LANGUAGE DEVELOPMENT RESEARCH

An Open Science Journal

Volume 1 | Issue 1 | December 2021
ISSN 0000-0000-00

About the journal

Language Development Research: An Open-Science Journal was established in 2020 to meet the field's need for a peer- reviewed journal that is committed to fully open science: LDR charges no fees for readers or authors, and mandates full sharing of materials, data and analysis code. The intended audience is all researchers and professionals with an interest in language development and related fields: first language acquisition; typical and atypical language development; the development of spoken, signed or written languages; second language learning; bi- and multilingualism; artificial language learning; adult psycholinguistics; computational modeling; communication in nonhuman animals etc. The journal is managed by its editorial board and is not owned or published by any public or private company, registered charity or nonprofit organization.

Child Language Data Exchange System

Language Development Research is the official journal of the **TalkBank system**, comprising the CHILDES, PhonBank, HomeBank, FluencyBank, Multilingualism and Clinical banks, the CLAN software (used by hundreds of researchers worldwide to analyze children's spontaneous speech data), and the Info-CHILDES mailing list, the de-facto mailing list for the field of child language development with over 1,600 subscribers.

Diamond Open Access

Language Development Research is published using the Diamond Open Access model (also known as “Platinum” or “Universal” OA). The journal does not charge users for access (e.g., subscription or download fees) or authors for publication (e.g., article processing charges).

Clarifying revision made upon discovery of errata. Reissued on 20 December 2023.

Hosting

The **Carnegie Mellon University Library Publishing Service** (LPS) hosts the journal on a Janeway Publishing Platform with its manuscript management system (MMS) used for author submissions.

License

Peer Review and Submissions

All submissions are reviewed by a minimum of two peer reviewers, and one of our Action Editors, all well- established senior researchers, chosen to represent a wide range of theoretical and methodological expertise. Action Editors select peer reviewers based on their expertise and experience in publishing papers in the relevant topic area.

Submissions and Publication Cycle

We invite submissions that meet our criteria for rigour, without regard to the perceived novelty or importance of the findings. We publish general and special-topic articles (“Special Collections”) on a rolling basis to ensure rapid, cost-free publication for authors.

Language Development Research is published once a year, in December, with each issue containing the articles produced over the previous 12 months. Individual articles are published online as soon as they are produced. For citation purposes, articles are identified by the year of first publication and digital object identifier (DOI).

Table of Contents

Volume 1, Issue 1, 31 December 2021

1
Language Development Research Editorial: Why do we need another journal? Language Development Research

Ben Ambridge

doi: 10.34842/yxh6-kz16

9
Features of lexical richness in children's books: Comparisons with child-directed speech

Nicola Dawson, Yaling Hsiao, Alvin Wei Ming Tan, Nilanjana Banerji, and Kate Nation

doi: 10.34842/5we1-yk94

55
Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories

Benjamin E. deMayo, Danielle Kellier, Mika Braginsky, Christina Bergmann, Cielke Hendriks, Caroline F. Rowland, Michael C. Frank, and Virginia A. Marchman

doi: 10.34758/kr8e-w591

99
Passive sentence reversal errors in autism: Replicating Ambridge, Bidgood, and Thomas (2020)

Samuel Jones, Madeline Dooley, and Ben Ambridge

doi: 10.34842/g1zk-3715

123
Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? — A computational investigation

Khazar Khorrami and Okko Räsänen

doi: 10.34842/w3vw-s845

193
Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study

Giulia Bovolenta and Emma Marsden

doi: 10.34842/c7t4-pz50

245
Children's language abilities predict success in remote communication contexts

Karla K. McGregor, Ronald Pomper, Nichole Eden, Timothy Arbisi-Kelm, Nancy Ohlmann, Shivani Gajre, and Erin Smolak

doi: 10.34758/rzsq-zd16

283
Sleep behaviour in children with language disorder

Victoria C.P. Knowland, Mohreet Rauni, M. Gareth Gaskell, Sarah Walker, Elaine van Rijn, Lisa-Marie Henderson, and Courtenay Norbury

doi: 10.34842/fs09-ga12

Language Development Research Editorial: Why do we need another journal?

Ben Ambridge
University of Liverpool, UK

Abstract: *Language Development Research* is a platinum Open Access journal that commits to publishing "any empirical or theoretical paper that is relevant to the field of language development and that meets our criteria for rigour, without regard to the perceived novelty or importance of the findings". This commitment is designed to reduce publication bias and incentives to engage in questionable research practices.

Keywords: Language Development Research; publication bias; questionable research practices; HARKing; *p*-hacking.

Corresponding author(s): Ben Ambridge, Department of Psychology, University of Liverpool, Bedford St South, Liverpool, L79 7ZA, UK. Email: Ben.Ambridge@Liverpool.ac.uk.

ORCID ID(s): https://orcid.org/0000-0003-2389-8477

Citation: Ambridge, B. (2021). Language Development Research Editorial: Why do we need another journal? *Language Development Research*, 1(1), p 1-8. DOI: 10.34842/yxh6-kz16

Why does the field of language development need a new journal? On the face of it, we are already well served. In addition to the non-specialist journals, we have several general child-language journals (i.e., journals concerned primarily with language acquisition and development in the pre-school and early-school years), plus specialist journals focussing on language disorders, second language learning, and bilingualism, as well as various linguistics journals.

The problem is that the vast majority of these journals, including – in my estimation – all existing language-development journals, are *selective* journals. That is, they endeavour to publish the best of the submissions that they receive. Exactly what constitutes the "best" is rarely made explicit, but selectivity is implicit in the superlatives in journal "About" statements, and in the hierarchy of journals in the heads of seasoned researchers.

Publishing only those papers that make a sufficiently novel or important contribution sounds laudable, until we consider the flip side: a reluctance to publish papers that don't reach a journal's (implicit) criteria for novelty, importance or broad interest; for example, because they replicate or extend a previous study, or because they report null findings, or findings that are simply unclear or messy.

Selectivity – selection on the basis of factors other than scientific rigour – distorts the scientific literature by introducing three major biases into the publication process (de Vries, Roest, de Jonge, Cuijpers, Munafò & Bastiaansen, 2018): (a) *publication bias*, whereby studies with null results are rejected or never submitted in the first place, (b) *outcome-reporting bias*, whereby researchers drop groups, conditions or sub-studies that fail to show a clear and/or desired effect and (c) *spin*, drawing conclusions that are not merited by the findings. Finally, if null findings do make it into the literature, they are less likely to be cited. Using the example of antidepressant drugs, de Vries et al (2018) show how these biases translate an evidence base that is, in reality, almost exactly 50/50 (the FDA classified 53/105 trials as positive) into a literature that offers overwhelmingly positive support for these drugs' efficacy (see Figure 1, reproduced from Figure 1 in de Vries et al, 2018).

Perhaps most seriously of all, selectivity all but compels fundamentally-honest researchers to engage in questionable research practices (John, Loewenstein & Prelec, 2012) such as *p-hacking* (Simons, Nelson & Simonsohn, 2011) – rerunning analyses with different coding, exclusions, covariates, transformations, statistical tests, sample sizes, and so on – and *hypothesising after results are known* (HARKing; Kerr, 1998), "reframing" the paper around a serendipitous finding that was not originally the question of primary interest (or even, in some cases, "fishing" or "data mining": collecting data in a purely exploratory fashion and only afterwards formulating theoretical claims or hypotheses). Sometimes these practices are intentional. Sometimes, and with the best of intentions, journal reviewers and editors even request them explicitly. Sometimes they are entirely unintentional. After all, decisions have to be made regarding coding, exclusions, transformations and so on, and if one set of decisions allows us to see the otherwise-obscured effect that we confidently expected to be there all along, we are likely to genuinely believe that this is the correct one.

