1), 103–121. https://doi.org/10.1348/026151004772901131

Dale, P., Simonoff, E., Bishop, D., Eley, T., Oliver, B., Price, T., … Plomin, R. (1998). Genetic influence on language delay in two-year-old children. *Nature Neuroscience*, *1*(4), 324–328. https://doi.org/10.1038/1142

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, *143*, 77–86. https://doi.org/10.1016/j.cognition.2015.06.003

DeCasper, A., & Fifer, W. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176. https://doi.org/10.1126/science.7375928

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), Phonetics and phonology in language comprehension and production: Differences and similarities. New York: Mouton de Gruyter. https://doi.org/10.1515/9783110895094.9

Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language*, *36*(01), 173–200. https://doi.org/10.1017/S0305000908008921

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2002). Diagnostic Evaluation of Articulation and Phonology–U.S. Edition (DEAP) Technical Report. San Antonio, TX: Pearson.

Dollaghan, C. (1998). Spoken word recognition in children with and without specific language impairment. *Applied Psycholinguistics*, *19*(2), 193–207. https://doi.org/10.1017/S0142716400010031

Dollaghan, C. A., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136–1146. http://www.ncbi.nlm.nih.gov/pubmed/9771635

Edwards, J., & Lahey, M. (1996). Auditory lexical decision of children with specific language impairment. *Journal of Speech and Hearing Research*, *39*, 1263–1273.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The Interaction Between Vocabulary Size and Phonotactic Probability Effects on Children's Production Accuracy and Fluency in Nonword Repetition. *Journal of Speech, Language, and Hearing Research, 47*(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

Farquharson, K., Centanni, T. M., Franzluebbers, C. E., & Hogan, T. P. (2014). Phonological and lexical influences on phonological awareness in children with specific language impairment and dyslexia. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00838

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language*, *51*, 419–439. https://doi.org/10.1017/CBO9780511980503.007

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Field, A. P. (2013). Meta-analysis in clinical psychology research. In P. C. Kendall & J. S. Comer (Eds.), *The Oxford Handbook of Research Strategies for Clinical Psychology* (pp. 317–335). Oxford: Oxford University Press.

Fisher, C., Church, B. A., & Chambers, K. E. (2004). Learning to Identify Spoken Words. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon* (pp. 3–40). Massachusetts: The MIT Press.

Floccia, C. (2017). Data collected with the Oxford CDI over a course of 5 years in Plymouth Babylab, UK. With the permission of Plunkett, K. and the Oxford CDI from Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language, 27,* 689-705. Retrieved from: http://centaur.reading.ac.uk/4542/1/Hamilton.Plunkett.Schafer.pdf

Fourtassi, A., Bian, Y., & Frank, M. C. (2018). *Word learning as network growth: A cross-linguistic analysis*. Unpublished manuscript, Language and Cognition Lab, Stanford University, Stanford, California. Retrieved from http://langcog.stanford.edu/papers_new/fourtassi-2018-cogsci.pdf

Fox, J. & Weisberg, S. (2011). *An {R} Companion to Applied Regression*, Second Edition. Thousand Oaks California: Sage. http://socserv.socsci.mcmaster.ca/ jfox/Books/Companion

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(03), 677–694. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29*, 336–360. https://doi.org/10.1016/0749-596X(90)90004-J

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 25*(1), 84–95. https://doi.org/10.1037/0278-7393.25.1.84

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, *33*(6), 966–979. https://doi.org/10.1037/0012-1649.33.6.966

Gathercole, S. E., Service, E., Hitch, G. J., Adams, A. M., & Martin, A. J. (1999). Phonological short-term memory and vocabulary development: Further evidence on the nature of the relationship. *Applied Cognitive Psychology, 13*(1), 65–77. https://doi.org/10.1002/(SICI)1099-0720(199902)13:1<65::AID-ACP548>3.0.CO;2-O

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not Itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Ghetti, S., & Lee, J. (2011). Children's episodic memory. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(4), 365–373. https://doi.org/10.1002/wcs.114

Gierut, J. A., & Dale, R. A. (2007). Comparability of lexical corpora: Word frequency in phonological generalization. *Clinical Linguistics and Phonetics*, *21*(6), 423–433. Retrieved from: https://doi.org/10.1080/02699200701299891

Gierut, J. A., & Morrisette, M. L. (2012). Density, frequency and the expressive phonology of children with phonological delay. *Journal of Child Language*, *39*(4), 804–834. https://doi.org/10.1017/S0305000911000304

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and* *meta-analysis (2nd ed.)* (pp. 357–376). New York: Russell Sage Foundation.

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 649–660. https://doi.org/10.1037/0096-1523.25.3.649

Goffman, L., Gerken, L., & Lucchesi, J. (2007). Relations Between Segmental and Motor Variability in Prosodically Complex Nonword Sequences. *Journal of Speech, Language, and Hearing Research*, *50*(2), 444–58. https://doi.org/10.1044/1092-4388(2007/031)

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Massachusetts: MIT Press. Retrieved from https://www.deeplearningbook.org

Goodman, E., & Bates, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515-531.

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the nonword repetition performance of children with and without specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)

Grimm, R., & Tulkens, S. (2015). PyPatPho: A phonological pattern generator. GitHub repository. Retrieved from https://github.com/RobGrimm/PyPatPho

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. https://doi.org/10.3758/BF03204386

Grunwell, P. (1992). Assessment of child phonology in the clinical context. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 457 – 483). Timonium, MD: York.

Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin, A., Mazuka, R., & Dupoux, E. (2018). Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science, 42*(5), 1586–1617. https://doi.org/10.1111/cogs.12616

Gupta, P. (2012). Word learning as the confluence of memory mechanisms: Computational and neural evidence. In M. Faust (Ed.), *The Handbook of the Neuropsychology of Language* (pp. 146–163). Oxford, UK: Wiley-Blackwell. https://doi.org/10.1002/9781118432501.ch8

Gupta, P., & Cohen, N. J. (2002). Theoretical and computational analysis of skill learning, repetition priming, and procedural memory. *Psychological Review*, *109*(2), 401–448. https://doi.org/10.1037/0033-295X.109.2.401

Gupta, P., & Tisdale, J. (2009). Word learning, phonological short-term memory, phonotactic probability and long-term memory: towards an integrated framework. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3755–3771. https://doi.org/10.1098/rstb.2009.0132

H2O.ai. (2016). R Interface for H2O, R package. GitHub repository.

Haebig, E., Kaushanskaya, M., & Ellis Weismer, S. (2015). Lexical processing in school-age children with autism spectrum disorder and children with specific language impairment: The role of semantics. *Journal of Autism and Developmental Disorders*, *45*(12), 4109–4123. https://doi.org/10.1007/s10803-015-2534-2

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate Data Analysis* (3rd edition). New York: Macmillan.

Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language, 27,* 689-705. Retrieved from: http://centaur.reading.ac.uk/4542/1/Hamilton.Plunkett.Schafer.pdf

Hammer, C. S., Morgan, P., Farkas, G., Hillemeier, M., Bitetti, D., & Maczuga, S.