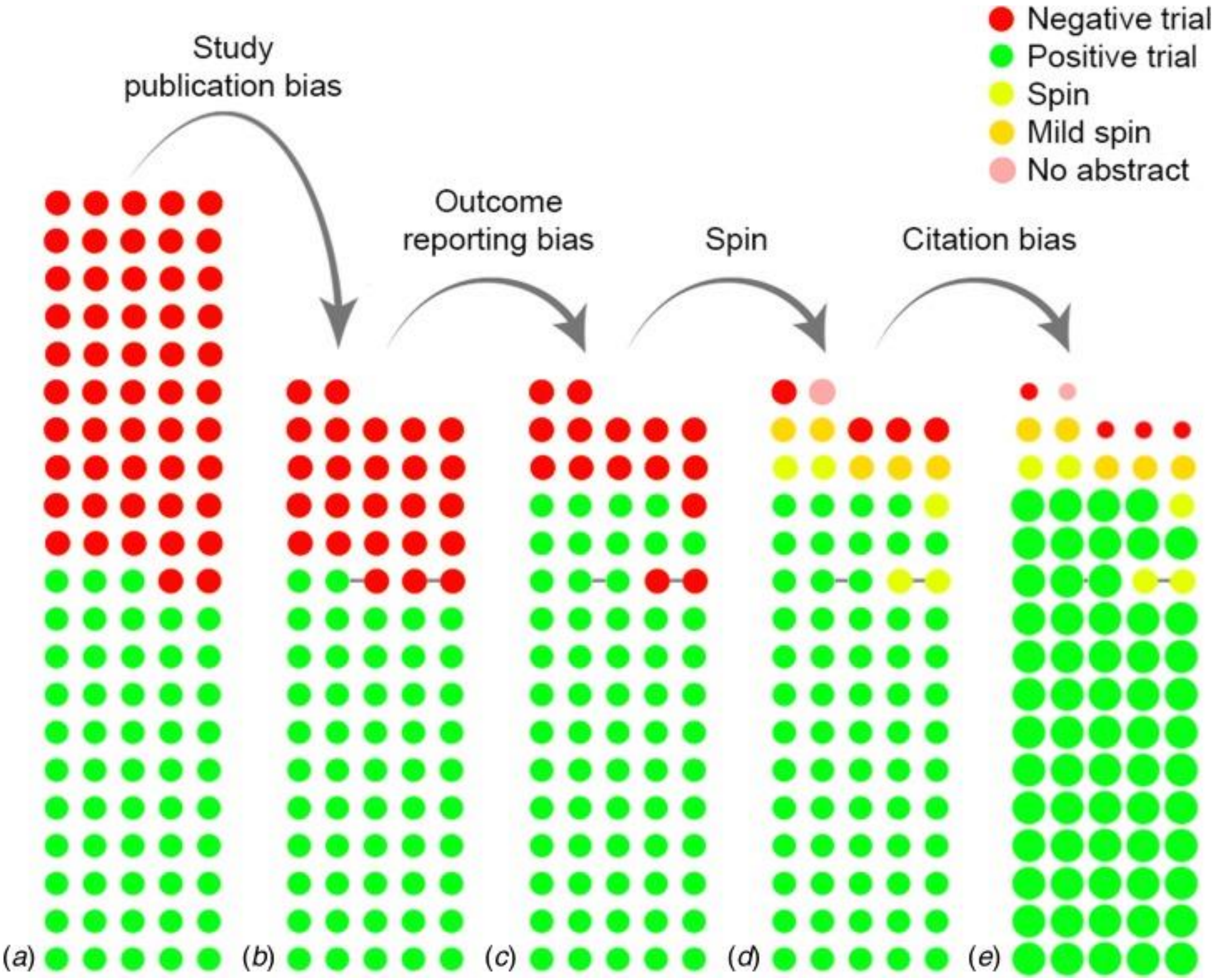


Figure 1. How *publication bias, outcome-reporting bias, spin* and *citation bias* skew the evidence base (from de Vries et al, 2018, creative commons licence).

The good news is that, at least in some fields, we seem to be moving in the right direction. Eason, Hamlin and Sommerville's (2017) survey of infancy researchers found that relatively few reported adding participants until *p* is <0.05 (2%), adding participants until they are confident that there is or is not an effect (11%), excluding dependent measures that yielded nonsignificant results (5%) or results that were inconsistent with the initial hypothesis (1%), exploring different transformations of their data and using the most favourable one (1%), or planning statistical analyses only once the data are in hand (5%). These are encouraging findings. How, then, can we ensure that all subfields of language development research make similar progress, and that the many "null" findings that are likely to appear as a result of these more stringent research practices are published? The answer, in my view, is to stop basing publication decisions on studies' findings, thereby removing a major incentive to selectively report, HARK or *p*-hack. But how?

One way to do so is via registered reports, whereby studies are reviewed, and accepted in principle, based on their methods and analysis plans, before any data are collected (Chambers, 2013). To their credit, several journals in our field now offer this format. This is an entirely positive development, and we offer the registered-report option too. Indeed, although the format is relatively young, there is already some evidence to suggest that registered reports reduce publication bias quite dramatically. Allen and Mehler (2019) report that around 60% of registered reports in the domains of biomedical and psychological science produce "null" findings, as opposed to around 12% for traditional articles. Focussing on psychology, Scheel, Schijen and Lakens (submitted) find null rates of 56% and just 4% for registered reports and traditional articles respectively. These findings are dramatic, but the very low rates of null findings in traditional articles suggest that registered reports cannot solve the problem of publication bias alone, if journals continue to apply criteria of novelty or importance to articles outside the registered report stream.

A second way to avoid basing publishing decision on studies' findings is by committing to "publish any empirical or theoretical paper that is relevant to the field...and that meets our criteria for rigour, without regard to the perceived novelty or importance of the findings", as set out in *Language Development Research*'s policies and procedures. There already exist several general journals with similar policies – *Royal Society Open Science, Frontiers* and, to some extent, *PLOS ONE* (2020; though "Submissions that replicate or are derivative of existing work will likely be rejected if authors do not provide adequate justification") – but these are general journals that are not necessarily familiar to many language-development researchers. More problematically, all have article processing charges upwards of $1,000, for most article types.

Yet even this commitment may not go far enough. Chris Chambers, a former editor of *PLOS ONE,* notes that, in his experience, "When expert reviewers see null results, they are more likely to go on the hunt for imperfections in the methodology or rationale. This bias is especially insidious because although it is thoroughly results-driven, it requires no explicit reference to the results at all" (Chambers, 2020). The third and final way, then, in which *Language Development Research* strives to avoid basing publishing decision on studies' findings is by offering a results-redacted format. This format allows authors to submit for peer-review articles with no Results or Discussion sections, even if – unlike for registered reports – the data have been analysed and these sections written. Our intention is that this format will allow peer-reviewers and action editors to evaluate papers solely on the basis of their theoretical and empirical rigour, without being unconsciously swayed by the results.

We will not, however, be requiring *all* empirical articles to use either the registered-report or results-redacted format. As Whitaker and Guest (2020) point out, invoking the "buffet model" of Bergmann (2019; as cited in Whitaker & Guest, 2020: 35), "Binging from the many different topics that fall under open scholarship will leave you feeling overwhelmed and exhausted". We take the view, then, that it is better to accept conventional, results-included articles than to force would-be *LDR* authors to "bite off more than they can chew" and risk driving them back to traditional "closed" journals.

Similarly, while we generally require all experimental materials, data and analysis code to be made available in a public repository prior to publication, exemptions will be granted when this is required to ensure participant confidentiality (particularly with hard-to-reach samples or clinical groups), to comply with local laws and regulations, or for copyright reasons (e.g., when researchers use a copyrighted standardized test). While open-science hardliners might take the view that researchers should not rely on data that cannot be legally or feasibly anonymized (e.g., certain video recordings) or use copyrighted tests, we take the "buffet" view: Some open science is better than no open science, and little would be gained by driving such papers to traditional "closed" journals. It is important to note at this point that the policies and procedures summarized here (and approved by our Editorial Board) will be kept under review, and evolve in line with discussions of open science practices both in our field and more generally.

In the meantime, our commitment to publishing any relevant paper that meets our criteria for rigour, though motivated primarily by openness and transparency, brings with it some additional – perhaps unexpected – benefits. First, because we do not screen papers for potential impact, or for their appeal to a wide readership, "relevance to the field of language development (typical and atypical, mono-, bi- and multi-lingual) is broadly construed so as to include, for example, studies of second language learning (or artificial language learning) in older children or adults, studies of nonhuman animals, computational modelling studies, studies or theories of the adult endpoint etc., provided that they are relevant to the issue of language development". Second, for the same reason, we need not impose any restrictions on the types of article that we publish. In addition to registered reports, results-redacted papers and "regular" empirical papers, we will consider literature reviews, systematic-reviews, meta-analyses, papers that present new research or analysis tools, theoretical articles, responses to previous articles, book reviews, and even new types of papers that have yet to be devised. Third, unlike journals that are restricted to a fixed number of issues and pages per year, *Language Development Research* has no need to impose any limits with regard to the number of words, pages or references in a given article.

Fourth, we very much hope that, by not imposing criteria of impact or broad interest, *LDR* will be accessible to, and inclusive of, researchers who study and/or belong to under-represented populations. On the subject of representativeness, I note that while our current team of Action Editors is relatively representative of the field in terms of gender (with 5/7 female researchers), and is not entirely Anglophone (3/7 have a first language that is not English), they are drawn entirely from WEIRD societies (Western, educated, industrialized, rich and democratic; Henrich, Heine & Norenzayan, 2010), specifically the USA, UK and France. As a member of just about every privileged category that exists, all I can say is that I am aware of the issue of representativeness, and will do my best to address it. With regard to inclusivity and accessibility, we have taken some very small steps, by requiring alternative text for figures and allowing abstracts in multiple languages, but we must do more. In the meantime, key to inclusivity and accessibility is our commitment that the journal will always be free of charge to both readers and authors (i.e., “diamond” or “platinum” open access).

How can we survive with no income? Simple: We have no expenditure. The journal runs on the open-source Janeway platform and is hosted for free by Carnegie Mellon University’s Library Publishing Service. For this, we must thank my co-founder Brian MacWhinney, who – via the Child Language Dates Exchange System (https://childes.talkbank.org/) – pioneered Open Science before the term was coined, and who kindly agreed to make *LDR* the official journal of the Talkbank system, which includes the *info-CHILDES* mailing list: the de-facto mailing list for our field. Of course, Carnegie Mellon are bearing some costs; not least the time of Rikk Mulligan, lead of the Library Publishing Service, who put in many hours setting up the journal. But the total cost to Carnegie Mellon can be no greater than a handful of APCs, let alone journal subscriptions.

In my view, then, the model we are adopting for *LDR*, whereby journal hosting costs are borne by universities in lieu of savings elsewhere, is one that can and should be replicated in other fields. After all, via our salaries, our institutions are already funding the writing, reviewing and editing of journal articles; there is no reason for them to baulk at the final financial hurdle of hosting them. We can *do* this. For the good of our field, for the good of science, we *have* to do this.

References

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology, 17*(5), e3000246. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000246

Chambers, C.D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex, 49*, 609-610. 10.1016/j.cortex.2012.12.016 .

Chambers, C. D. (2020). Frontloading selectivity: A third way in scientific publishing?. *PLoS Biology, 18*(3), e3000693. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000693

De Vries, Y. A., Roest, A. M., de Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. *Psychological Medicine*, *48*(15), 2453-2455. 10.1017/S0033291718001873.

Eason, A. E., Hamlin, J. K., & Sommerville, J. A. (2017). A survey of common practices in infancy research: Description of policies, consistency across and within labs, and suggestions for improvements. *Infancy*, *22*(4), 470-491. https://onlinelibrary.wiley.com/doi/abs/10.1111/infa.12183

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences, 33*(2-3), 61-83. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/weirdest-people-in-the-world/BF84F7517D56AFF7B7EB58411A554C17

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://journals.sagepub.com/doi/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. 10.1207/s15327957pspr0203_4.

PLOS ONE (2020). *Criteria for Publication*. Retrieved from https://journals.plos.org/plosone/s/
criteria-for-publication.

Ritchie, S. (2020). *Science Fictions: Exposing fraud, bias, negligence and hype in science.* Bodley Head.

Scheel, A. M., Schijen, M., & Lakens, D. (submitted). An excess of positive results: Comparing the standard psychology literature with registered reports. https://psyarxiv.com/p6e9c/download?format=pdf

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. 10.1177/0956797611417632

Whitaker, K. and Guest, O. (2020). #bropen science is broken science. *The Psychologist*, *33,* 34-37. https://thepsychologist.bps.org.uk/volume-33/november-2020/bropenscience-broken-science

License

Features of lexical richness in children's books: Comparisons with child-directed speech

Nicola Dawson
Yaling Hsiao
Alvin Wei Ming Tan
University of Oxford, UK

Nilanjana Banerji
Oxford University Press, UK

Kate Nation
University of Oxford, UK

Abstract: Access to children's books via shared reading may be a particularly rich source of linguistic input in the early years. To understand how exposure to book language supports children's learning, it is important to identify how book language differs to everyday conversation. We created a picture book corpus from 160 texts commonly read to children aged 0-5 years (around 320,000 words). We first quantified how the language of children's books differs from child-directed speech (compiled from 10 corpora in the CHILDES UK database, around 3.8 million words) on measures of lexical richness (diversity, density, sophistication), part of speech distributions, and structural properties. We also identified the words occurring in children's books that are most uniquely representative of book language. We found that children's book language is lexically denser, more lexically diverse, and comprises a larger proportion of rarer word types compared to child-directed speech. Nouns and adjectives are more common in book language whereas pronouns are more common in child-directed speech. Book words are more structurally complex in relation to both number of phonemes and morphological structure. They are also later acquired, more abstract, and more emotionally arousing than the words more common in child-directed speech. Written language provides unique linguistic input even in the pre-school years, well before children can read for themselves.