(2017). Late Talkers: A population-based study of risk factors and school readiness consequences. *Journal of Speech, Language, and Hearing Research*, *60*(3), 607–626. https://doi.org/10.1044/2016_JSLHR-L-15-0417

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hedlund, G., & Rose, Y. (2019). Phon 3.0. [Computer software].Available from https://www.phon.ca/phon-manual/misc/Welcome.html

Hogan, T. P., Bowles, R. P., Catts, H. W., & Storkel, H. L. (2011). The influence of neighborhood density and word frequency on phoneme awareness in 2nd and 4th grades. *Journal of Communication Disorders*, *44*(1), 49–58. https://doi.org/10.1016/j.jcomdis.2010.07.002

Hollich, G., Jusczyk, P. W., & Luce, P. A. (2002). Lexical neighborhood effects in 17-month-old word learning. In B, Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 314–323). Boston, MA: Cascadilla Press.

Holm, A., Crosbie, S., & Dodd, B. (2007). Differentiating normal variability from inconsistency in children's speech: Normative data. International Journal of Language and Communication Disorders, *42*(4), 467–86. https://doi.org/10.1080/13682820600988967

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100–116. https://doi.org/10.1016/j.jml.2010.02.003

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology. Human Perception and Performance*, *26*(5), 1570–1582. http://www.ncbi.nlm.nih.gov/pubmed/11039485

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/0096-1523.26.5.1570

Hunt, R. R., & Worthen, J. B. (Eds.). (2006). Distinctiveness and Memory. Oxford University Press. Retrieved from: https://doi.org/10.1093/acprof:oso/9780195169669.001.0001

Ingram, D. (2002). The measurement of whole-word productions. *Journal of Child Language*, *29*(04). https://doi.org/10.1017/S0305000902005275

James, D., Van Steenbrugge, W., & Chiveralls, K. (1994). Underlying deficits in language-disordered children with central auditory processing difficulties. *Applied Psycholinguistics*, *15*(3), 311–328. https://doi.org/10.1017/S0142716400065917

Jimenez, E., & Hills, T. (2017). Network analysis of a large sample of typical and late talkers. In In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2302–2307). Austin, TX: Cognitive Science Society. Retrieved from https://mindmodeling.org/cogsci2017/papers/0438/paper0438.pdf

John, F., Weisberg, S., Adler, D., Bates, D., Baud-bovy, G., Ellison, S., … Venables, W. (2017). Package 'car.' *CRAN Repository*. https://doi.org/10.1177/0049124105277200

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.

Jones, S. D. (2019). Accuracy and variability in early spontaneous word production. *First Language, 40*(2), 128–150. Repository: https://osf.io/w9y27/

Jones, S. D., & Brandt, S. (2018). Auditory lexical decisions in developmental language disorder: A meta-analysis of behavioral studies. *Journal of Speech, Language, and Hearing Research*, *61*(7), 1766–1783. https://doi.org/10.1044/2018_JSLHR-L-17-0447. Repository: https://osf.io/2cvnm/

Jones, S. D. &, Brandt, S. (2019a). Do children really acquire dense neighbourhoods? *Journal of Child Language, 46*(6), 1260–1273. https://doi.org/10.1017/S0305000919000473

Jones, S. D., & Brandt, S. (2019b). Neighborhood density and word production in delayed and advanced learners. *Journal of Speech, Language, and Hearing Research*, 1–8. https://doi.org/10.1044/2019_JSLHR-L-18-0468. Repository: https://osf.io/p8ax4/

Jones, S. D., & Brandt, S. (2020). Density and distinctiveness in early word learning: Evidence from neural network simulations. *Cognitive Science, 44*: Article e12812. Repository: https://osf.io/2qk5j/

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

Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *53*(3), 739–756. https://doi.org/10.1044/1092-4388(2009/08-0248)

Kent, R. D. (1992). The biology of phonological developement. In C. A. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 65–90). Timonium, MD: York Press.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, *55*(4), 306–353. https://doi.org/10.1016/j.cogpsych.2007.01.001

Leonard, L. B. (2009). Is expressive language disorder an accurate diagnostic category? *American Journal of Speech-Language Pathology*, *18*(2), 115–123. https://doi.org/10.1044/1058-0360(2008/08-0064)

Leonard, L. B. (2014). *Children with specific language impairment* (2nd ed.). Massachusetts: MIT.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195. https://doi.org/10.1016/j.cognition.2016.04.003

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 408–415. https://doi.org/10.3758/BF03195469

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Lum, J. A. G., Conti-Ramsden, G., Morgan, A. T., & Ullman, M. T. (2014). Procedural learning deficits in specific language impairment (SLI): A meta-analysis of serial reaction time task performance. *Cortex*, *51*(1), 1–10. https://doi.org/10.1016/j.cortex.2013.10.011

Macrae, T. (2013). Lexical and child-related factors in word variability and accuracy in infants. In *Clinical Linguistics and Phonetics, 27*(6–7), 497–507. https://doi.org/10.3109/02699206.2012.752867

Macrae, T., & Sosa, A. V. (2015). Predictors of token-to-token inconsistency in preschool children with typical speech-language development. *Clinical Linguistics & Phonetics*, *29*(12), 922–937. https://doi.org/10.3109/02699206.2015.1063085

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs. The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs (3rd ed.).* (3rd ed.). New York: Psychology Press (Taylor and Francis group).

Maillart, C., Schelstraete, M.-A., & Hupet, M. (2004). Phonological representations in children with SLI: A study of French. *Journal of Speech, Language, and Hearing Research*, *47*(1), 187–198. https://doi.org/10.1044/1092-4388(2004/016)

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*. https://doi.org/10.1016/0010-0277(80)90015-3

McArthur, G. M., & Bishop, D. V. M. (2005). Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. *Brain and Language*, *94*(3), 260–273. https://doi.org/10.1016/j.bandl.2005.01.002

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. London: Taylor and Francis. https://doi.org/10.3102/1076998616659752

McKean, C., Letts, C., & Howard, D. (2014). Triggering word learning in children with Language Impairment: The effect of phonotactic probability and neighbourhood density. *Journal of Child Language*, *41*(6), 1224–1248. https://doi.org/10.1017/S0305000913000445

McLeod, S., & Hewett, S. R. (2008). Variability in the production of words containing consonant clusters by typical 2- and 3-year-old children. *Folia Phoniatrica et Logopaedica*, *60*(4), 163–172. https://doi.org/10.1159/000127835

Melby-Lervåg, M., Lervåg, A., Lyster, S. A. H., Klem, M., Hagtvet, B., & Hulme, C. (2012). Nonword-repetition ability does not appear to be a causal influence on children's vocabulary development. *Psychological Science, 23*(10), 1092–1098. https://doi.org/10.1177/0956797612443833

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science*, *271*(5245), 77–81. https://doi.org/10.1126/science.271.5245.77

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representations: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89–120). Mahwah, NJ: Lawrence Erlbaum.