Keywords: lexical richness; book language; child-directed speech; language acquisition; literacy

Corresponding author: Nicola Dawson, Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom. Email: nicola.dawson@psy.ox.ac.uk.

ORCID IDs: Nicola Dawson https://orcid.org/0000-0001-7167-6081; Yaling Hsiao https://orcid.org/0000-0003-3986-5178; Alvin Wei Ming Tan https://orcid.org/0000-0001-5551-7507; Kate Nation https://orcid.org/0000-0001-5048-6107

Citation: Dawson, N., Hsiao, Y., Tan, A.W.M., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*, 1(1), p 9-53. DOI: 10.34842/5we1-yk94

Introduction

Children learn from the language they hear (e.g., Cameron-Faulkner et al., 2003; Weisleder & Fernald, 2013). Evidence from longitudinal studies and computational modelling shows that children who experience greater amounts of sophisticated and diverse child-directed talk develop larger vocabularies and better reading skills, and are at an advantage in early school achievement (Chang & Monaghan, 2019; Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 1991, 2010; Jones & Rowland, 2017; Pan et al., 2005; Rowe, 2008, 2012). Yet children's language experiences in the early years vary widely. These differences have been linked to caregiver language competence and socio-economic status (Hart & Risley, 1995; Hoff, 2003; Huttenlocher et al., 2010; Weisleder & Fernald, 2013), but language use may vary within, as well as between, home environments. Shared reading might be a particularly important source of language input, not least because it elicits more complex language and more words per minute from caregivers compared to other contexts, such as mealtimes and play (Demir-Lira et al., 2019; Weizman & Snow, 2001). In this paper we investigate in detail the language of children's books to specify the quantity and nature of lexical input they offer, relative to the language that children encounter via everyday speech.

Corpus analyses consistently demonstrate that written language departs from spoken language in several ways. These differences are well-documented in texts and speech aimed at adults. Overall, written language tends to be more syntactically complex and more lexically diverse than spoken language (Malvern et al., 2004; Roland et al., 2007), although patterns of language use may also reflect other factors, such as formality and genre (Biber, 1993). In part, linguistic differences across modality reflect the decontextualized nature of written language. As spoken language typically takes place in the 'here and now', communication is supported through gesture, facial expression and intonation. Spoken utterances that are incomplete or ambiguous may not pose a barrier to comprehension if meaning is apparent from the communication context. Speech may also be adapted in the moment to rectify breakdowns in communication (Clark, 2020; Healey, de Ruiter, et al., 2018; Healey, Mills, et al., 2018). In the absence of these nonverbal cues and bi-directional dynamics, written language depends more on choice of words and sentence structures to communicate information effectively (Snow, 2010).

Turning to children, books provide exposure to syntactic structures that occur rarely in speech. Montag (2019) showed that even in texts targeted at very young children (i.e. picture books), passive sentences and relative clauses occurred more frequently than in child-directed speech. Similar findings were reported by Cameron-Faulkner and Noble (2013), who found that canonical sentence structures (comprising subject-verb-[object]) and complex sentence constructions (containing two or more lexical verbs) were more frequent in children's books than child-directed speech, whereas questions were more common in speech than in books. Differences also emerge at the lexical level. Montag et al. (2015) calculated type-token ratio curves for a corpus of picture books and a corpus of child-directed speech, revealing that books contained more unique word types than speech at any given sample size. This pattern held true both at the corpus level, and in comparisons between individual books and conversations. Strikingly, even when compared to speech between two adults, children's picture books contain more unique rare word types (Massaro, 2015).

Together, these corpus comparisons suggest that children who frequently participate in shared reading activities are regularly exposed to more advanced linguistic content than children who do not. These differences matter, given that language input is closely tied to language development and that regular access to books in the early years is not universal across children (Hart & Risley, 1995). Identifying and characterising common linguistic properties of children's books is an important starting point for understanding the impact of variation in access to books on children's language development. To this end, we introduce a new children's picture book corpus and identify critical properties of book language, focusing on its lexical content.

Lexical richness broadly refers to the quality of words in a language sample. It encompasses a number of measurable lexical properties, including lexical diversity, lexical density and lexical sophistication (Jarvis, 2013; Malvern et al., 2004; Read, 2000). Lexical diversity provides an indication of vocabulary breadth and is usually measured using type-token ratios (or type-token ratio curves; Montag et al., 2015, 2018). Lexical diversity tells us about the range of words in a text and has been widely adopted as a measure of language quality or proficiency (e.g., Malvern et al., 2004). Measures of lexical density capture the proportion of lexical items (usually defined as nouns, lexical verbs, adjectives and adverbs derived from adjectives) in a language sample relative to the total number of words (Ure, 1971). A higher proportion of lexical items in a language sample is indicative of denser information content compared to a sample with a higher proportion of function words (e.g., prepositions, conjunctions and pronouns). Lexical density is highly correlated with lexical diversity (Johansson, 2008), but conceptually, they measure distinct features. Hypothetically, it is possible for a text to have a high density of lexical items that are repeated frequently, or conversely, a text that uses a diverse range of vocabulary, but includes a high proportion of function words.

Like lexical density, measures of lexical sophistication shed light on the types of words contained within a language sample, and in particular, whether those words are skewed towards one end of the frequency distribution. One approach is to calculate the number of unique word types within a corpus after having accounted for the most frequent word types according to a general language corpus (Massaro, 2015). Adopting this method, Massaro reported that children's picture books contained around three times the number of rare word types of child-directed speech, and around one-and-a-half times the number observed in adult-adult speech. Alternatively, cumulative proportions of word tokens in a given corpus can be plotted against the rank frequency of those words in a general language corpus, providing additional information on the frequency distributions of the most common words across different corpora (Hayes, 1988; Hayes & Ahrens, 1988).

In summary, existing evidence indicates that children's books are more lexically diverse (Montag et al., 2015) and contain a higher proportion of rarer word types (Massaro, 2015) than child-directed speech. This indicates that the language of children's books is disproportionately skewed towards lexical items from the lower end of the frequency distribution. However, little is currently known about the properties of these words and lexical density has not been directly compared across these sources. This matters when we consider that children who are read to less frequently in the early years will gain less exposure to such words. Our aim here is to identify words that are relatively common in children's books, but which appear infrequently in child-directed speech, and to analyse their lexical properties. This will allow us to highlight the types of words that may be particularly impacted by variation in exposure to books in the early years.

Aims and Hypotheses

We created a new children's picture book corpus and selected samples of child-directed speech from the UK CHILDES corpora. This allowed us to fulfil three aims. Our first aim was to replicate Montag et al.'s (2015) analysis of lexical diversity in a new and larger set of children's picture books. Our second aim was to extend cross-modality comparisons to other measures of lexical richness (lexical density and lexical sophistication), along with part of speech distributions, word length, and morphological complexity. Our third aim was to identify the words most uniquely representative of children's books, and to examine how they differ from words more common in child-directed speech in relation to key psycholinguistic properties, namely age of acquisition (the age at which a word is learned; Kuperman et al., 2012), concreteness (the extent to which a word references a perceptible entity, or conversely, how abstract it is; Brysbaert et al., 2014), arousal (the intensity of emotion elicited by a word; Warriner et al., 2013), and valence (how pleasant a word is judged to be; Warriner et al., 2013). If the words most typical of books are more advanced and more abstract than words more common to child-directed speech, then children who regularly participate in shared reading activities will have more opportunity to encode the phonological forms and meanings of such words, and to experience them across diverse contexts. These experiences not only enhance oral vocabulary knowledge (Weizman & Snow, 2001), but also lay the foundations for reading development (Gough & Tunmer, 1986; Perfetti & Hart, 2002), even before children are able to read independently.

Following Montag et al. (2015), we predicted that our set of children's picture books would contain more diverse vocabulary than child-directed speech targeted at a similar age range. We also predicted that books would contain a higher proportion of content words, and more sophisticated vocabulary, relative to speech (Massaro, 2015, 2017). We further anticipated that differences would emerge in structural complexity and part of speech distributions. If the vocabulary of picture books is more sophisticated than that of child-directed speech, we would expect these words to be longer, and for books to contain a higher proportion of morphologically complex words. Given previous comparisons of written and spoken material in adult language samples, we expected differences to emerge in part of speech distributions across children's books and child-directed speech, in particular in the balance of nouns and pronouns (Biber et al., 1998; Hudson, 1994). Finally, we predicted that the words we identified as most representative of 'book language' would have a higher age of acquisition, and would be more abstract, more emotionally arousing, and evoke stronger positive and negative emotions than the words more typical of child-directed speech.

We present our findings in two parts. First, we describe our corpora and the methods used to compare lexical richness across book language and child-directed speech. We then introduce the keyword methodology used to identify words most and least representative of book language before comparing their psycholinguistic properties.

Method

Corpora and Corpus Processing

Picture Book Corpus

The picture book corpus comprised 160 children's fiction books with a total word count of 319,435. These books were purchased for the purposes of this research, and were selected to be representative of the type of reading material children encounter in shared reading contexts in the UK. To this end, we generated an initial list of titles with a target age range of 0-7 years from a combination of retailer bestseller lists and recommendations from literacy charities, book review sites, and teachers. The final list included the titles that were cited most frequently across these sources (see Appendix A for the final selection of book titles; the full corpus can be found at https://osf.io/zta29/). The vast majority of books in the corpus were picture books, but a small number of longer texts that might be read to young children were also included (e.g., The BFG). The content of these books was transcribed as plain text files by undergraduate psychology students. We included text that appeared in illustrations and appendages (for example, text in speech bubbles) in the transcription on the basis that caregivers would likely read these words aloud in addition to the main body of text.

The plain text files containing the transcribed picture books were converted to CHAT Transcription Format (.cha) files so that they could be processed using Computerised Language Analysis (CLAN) software (MacWhinney, 2000). The 'mor' function in CLAN was used to lemmatise and generate part-of-speech tagging for all words within the corpus. The output .cha files were then converted to XML and parsed using the XML package in R (R Developement Core Team, 2017), with the data outputted to .csv files, which were used in subsequent analyses.

Spoken Language Corpus

This was generated from 10 corpora from the English-UK section of the CHILDES database (MacWhinney, 2000). The sample comprised all suitable corpora from this collection, with the exception of those that focused on specific populations (e.g., children with language impairments). The final set of 10 corpora (see Appendix B; the full set of corpora are accessible via the link above) contained transcripts of interactions between 190 different children aged 6 weeks to 6 years and their caregivers, siblings, other family members and research investigators. Recordings took place across a variety of contexts, but typically involved structured and free play activities between children and their caregivers, as well as everyday routines such as mealtimes and bedtimes. Across all recordings, utterances produced by the child were filtered out, such that the final dataset contained only talk directed to the child for a total word count of 3,853,976. The CHILDES corpora were downloaded in CHAT format and had already been processed using CLAN. As above, these files were converted to XML and parsed using R, with data outputted to .csv files in the same format as the picture book corpus.

Procedure

(i) Corpus Comparisons

Lexical Diversity. Following Montag et al. (2015), we calculated type-token ratio curves to show the number of unique word types in each corpus at various token sample sizes. We took this approach because type-token ratios decrease as the number of tokens in a sample increases: the more words there are in a language sample, the more likely it is that words will be repeated (Montag et al., 2018). Because our spoken language corpus is considerably larger than our picture book corpus, it was not possible to compare the two corpora on a single measure of lexical diversity. We adopted Montag et al.'s (2015) method of calculating type-token ratios for multiple random samples from each corpus, ranging from 100 to 50,000 words in size, and increasing in increments of 100 words each time. One hundred simulations were generated at each sample size, each based on a new random sample, and type-token ratios were calculated as the mean type count across the 100 simulations divided by the sample size.

Lexical Density. Each lemma token was coded as 'lexical' or 'non-lexical'. Lexical lemmas were defined as nouns (excluding proper nouns and pronouns), adjectives, verbs (excluding modal verbs, such as 'do', 'will', 'can', 'must', 'shall', 'may', and auxiliary verbs 'be', 'have' and 'get') and adverbs derived from adjectives (e.g., 'fast' and 'happily'). All other tokens were coded as 'non-lexical'. We calculated lexical density by dividing the number of lexical items by the total number of lemmas in each individual text or conversation (Berman & Nir, 2010; Strömqvist et al., 2002).

Lexical Sophistication. Following Hayes (1988; also Hayes & Ahrens, 1988), we generated cumulative frequency curves showing the proportion of each corpus accounted for by the 1,000 most common words in English. We decided to use the SUBTLEX-UK database as our reference, which lists frequencies for around 160,000 words generated from subtitles of British television programmes (van Heuven et al., 2014). We chose this as our reference database for two reasons. Firstly, these frequencies have been shown to explain 4% more variance in word processing times than other large general language corpora (e.g., the British National Corpus; van Heuven et al., 2014). Secondly, we reasoned that television subtitles represent a hybrid between written and spoken language as they typically record scripted speech, and therefore this approach would not be biased towards one modality over the other.

Our analysis was based on the cleaned version of the SUBTLEX-UK frequency list, with digits and non-alphanumerical symbols removed. We further eliminated all proper nouns from the list, and then ranked the list by token frequency across all broadcasts and selected the top 1,000 words. We calculated the cumulative proportion of tokens in the picture book and spoken language corpora accounted for by the 1,000 most common words in the reference list. We noted some inconsistencies in the tokenised forms of words between the SUBTLEX list and our corpora processed by CLAN (for example, contracted forms such as *n't* in the word *wasn't* was listed as a token in the SUBTLEX list, but not in our corpora). This meant that a number of items in the 1,000 most common words returned a frequency of 0 or a very low frequency in the picture book and spoken corpora. Therefore, we checked all entries in the SUBTLEX list that occurred with 0 frequency in either corpus to ensure that this was truly due to non-occurrence, and not inconsistency in tokenisation. In the case of inconsistency, we manually corrected the relevant entries in our corpora to align with the tokenised form in the SUBTLEX list. Finally, we plotted cumulative frequencies as a proportion of total corpus size against rank order of the 1,000 most common words.

Part of Speech. The automatic part of speech tags generated by CLAN were combined into broad lexical categories. For example, CLAN provides a unique tag for each different type of pronoun: these were reclassified for the purposes of our analysis as 'pronouns'. Our focus was on the major parts of speech, including nouns, lexical verbs, adjectives, adverbs, pronouns and determiners. All other tags, including modal and auxiliary verbs, proper nouns and communicators (e.g., 'ah') were coded as miscellaneous.