Mody, M., Studdert-Kennedy, M., & Brady, S. (1997). Speech perception deficits in poor readers: auditory processing or phonological coding? *Journal of Experimental Child Psychology*, *64*(2), 199–231. https://doi.org/10.1006/jecp.1996.2343

Montgomery, J. W. (1999). Recognition of gated words by children with specific language impairment: An examination of lexical mapping. *Journal of Speech, Language, and Hearing Research*, *43*(3), 735–743.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., . . . Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisuisition for 4,300 dutch words. *Behaviour Research Methods*, *45*(1), 169–177. Retrieved from: https://www.ncbi.nlm.nih.gov/pubmed/22956359

Munson, B., Edwards, J., & Beckman, M. E. (2005). Relationships between nonword repetition accuracy and other measures of linguistic development in children with phonological disorders. *Journal of Speech, Language, and Hearing Research*, *48*(1), 61–78. https://doi.org/10.1044/1092-4388(2005/006)

Nivedita, M., & Borovsky, A. (2018). Building a lexical network. In G. Westermann & M. Nivedita (Eds.), *Early word learning* (pp. 57–69). Oxon: Routledge.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., … Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157–159. https://doi.org/10.2307/1164923

Ota, M., & Green, S. J. (2013). Input frequency and lexical variability in phonological development: A survival analysis of word-initial cluster production. *Journal of Child Language*, *40*(03), 539–566. https://doi.org/10.1017/S0305000912000074

Pan, Y, & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology and Infection, 136*, 421–431.

Partanen, E., Kujala, T., Naatanen, R., Liitola, A., Sambeth, A., & Huotilainen, M. (2013). Learning-induced neural plasticity of speech processing before birth. *Proceedings of the National Academy of Sciences*, *110*(37), 15145–15150. https://doi.org/10.1073/pnas.1302159110

Pater, J., Stager, C., & Werker, J. F. (2004). The perceptual acquisition of phonological contrasts. *Language*, *80*(3), 384–402. https://doi.org/10.1353/lan.2004.0141

Perry, L. K., Perlman, M., Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE, 10*(9). Retrieved from: https://doi.org/10.1371/journal.pone.0137147

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*(1), annurev-linguist-030514-125050.

https://doi.org/10.1146/annurev-linguist-030514-125050

Pizzioli, F., & Schelstraete, M.-A. (2007). Auditory lexical decision in children with specific language impairment. *Proceedings of the 31st Boston University Conference on Language Development*. Retrieved from: http://www.bu.edu/buckd/files/2011/05/31-Pizzioli1.pdf

Pizzioli, F., & Schelstraete, M.-A. (2011). Lexico-semantic processing in children with specific language impairment: The overactivation hypothesis. *Journal of Communication Disorders*, *44*(1), 75–90. https://doi.org/10.1016/j.jcomdis.2010.07.004

Pizzioli, F., & Schelstraete, M.-A. (2013). Real-time sentence processing in children with specific language impairment: The contribution of lexicosemantic, syntactic, and world-knowledge information. *Applied Psycholinguistics*, *34*, 1–30. https://doi.org/10.1017/S014271641100066X

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology 25*(2), 143–170. https://doi.org/10.1016/j.newideapsych.2007.02.001

Python Software Foundation. (2013). Python language reference. *Python Software Foundation*. https://doi.org/https://www.python.org/

Quémart, P., & Maillart, C. (2016). The sensitivity of children with SLI to phonotactic probabilities during lexical access. *Journal of Communication Disorders*, *61*, 48–59. https://doi.org/10.1016/j.jcomdis.2016.03.005

Quine, W. V. O. (1960). *Word and Object*. Massachusetts: MIT Press.

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from: http://www.R-project.org/

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences 53*(Part A), 58–74. https://doi.org/10.1016/j.langsci.2015.08.002

Ramus, F., Marshall, C. R., Rosen, S., & van der Lely, H. K. J. (2013). Phonological deficits in specific language impairment and developmental dyslexia: Towards a multidimensional model. *Brain*, *136*(2), 630–645.

https://doi.org/10.1093/brain/aws356

Rispens, J., & Baker, A. (2012). Nonword repetition: The relative contributions of phonological short-term memory and phonological representations in children with language and reading impairment. *Journal of Speech, Language, and Hearing Research*, *55*(3), 683–694. https://doi.org/10.1044/1092-4388(2011/10-0263)

Robey, R. R., & Dalebout, S. D. (1998). A tutorial on conducting meta-analyses of clinical outcome research. *Journal of Speech, Language, and Hearing Research*, *41*, 1227–1241. https://doi.org/1092-4388/98/4106-1227

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464–468. https://doi.org/10.1111/j.1095-8649.2006.01157.x

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rowe, M. L., & Leech, K. A. (2017). Individual differences in early word learning. In I. G. Westermann & N. Mani (Eds.), *Early word learning*. Abingdon, Oxon: CRC Press – Taylor & Francis.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1419773112

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986). Parallel distributed Processing: Explorations in the Microstructure of Cognition (Volume 1: Foundations). Cambridge, MA: MIT Press.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(2), 1928–1941. Retrieved from https://psyarxiv.com/93mwx

Schwartz, R. G., & Leonard, L. B. (1982). Do children pick and choose? An examination of phonological selection and avoidance in early lexical acquisition.

Journal of Child Language, 9(2), 319–336. https://doi.org/10.1017/S0305000900004748

Schwartz, R. G., Leonard, L. B., Frome Loeb, D., Swanson, L. A., & Loeb, D. M. (1987). Attempted sounds are sometimes not: An expanded view of phonological selection and avoidance. *Journal of Child Language, 14(3), 411*–418. https://doi.org/10.1017/S0305000900010205

Skousen, R. (1989). *Analogical Modeling of Language.* Dordrecht: Kluwer.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Snowling, M., Chiat, S., & Hulme, C. (1991). Words, nonwords, and phonological processes: Some comments on Gathercole, Willis, Emslie, and Baddeley. *Applied Psycholinguistics 12*(3), 369–373. https://doi.org/10.1017/S0142716400009279

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, *40*(1), 83–108. https://doi.org/10.1006/jmla.1998.2614

Sosa, A. V. (2015). Intraword variability in typical speech development. *American Journal of Speech-Language Pathology*, *24*(1), 24–35. https://doi.org/10.1044/2014_AJSLP-13-0148

Sosa, A. V., & Stoel-Gammon, C. (2006). Patterns of intra-word phonological variability during the second year of life. *Journal of Child Language*, *33*(1), 31–50. https://doi.org/10.1017/S0305000905007166

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and phonological effects in early word production. *Journal of Speech Language and Hearing Research*, *55*(2), 596–608. https://doi.org/10.1044/1092-4388(2011/10-0113)

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102

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

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

Stokes, S. F., Kern, S., & Dos Santos, C. (2012). Extended statistical learning as an account for slow vocabulary growth. *Journal of Child Language*, *39*(1), 105–129. https://doi.org/10.1017/S0305000911000031

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language*, *29*(2), 251–274. https://doi.org/10.1017/S0305000902005032

Storkel, H. L. (2004). Do children acquire dense neighbourhoods? An investigation of similarity neighbourhoods in lexical acquisition. *Applied Psycholinguistics, 25*(2), 201–221. Retrieved from: https://doi.org/10.1017/S0142716404001109

Storkel, H. L. (2006). Do children still pick and choose? The relationship between phonological knowledge and lexical acquisition beyond 50 words. *Clinical Linguistics and Phonetics*, *20*(7–8), 523–529. https://doi.org/10.1080/02699200500266349

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

Storkel, H. L., & Lee, S. (2011). The independent effects of phonotactic probability and neighborhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, *26*(2), 191–211. https://doi.org/10.1080/01690961003787609

Strong, G. K., Torgerson, C. J., Torgerson, D., & Hulme, C. (2011). A systematic meta-analytic review of evidence for the effectiveness of the 'Fast ForWord' language intervention program. *Journal of Child Psychology and Psychiatry*, *52*(3), 224–235. https://doi.org/10.1111/j.1469-7610.2010.02329.x

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin and Review*, *18*(3), 605–11. https://doi.org/10.3758/s13423-011-0078-9

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound.