Structural Properties. We calculated word length in number of phonemes using the Carnegie Mellon Pronouncing Dictionary (Carnegie Mellon University, 2014) as the reference database. Data on number of phonemes were available for 84% words in the picture book corpus and 79% of the words in the spoken language corpus. We also recorded the morphological structure of the words in each text or conversation. We calculated the percentage of morphologically complex lemmas in each text or conversation (i.e. ignoring inflected word forms), and recorded whether complex words were derivations (e.g., *teacher*), compounds (e.g., *football*) or words that were formed through both compound and derivational processes (e.g., *footballer*). Our coding of morphological structure was based on information available in the MorphoLex (Sánchez-Gutiérrez et al., 2018) and MorphoQuantics (Laws & Ryder, 2014) databases. Lemmatised forms output by CLAN were checked for errors (e.g., stems that comprised only one segment of a compound – *foot* instead of *football*) and inconsistencies with lemmatised forms in our morphology reference databases (for example, we included inflectional suffixes in the lemmatised form of nouns derived from verbs – *the writing on the page* – or participle adjectives, such as *the painted bench*). Any identified errors or inconsistencies were manually corrected. Morphological information was available for 97% of the words in the picture book corpus and 95% of the words in the spoken language corpus.

(ii) Keyword Analysis

We followed the method outlined by Kilgarriff (2009; see also Kilgarriff, 2001) to identify the words most representative of the picture book corpus. We started by filtering out tokens tagged as proper nouns or letters, and we also removed tokens with missing part of speech information. We then mapped the remaining tokens to the list of corrected lemmas used in the analysis of morphology, with the exception that inflectional suffixes (-*ed* and -*ing*) were removed to align with lemmatised forms in the age of acquisition, concreteness and affective ratings (see below).

Taking the picture book corpus as the focus corpus, and the spoken language corpus as the reference corpus, we calculated a keyness score for each word that appeared in the former. The keyness score for a given word is the ratio of normalised frequency in the focus corpus to normalised frequency in the reference corpus. We used average reduced frequencies in place of raw frequencies to account for the dispersion of a word across the corpus. This is an adjusted frequency measure which is based on the distances between consecutive occurrences of a given word in a corpus (Hlaváčová, 2006; Savický & Hlaváčová, 2002). This approach addresses the issue of 'burstiness': words that occur with high concentration within a small section of a corpus (e.g., within the same document), but sparsely elsewhere. Two words with the same raw frequency may differ on average reduced frequency if one is more evenly distributed across the corpus than the other. For a word that is completely evenly distributed, the average reduced frequency will be equivalent to the raw frequency.

A keyness score of 1 means that a word appears with equal frequency (per million) in each corpus, whereas a score greater than 1 indicates that the word occurs more frequently in the focus corpus than the reference corpus, and a score below 1 indicates that the word occurs less frequently in the focus corpus than the reference corpus. Given the problem of calculating ratios for words occurring in the focus corpus, but not at all in the reference corpus, we added a constant of 10 to all normalised frequencies before calculating keyness. We selected this value as the constant because it focuses the keyword analysis on the lower end of the frequency spectrum (Kilgarriff, 2009), which we considered to be important when identifying the words that children were unlikely to encounter in everyday conversation, but which they would experience through regular exposure to book language. We have included output from additional keyword analyses in Supplementary Materials (available on the OSF project page https://osf.io/zta29/) which focus on keywords in higher frequency ranges.

Once we had generated a keyness score for each item in the picture book corpus, we ranked them and selected the 500 words with highest keyness scores (i.e. the words most representative of the book language corpus; hereafter ‘book+ words’), and the 500 words with the lowest keyness scores (the words least representative of books; hereafter ‘book— words’). We chose to focus on 500 words from each end of the spectrum as this was approximately the largest sample for which all words in the book— set had a keyness score of less than 1, indicating that they occurred with greater relative frequency in the spoken language corpus compared to the picture book corpus. See Appendix C for a reduced list of the 50 book+ and 50 book— words with the most extreme keyness scores.

We then compared the two sets of words on a number of psycholinguistic properties to examine what characterises the words that children experience through book language, and how they differ to words more typical of child-directed speech.

Age of acquisition. We analysed the age at which our two sets of words are typically acquired using ratings from Kuperman et al. (2012). These norms are generated by asking adults to rate the age at which they think they learned a word, with lower ratings indicating that a word is acquired earlier in development.

Concreteness. This was based on ratings from adults (Brysbaert et al., 2014), where participants were asked to rate the extent to which a word refers to something perceptible (i.e. something that can be directly experienced via any of the five senses), or conversely, the extent to which a word’s meaning is defined using other words. Ratings range from 1 for words that are highly abstract (e.g., *would*) to 5 for words that are highly concrete (e.g., *apple*).

Arousal. We examined emotional arousal using norms from Warriner et al. (2013). Participants in this study were asked to rate the intensity of emotion elicited by a given word, ranging from 1 for ‘calm’ (e.g., *librarian*) to 9 for ‘excited’ (e.g., *insanity*).

Valence. Our valence ratings were also taken from Warriner et al. (2013). These ratings indicate the extent to which a word evokes positive or negative emotions, and also range from 1-9 where 1 represents 'sad' (e.g., *murder*), and 9 'happy' (e.g., *sunshine*). Because our hypothesis relates to the extremity of valence ratings, rather than the direction of the effect, we transformed the mean valence rating for each word by centring it at the midpoint of the scale (i.e. 5, representing a neutral response), and calculating deviation from that point irrespective of direction. For example, a mean rating of 5 was allocated a score of 0, and mean ratings of 4 and 6 were each scored as 1.

Results

(i) Corpus Comparisons

Lexical Diversity

The mean number of word types at each sample size for the picture book and spoken language corpora are presented in Figure 1. The data show that, at any given sample size, the picture book corpus contains a greater number of unique word types than the spoken language corpus. Differences also emerge in the slopes of the lines. The picture book corpus shows a steeper type-token ratio curve compared to the spoken language corpus, indicating a greater increase in unique word types per unit increase in word tokens.