Developmental Science, *8*(5), 432–443. https://doi.org/10.1111/j.1467-7687.2005.00432.x

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., … Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, *271*(5245), 81–84. https://doi.org/10.1126/science.271.5245.81

Tallal, Paula, & Piercy, M. (1973). Developmental aphasia: Impaired rate of non-verbal processing as a function of sensory modality. *Neuropsychologia, 11*(4), 389–398. https://doi.org/10.1016/0028-3932(73)90025-0

Tallal, Paula, Stark, R. E., & Curtiss, B. (1976). Relation between speech perception and speech production impairment in children with developmental dysphasia. *Brain and Language, 3*(2), 305–317. https://doi.org/10.1016/0093-934X(76)90025-0

Thal, D. J., Marchman, V. A., & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued delay. In L. Rescorla & P. Dale (Eds.), *Late Talkers: Language Development, Interventions, and Outcomes*. Baltimore, MD: Brookes.

Theakston, A. L., Lieven, E. V, Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language, 28*(1), 127–152. https://doi.org/10.1017/S0305000900004608

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Massachusetts: Harvard University Press.

Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007). Procedural learning in adolescents with and without specific language impairment. *Language Learning and Development*, *3*(4), 269–293. https://doi.org/10.1080/15475440701377477

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods, 51*(3), 1187–1204. https://doi.org/10.3758/s13428-018-1056-1

Twomey, K. E., & Westermann, G. (2017). Curiosity-based learning in infants: A neurocomputational approach. *Developmental Science*, *21*(4), e12629. https://doi.org/10.1111/desc.12629

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex*, *41*(3), 399–433. https://doi.org/10.1016/S0010-9452(08)70276-4

Van Der Feest, S. V. H., & Fikkert, P. (2015). Building phonological lexical representations. *Phonology*, *32*(2), 207–239. https://doi.org/10.1017/S0952675715000135

Van Der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, 6, 111-122. https://CRAN.R-project.org/package=stringdist.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*(6), 1176-1190. Retrived from: http://crr.ugent.be/ papers/SUBTLEX-UK_ms.pdf

Vance, M. (2008). Short-term memory in children with developmental language disorder. In C. F. Norbury, J. B. Tomblin, & D. V. M. Bishop (Eds.), *Understanding Developmental Language Disorders: From Theory to Practice* (pp. 23–38). Hove, England, UK: Psychology Press.

Ventura, P., Kolinsky, R., Fernandes, S., Querido, L., & Morais, J. (2007). Lexical restructuring in the absence of literacy. *Cognition*, *105*(2), 334–361. https://doi.org/10.1016/j.cognition.2006.10.002

Viechtbauer, W. (2010). Conducting meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.

https://doi.org/10.1044/1092-4388(2008/030)

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

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the Acquisition of Phonological Word Forms with Computational Experiments. *Language and Speech*, *56*(4), 493–527. https://doi.org/10.1177/0023830912460513

Vitevitch, M. S., & Storkel, H. L. (2013). Examining the acquisition of phonological word forms with computational experiments. *Language and Speech*, *56*(4), 493–527. https://doi.org/10.1177/0023830912460513

Vitevitch, M. S., Storkel, H. L., Francisco, A. C., Evans, K. J., & Goldstein, R. (2014). The influence of known-word frequency on the acquisition of new neighbours in adults: evidence for exemplar representations in word learning. *Language, Cognition and Neuroscience*, *29*(10), 1311–1316. https://doi.org/10.1080/23273798.2014.912342

Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*(3), 270–276. http://www.ncbi.nlm.nih.gov/pubmed/15595367

Walley, A. C. (1993). The role of vocabulary development in children's spoken word recognition and segmentation ability. *Developmental Review*, *13*(3), 286–350. https://doi.org/10.1006/drev.1993.1015

Walley, A. C., Metsala, J. L., & Garlock, V. M. (2003). Spoken vocabulary growth: Its role in the development of phoneme awareness and early reading ability. *Reading and Writing: An Interdisciplinary Journal*, *16*(1), 5–20. https://doi.org/10.1023/A:1021789804977

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. Retrieved from: https://doi.org/10.3758/s13428-012-0314-x

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401. https://doi.org/10.1002/wcs.1178

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2017). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21(2),* e12552. https://doi.org/10.1111/desc.12552

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*(1), 114–132. https://doi.org/10.1016/j.jml.2008.03.001

Wickham et al., (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Windsor, J., & Hwang, M. (1999). Children’s auditory lexical decisions: A limited processing capacity account of language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(4), 990–1002. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10912254

Appendix A Auditory Lexical Decisions in Developmental Language Disorder: A Meta-Analysis of Behavioural Studies

A.1 Study summaries

Table A.1.1: Summary characteristics of included studies. DLD = developmental language disorder; AMC = age-matched controls; LMC = language-matched controls; RT = reaction time; SD = standard deviation; nr = not reported. See primary literature for standardised test references.

Citation (date)	Groups (n)	Mean age (SD)	Standardised tests used/inclusion criteria	Stimuli (n)	Response type	Measure
James, Van Steenbrugge, and Chiveralls (1994)	DLD (6)	9;09 (0;11)	Linguistic: • Peabody Picture Vocabulary Test Revised (PPVT-R) • Test for the Reception of Grammar (TROG) • Neale Analysis of Reading Ability-Revised • Staggered Spondaic Word test (SSW)	Familiar words ($n = 40$) Viable non-words ($n = 40$)	Nr: Assumed verbal but unclear	Accuracy
	AMC (6)	9;09 (0;11)				
	LMC (6)	7;09 (0;10)				

			• Competing Sentence Test (CST) Non-linguistic: • Peripheral hearing status and non-verbal IQ (RCPM, see p. 314) in normal range			
Edwards and Lahey (1996)	DLD (46)	7:3 (1:7)	Independent diagnosis by certified SLT Linguistic: • Clinical Evaluation of Language Fundamentals Revised (CELF-R) • Peabody Picture Vocabulary Test Revised (PPVT-R) • Test of Language Development-2-Primary (TOLD-P) • Illinois Test of Psycholinguistic Abilities Non-linguistic: • Test of Non-verbal Intelligence (TONI) • Kauffman Brief Intelligence Test	Familiar words (*n* = 20) Legal non-words (*n* = 20)	Verbal; ‘yes’/‘no’	Response time
	Subgroup 1: DLD-expressive (10)	7;0 (1;3)				
	Subgroup 2: DLD-mix (20)	8;0 (1;3)				
	AMC (46)	7:3 (1:6)				

Windsor and Hwang (1999)	DLD (20)	11;4 (nr)	Linguistic: - Peabody Picture Vocabulary Test Revised (PPVT-R) - Test of Language Development-Intermediate (TOLD-I) Non-linguistic: - Test of Non-verbal Intelligence (TONI) - Hearing screening test	Study A: Real word derivatives (n = 20) Pseudo derivatives (n = 20) Foils (n = 20) Study B: Phonologically transparent (PT) real derivatives (n = 15) Phonologically opaque (PO) real derivatives (n = 15) PT pseudo derivatives (n = 15) PO pseudo derivatives (n = 15) Foils (n = 15)	Button pressing on computer; ‘yes’/‘no’	Response time
	AMC (20)	11;4 (nr)				
	LMC (20)	9;0 (nr)				
Maillart, Schelstraete, and Hupet (2004)	DLD 1; lexical age 5;0 (7)	7;8 (nr)	Independent diagnosis by certified SLT Linguistic: - Peabody Picture Vocabulary	Real words (n = 24) High-similarity non-words (n = 12) Low-similarity non-words (n = 12)	Verbal; ‘yes’/‘no’ response to uttered (i.e. not pre-recorded) words	Accuracy
	DLD 1 LMC (16)	5;1 (nr)				
	DLD 2; lexical age 6;0 (10)	9;2 (nr)				