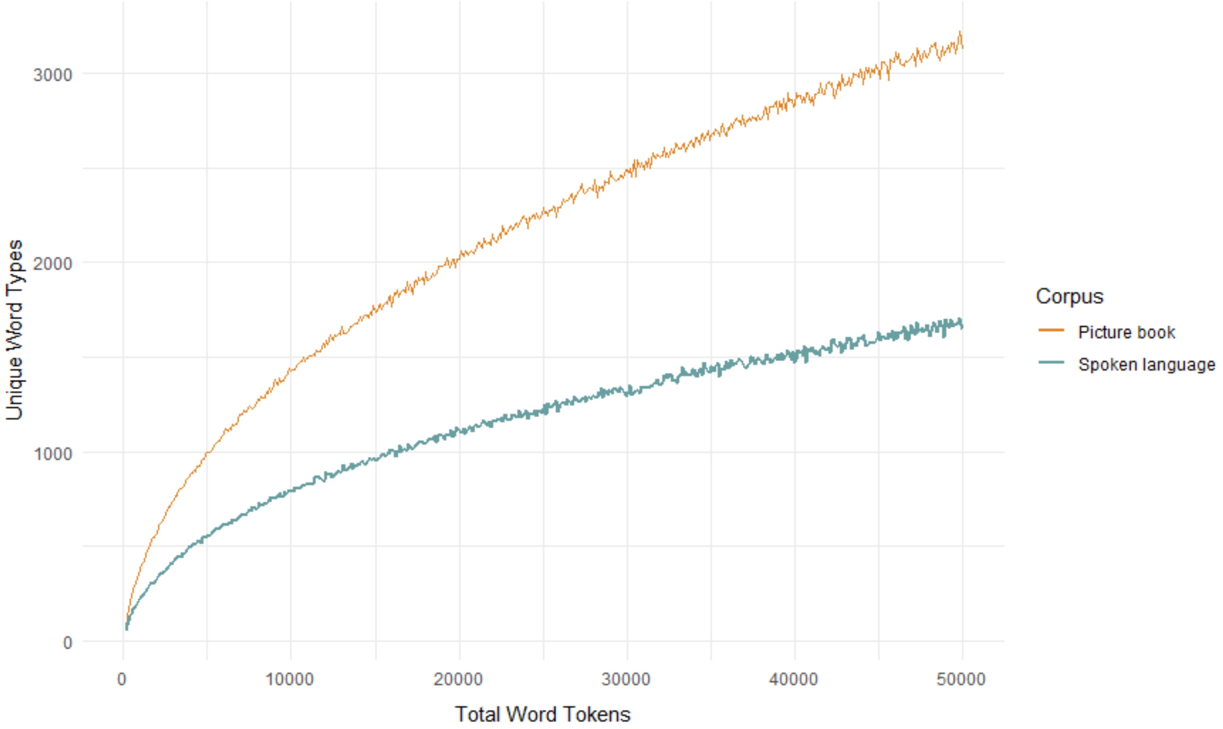


Figure 1. *Mean number of word types at different sized samples of word tokens randomly selected from the picture book and spoken language corpora*

Lexical Density

Figure 2 plots percentage lexical density for each individual text in the picture book corpus (n = 160), and each contiguous sample of child-directed speech in the spoken language corpus (n = 1616). The picture books contain a significantly higher percentage of content words (M = 43.77; SD = 7.00) compared to samples of child-directed speech (M = 28.56; SD = 2.65): t(163.55) = 27.29, p < .0001.

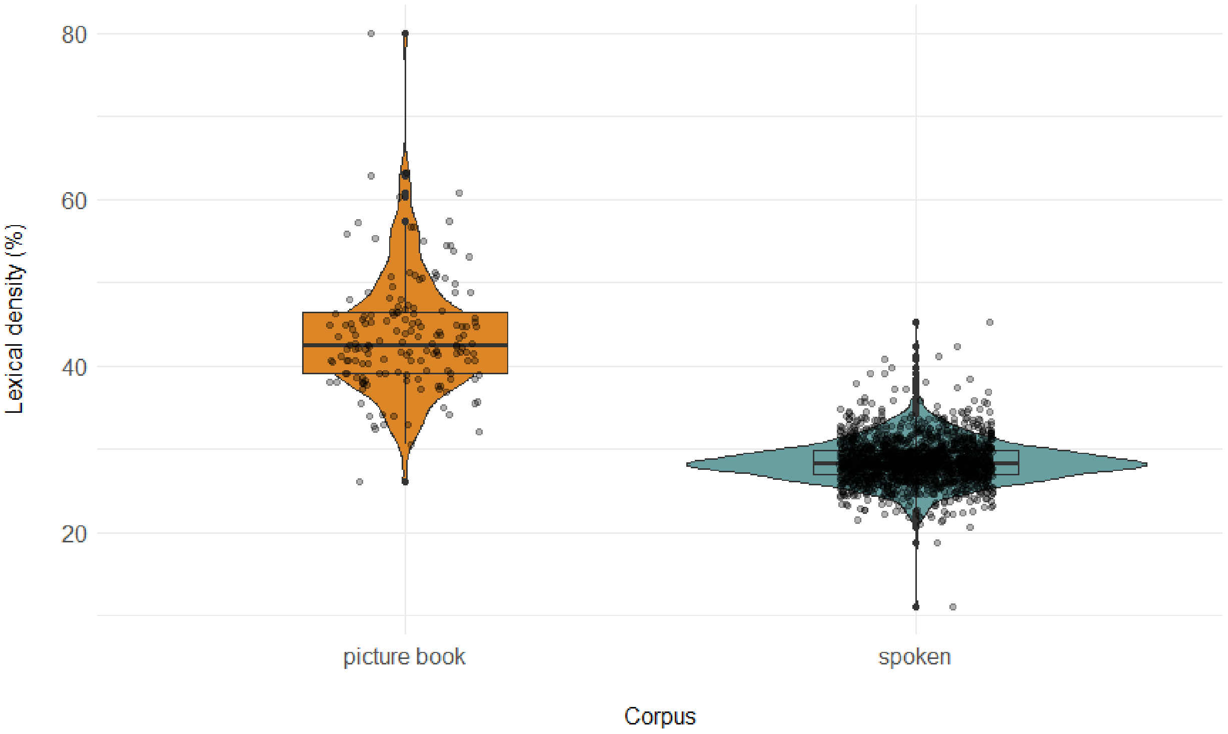


Figure 2. ***Percentage lexical density across picture book and spoken language corpora, plotted by individual document (picture book corpus) and conversation (spoken language corpus)***

We then examined whether lexical density varies by text genre. Specifically, we compared lexical density in texts written in a narrative style to those written in rhyme. It might be that rhyming texts would be more lexically dense than narrative texts, given the focus on imagery, rhythm and phonological properties of words. Texts adopting a partial rhyming structure were included in the 'rhyme' category, provided they were clearly written in verse. However, texts that were predominantly written in prose (e.g., a text comprising a collection of stories which included one story written in verse) were categorised as 'narrative'. Analysis revealed that percentage lexical density was indeed significantly greater in the rhyming texts (n = 62; M = 47.32; SD = 8.24) compared to the narrative texts (n = 98; M = 41.52; SD = 4.95): t(89.03) = -4.99; p < .0001 – see Figure 3). Inspection of the data distributions indicated an outlier in the set of rhyming texts with a lexical density score of 80%. We reanalysed the data without this outlier, but this did not alter the outcome. Note that while lexical density was greater in the rhyming texts, narrative texts (*M* = 41.52; *SD* = 4.95) were still more dense than child-directed speech (*M* = 28.56; *SD* = 2.65).