	DLD 2 LMC (21)	5;5 (nr)	Test Revised (PPVT-R) • French equivalent of the Test for the Reception of Grammar (TROG); ECOSSE Non-linguistic: • Leiter International Performance Scales • Echelle d'Intelligence Pour Enfants (3ème éd.) [Wechsler Intelligence Scale for Children (3rd ed.)]	Word-initial manipulation (*n* = 8) Word-medial manipulation (*n* = 8) Word-final manipulation (*n* = 8)		
	DLD 3; lexical age 7;0 (8)	9;2 (nr)				
	DLD 3 LMC (11)	6;10 (nr)				
Crosbie, Howard, and Dodd (2004)	DLD (15)	8;11 (0;9)	Independent diagnosis by certified SLT and educational psychologist Linguistic: • British Picture Vocabulary Scales • The German Test of Word Finding • South Tyneside Assessment of Phonology (STAP)	Real words (*n* = 20) Real word foils (*n* = 20) Legal non-words (*n* = 20) Phonotactically illegal non-words (*n* = 20)	Verbal; 'yes'/'no' response to pre-recorded words	Accuracy and response time
	AMC (15)	9;3 (0;7)				
	LMC (15)	6;10 (0;11)				

			Non-linguistic: • Test of Non-verbal Intelligence-3 (TONI-3)			
Befi-Lopes, Pereira, and Bento (2010)	Lexical age 4 DLD (5)	Range = 3;8-5;0	Independent diagnosis of SLI by certified SLT Linguistic: • Receptive vocabulary test from the Laboratory for Investigation in Language Development and Alterations, at the Faculdade de Medicina da Universidade de São Paulo (see p. 306).	Real words (*n* = 24) Non-words varying in degree, position, and type of modification (*n* = 24). See p. 309 for summary	Verbal; 'yes'/'no' response to pre-recorded words	Accuracy
	Lexical age 4 control (10)	Range = 4;1-4;11				
	Lexical age 5 DLD (6)	Range = 4;10-7;9				
	Lexical age 5 control (12)	Range = 5;1-5;7				
	Lexical age 6 DLD (7)	Range = 4;10-8;9				
	Lexical age 6 control (14)	Range = 6;0-6;8				
Pizzioli and Schelstraete (2013)	DLD (13)	10;1 (1;02)	Independent diagnosis by certified SLT Linguistic: • French equivalent of the Peabody Picture Vocabulary Test Revised (PPVT-R); EVIP • French equivalent of the Test for the Reception of Grammar	Target real words (*n* = 42) Filler real words (*n* = 60) Pseudowords (*n* = 68)	Button pressing: green = real word; red = non-word	Accuracy and response time
	AMC (13)	10;2 (1;0)				
	LMC (13)	7;7 (0;5)				

			(TROG); ECOSSE Non-linguistic: • Hearing assessment • Wechsler Intelligence Scale for Children-Revised • No history of neurological dysfunction or psychopathology			
Haebig, Kaushanskaya, and Ellis Weismer (2015)	DLD (28)	10;0 (1;5)	Linguistic: • The Clinical Evaluation of Language Fundamentals – fourth edition (CELF-4) • Peabody Picture Vocabulary Test, fourth edition (PPVT-4) Non-linguistic: • Wechsler Intelligence Scale for Children, fourth edition.	Words (*n* = 40) Non-words A; low semantic network (*n* = 20) Non-words B; high semantic network (*n* = 20)	Button pressing: smiling face = real word; frowning face = non-word	Accuracy and response time
	LMC (30)	9;7 (1;8)				
Quémart and Maillart (2016)	DLD (20)	10;1 (1;10)	Independent diagnosis by certified team of practitioners Linguistic:	Bi-syllabic words (*n* = 120) High phonotactic	Button pressing: smiley key = real word; red	Accuracy and response time
	AMC (20)	10;0 (1;10)				

	LMC (20)	7;4 (0;7)	• Evaluation du Langage Oral (ELO) • Langage Oral, Langage Ecrit, Mémoire et Attention (L2MA2) Non-linguistic: • Hearing assessment • Wechsler Intelligence Scale for Children	probability (PP) non-words (*n* = 60) Low PP non-words (*n* = 60)	‘X’ = non-word	

A.2 Forest plots

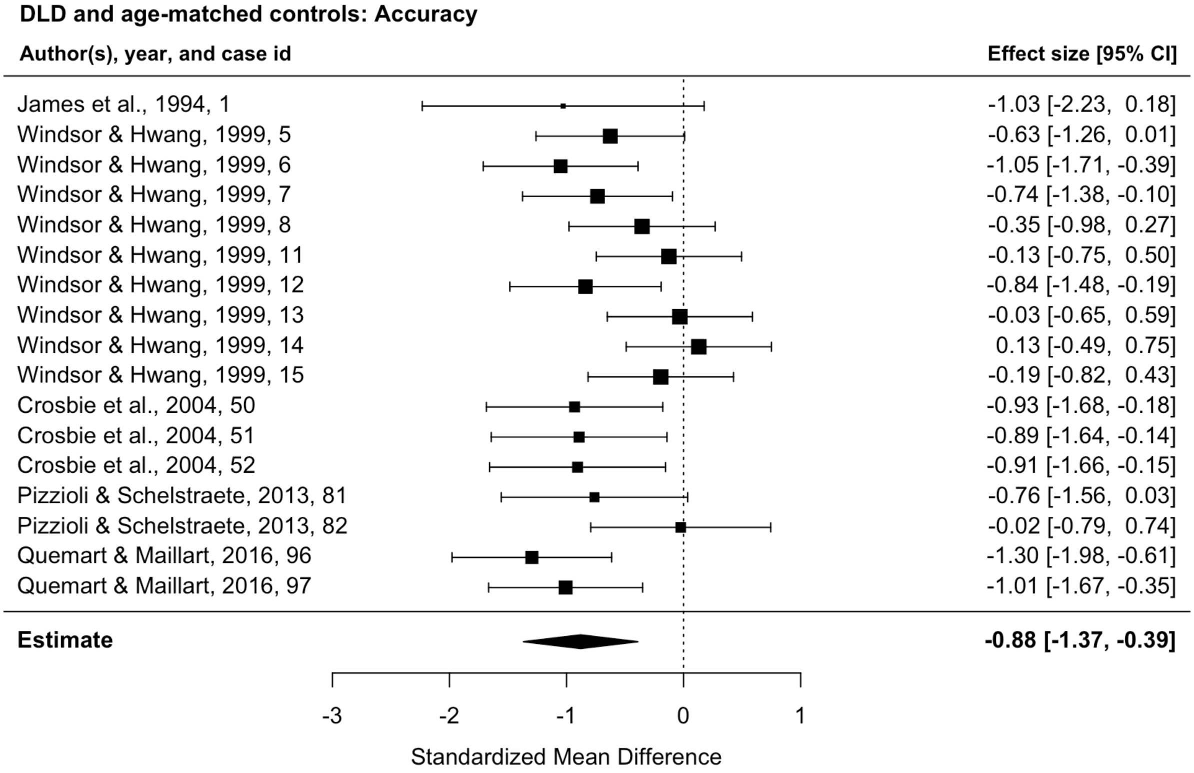

Figure A.2.1: Forest plot showing case and summary (i.e. 'Estimate') Hedges' *g* and 95% confidence intervals for the DLD/age-matched control comparison on the accuracy measure. Case id indexes individual effect sizes as listed in the master dataset. Negative effect sizes indicate children with DLD were less accurate than controls.

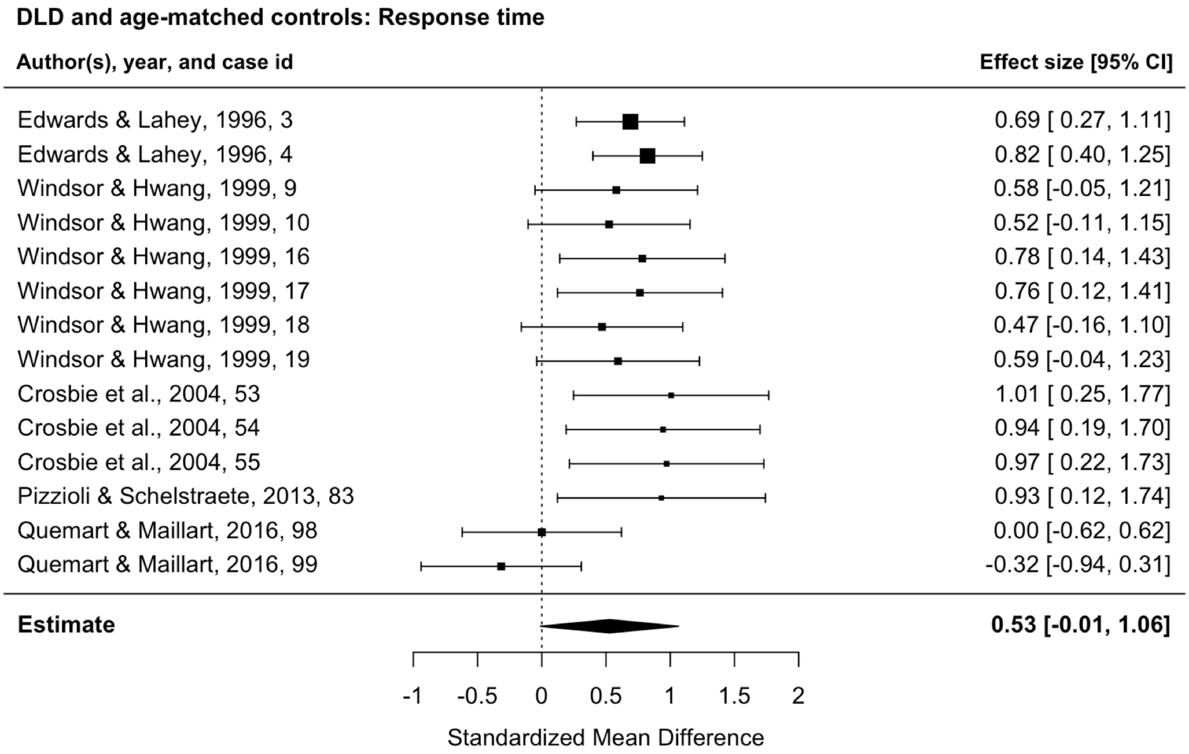

Figure A.2.2: Forest plot showing case and summary (i.e. ‘Estimate’) Hedges’ *g* and 95% confidence intervals for the DLD/age-matched control comparison on the response time measure. Case id indexes individual effect sizes as listed in the master dataset. Positive effect sizes indicate children with DLD took longer to respond than controls.

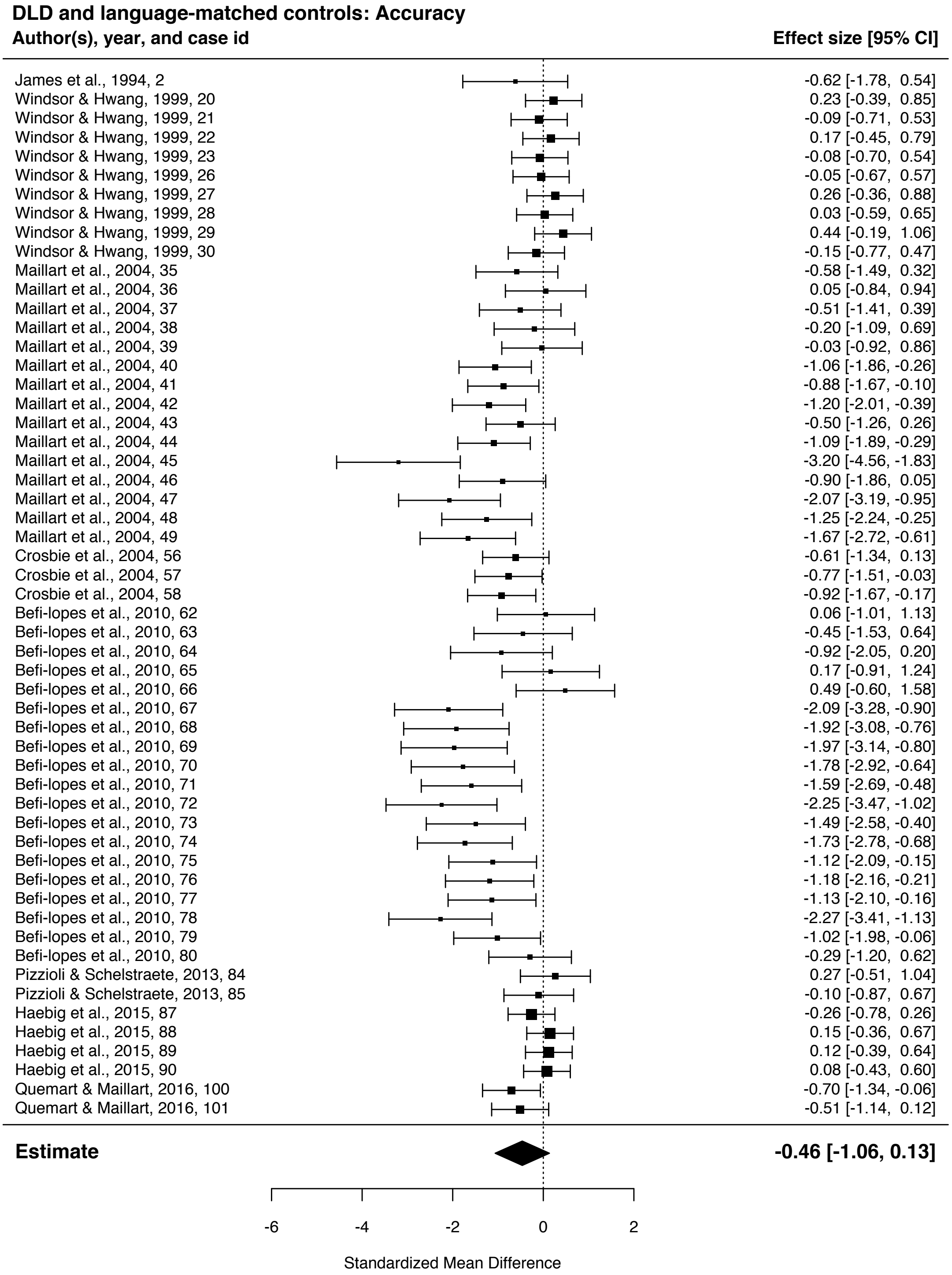

Figure A.2.3: Forest plot showing case and summary (i.e. ‘Estimate’) Hedges’ g and 95% confidence intervals for the DLD/language-matched control comparison on the accuracy measure. Case id indexes individual effect sizes as listed in the master dataset. Negative effect sizes indicate children with DLD were less accurate than controls.

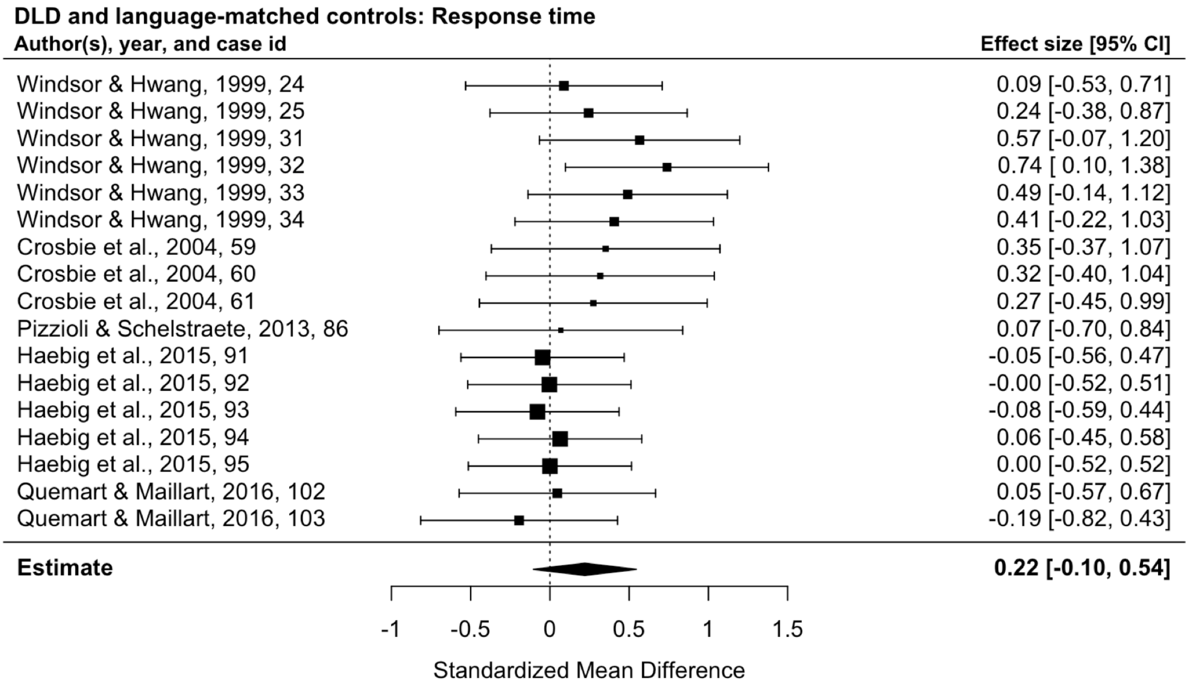

Figure A.2.4: Forest plot showing case and summary (i.e. ‘Estimate’) Hedges’ *g* and 95% confidence intervals for the DLD/language-matched control comparison on the response time measure. Case id indexes individual effect sizes as listed in the master dataset. Positive effect sizes indicate children with DLD took longer to respond than controls

Appendix B Do Children Really Acquire Dense Neighbourhoods?

B.1 Predictor correlations

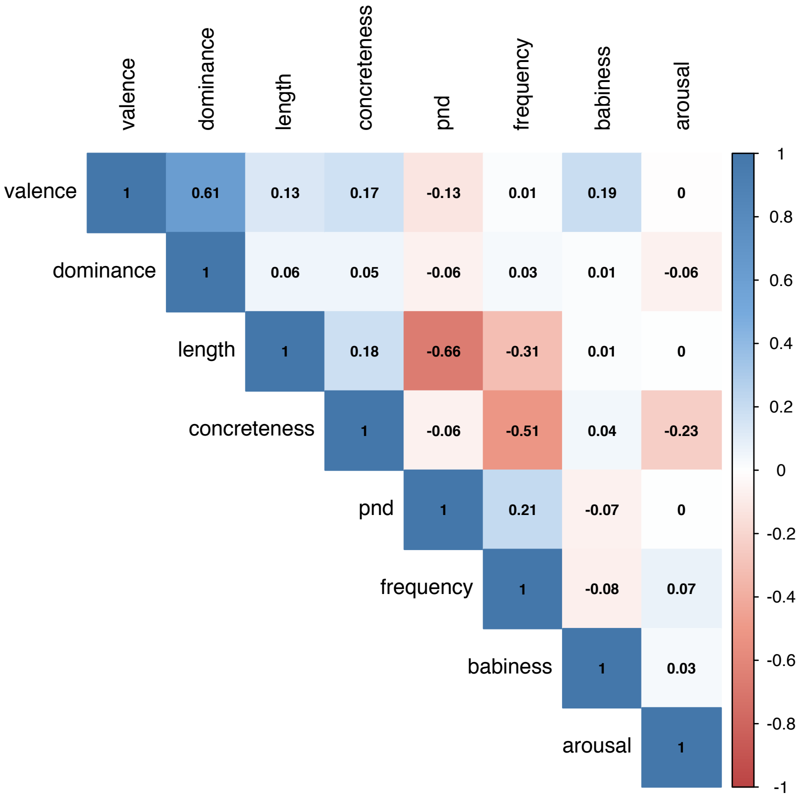

Figure B.1.1 Post-imputation Pearson correlations between predictors (pnd indicates phonological neighbourhood density).

B.2 Missing data and variance inflation factors

Table B.2.1: Rates of missing data and variance inflation factors for each predictor variable, calculated (using the car and lme4 packages in R) from the model: glmer(cbind(understands, produces) ~ length + pnd + frequency + babiness + concreteness + valence + arousal + dominance + (1 | word), family = binomial). Note that VIFs are shown for post-imputation values.

Predictor	Missing (%)	VIF
Frequency	5.5	1.50
Length	0	1.93
Babiness	22.73	1.08
Concreteness	4.55	1.52
Phonological neighbourhood density (PND)	3.35	1.83
Valence	18.18	1.79
Arousal	18.18	1.08
Dominance	18.18	1.63

B.3 Model summaries

Table B.3.1: Model summary for the understands outcome, showing term, estimate, standard error (Std. error), and lower and upper 95% confidence intervals (CI). CDS indicates child-directed speech. PND indicates phonological neighbourhood density.

Term: Understands	Estimate	Std. error	Lower 95% CI	Upper 95% CI
Intercept	-1.25	0.04	-1.32	-1.18
CDS frequency	0.12	0.04	0.05	0.21
PND	0	0.05	-0.1	0.09
Length (phonemes)	0.06	0.05	-0.03	0.15
Babiness	0.14	0.03	0.07	0.21
Concreteness	0.18	0.04	0.10	0.26
Valence	-0.02	0.05	-0.11	0.06
Arousal	-0.04	0.03	-0.11	0.02
Dominance	0.1	0.04	0.01	0.18
Age	0.11	0.02	0.08	0.15
Interactions				
CDS frequency: Age	-0.08	0.02	-0.13	-0.04
PND: Age	-0.02	0.03	-0.07	0.03
Length (phonemes): Age	0.01	0.03	-0.04	0.06
Babiness: Age	-0.06	0.02	-0.1	-0.03
Concreteness: Age	-0.15	0.02	-0.20	-0.11
Valence: Age	-0.03	0.02	-0.08	0.02
Arousal: Age	-0.01	0.02	-0.05	0.02
Dominance: Age	0.01	0.02	-0.03	0.06
Standard deviations (SD) and correlations (Corr)				
SD: Word intercept	0.67	0.02	0.62	0.71
SD: Age slope, word intercept	0.33	0.02	0.3	0.35
Corr: Age slope, word intercept	-0.58	0.04	-0.66	-0.51

Table B.3.2: Model summary for the produces outcome, showing term, estimate, standard error (Std. error), and lower and upper 95% confidence intervals (CI). CDS indicates child-directed speech. PND indicates phonological neighbourhood density.

Term: Produces	Estimate	Std. error	Lower 95% CI	Upper 95% CI
Intercept	-2.21	0.06	-2.33	-2.09
CDS frequency	0.2	0.07	0.07	0.34
PND	0.13	0.08	-0.03	0.28
Length (phonemes)	-0.07	0.08	-0.22	0.09
Babiness	0.17	0.06	0.06	0.28
Concreteness	0.42	0.07	0.29	0.56
Valence	0.09	0.07	-0.05	0.24
Arousal	0.06	0.06	-0.05	0.18
Dominance	-0.08	0.07	-0.21	0.06
Age	1.43	0.02	1.39	1.46
Interactions				
CDS frequency: Age	0.04	0.02	0.01	0.08
PND: Age	-0.01	0.02	-0.05	0.03
Length (phonemes): Age	0.04	0.02	-0.01	0.08
Babiness: Age	-0.04	0.01	-0.07	-0.01
Concreteness: Age	0.04	0.02	0.01	0.08
Valence: Age	-0.02	0.02	-0.06	0.02
Arousal: Age	-0.00	0.02	-0.03	0.03
Dominance: Age	0.03	0.02	-0.00	0.07
Standard deviations (SD) and correlations (Corr)				
SD: Word intercept	1.11	0.04	1.03	1.20
SD: Age slope, word intercept	0.18	0.02	0.15	0.21
Corr: Age slope, word intercept	-0.96	0.03	-1	-0.89

Appendix C Neighbourhood Density and Word Production in Delayed and Advanced Learners

C.1 Model summary

Table C.1.1: Model summary showing term (main effects and interactions), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI).

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-2.23	0.02	-2.27	-2.19
SD of random intercepts	0.45	0.01	0.41	0.48
Length	-0.19	0.01	-0.2	-0.17
Vocabulary size	1.27	0.02	1.24	1.31
Frequency	0.58	0.00	0.57	0.6
Babiness	0.33	0.01	0.32	0.34
Concreteness	0.78	0.01	0.77	0.8
Neighbourhood density	0.07	0.01	0.06	0.08

Term (interactions)	Estimate	Std. error	L-95% CI	U-95% CI
Length: Vocabulary	0.08	0.01	0.07	0.09
Frequency: Vocabulary	0.03	0.01	0.01	0.04
Babiness: Vocabulary	-0.03	0.01	-0.04	-0.02
Concreteness: Vocabulary	0.11	0.01	0.11	0.13
Neighbourhood density: Vocabulary	-0.02	0.01	-0.04	-0.02

Appendix D Accuracy and Variability in Early Word Production

D.1 Model Summaries

Table D.1.1: Model summary for the production accuracy outcome (model 1; m.1), showing term, estimate, standard error (SE), and lower and upper 95% confidence intervals (CI). PLD20 indicates phonological neighbourhood density (i.e. average 20-step phonological Levenshtein distance). Terms are grouped into main effects, interactions, and family specific parameters.

Term	Estimate	SE	Lower 95% CI	Upper 95% CI
Intercept	0.47	0.00	0.46	0.47
Frequency	-0.12	0.00	-0.12	-0.12
Age	-0.03	0.00	-0.03	-0.03
PLD20	0.10	0.00	0.10	0.10
Frequency: Age	0.00	0.00	0.00	0.00
Age: PLD20	-0.01	0.00	-0.01	0.00
Frequency: PLD20	-0.02	0.00	-0.03	-0.02
Sigma	0.46	0.00	0.46	0.46
Hu	0.22	0.00	0.21	0.22

Table D.1.2: Model summary for the production variability outcome (model 2; m.2), showing term, estimate, standard error (SE), and lower and upper 95% confidence intervals (CI). PLD20 indicates phonological neighbourhood density. Terms are grouped into main effects, interactions, and family specific parameters.

Term	Estimate	SE	Lower 95% CI	Upper 95% CI
Intercept	-0.56	0.00	-0.56	-0.55
Frequency	-0.48	0.00	-0.48	-0.47
Age	-0.02	0.00	-0.03	-0.02
PLD20	0.10	0.00	0.10	0.11
Frequency: Age	0.02	0.00	0.01	0.02
Age: PLD20	0.02	0.00	0.02	0.03
Frequency: PLD20	0.01	0.00	0.01	0.02
Phi	5.96	0.02	5.93	5.99
Zoi	0.23	0.00	0.23	0.23
Coi	0.31	0.00	0.30	0.31

Appendix E Density and Distinctiveness in Early Word Learning: Evidence from Neural Network Simulations

E.1 Model Summaries

Table E.1.1: Test phase model summary showing term (main effects, interactions, and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: Mean reconstruction error ~ Length + Frequency + PLD20 + PLD20 * Length + PLD20 * Frequency.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-3.38	0.01	-3.4	-3.36
PLD20	0.18	0.02	0.14	0.22
Length	0.04	0.02	0	0.07
Frequency	-0.02	0.01	-0.04	0
Term (interactions)	Estimate	Std. error	L-95% CI	U-95% CI
PLD20: Length	-0.01	0.01	-0.02	0
PLD20: Frequency	-0.04	0.01	-0.06	-0.02
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Sigma	0.23	0.01	0.22	0.24

Table E.1.2: Validation phase model summary showing term (main effects and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: Produces (%) ~ Mean squared error.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-1.21	0.03	-1.25	-1.16
Mean squared error	-0.03	0.03	-0.08	0.02
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Shape	1.98	0.11	1.8	2.16

Table E.1.3: Generalisation phase model summary showing term (main effects and family specific parameters), estimate, standard error (Std. error), and lower (L) and upper (U) 95% confidence intervals (CI). Model formula: *Mean reconstruction error ~ PLD20 + Length*.

Term (main effects)	Estimate	Std. error	L-95% CI	U-95% CI
Intercept	-0.01	0	-0.01	0
PLD20	0.02	0	0.01	0.02
Length	0	0	0	0.01
Term (family specific parameters)	Estimate	Std. error	L-95% CI	U-95% CI
Sigma	0.02	0	0.02	0.02
Alpha	1.83	0.4	1.19	2.45