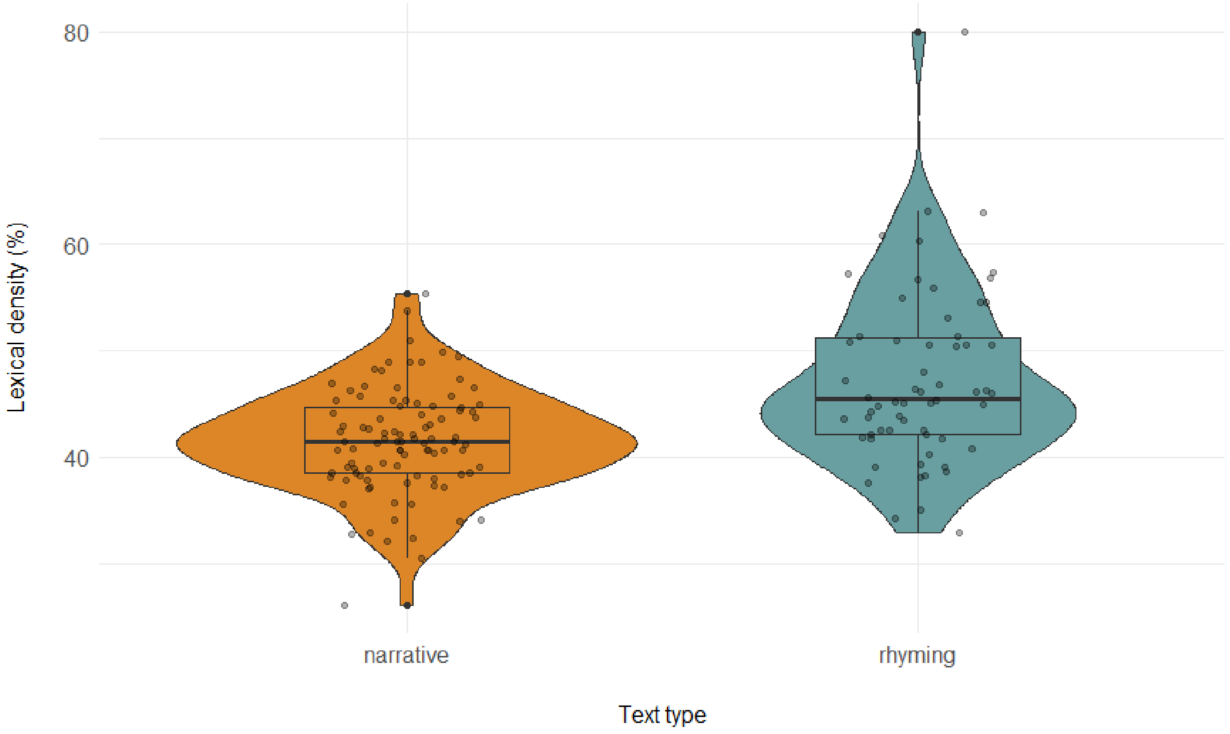


Figure 3. ***Lexical density by text type***

Finally, we examined whether differences in lexical density across the book and spoken language corpora were driven by a proportionate increase across all lexical word classes, or a higher concentration of words from a particular word class. To do this, we calculated the frequency of nouns, verbs, adjectives and adverbs as a percentage of total lexical items in each corpus (Figure 4). If greater lexical density in the picture book corpus is equally distributed across word class, then there should be little difference across corpora in the frequency of each part of speech as a proportion of total lexical items. However, Figure 4 indicates a greater relative proportion of nouns and adjectives in the picture book corpus, and a lower proportion of verbs.

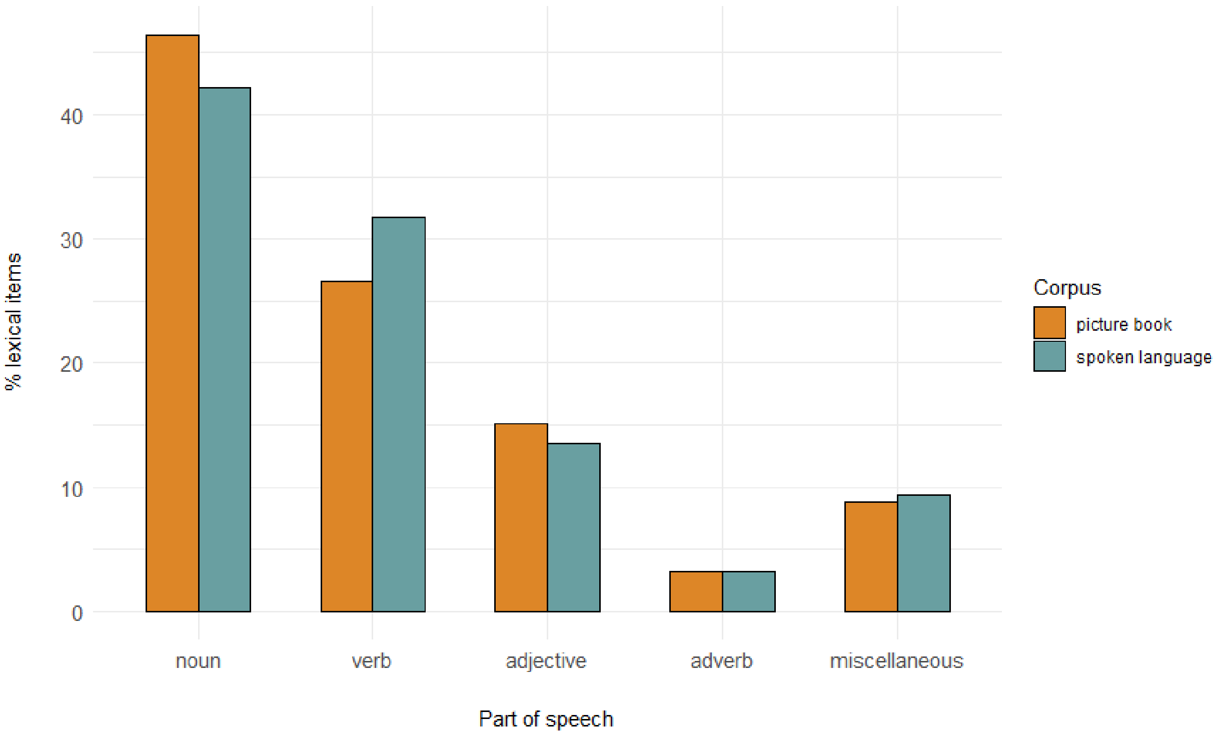


Figure 4. *Frequency of part of speech tags as a percentage of total content words in the picture book and spoken language corpora*

Lexical Sophistication

Figure 5 plots the cumulative proportion of total tokens in each corpus accounted for by the 1,000 most common words in English (with SUBTLEX-UK television subtitles as the reference database), ranked in order of frequency on the log10 scale. The intercept at the left y-axis shows the proportion of each corpus accounted for the most common word according to SUBTLEX frequencies (*the*): 5% of the picture book corpus, and 3% of the spoken corpus. The point at which the curve intersects the right y-axis shows the proportion of each corpus accounted for by the 1,000 most common words: 72% of the picture book corpus, and 79% of the spoken corpus. The curves show that the words in picture books and child-directed speech are differently distributed along the frequency spectrum. A higher proportion of words in child-directed speech are among the most common words in the language overall, whereas picture books contain a higher proportion of words that fall outside this set. Therefore, access to picture books increases the likelihood that children will experience rarer word types that they would not otherwise encounter through conversation alone.

The curves also reveal an interesting pattern about the distributions of the most common words across the two modalities. As expected, the 1,000 most frequent words accounted for a larger proportion of total tokens in the spoken language corpus compared to the book corpus, yet the most common words account for a higher proportion of words in the picture book corpus. Closer inspection of the top 10 words revealed that this effect was primarily driven by a higher proportion of articles (*the*, *a*) and conjunctions (*and*) in the book corpus, whereas the proportion of pronouns (*you*) and demonstratives (*that*) was greater in the spoken language corpus. We examine part of speech distributions in more detail next.

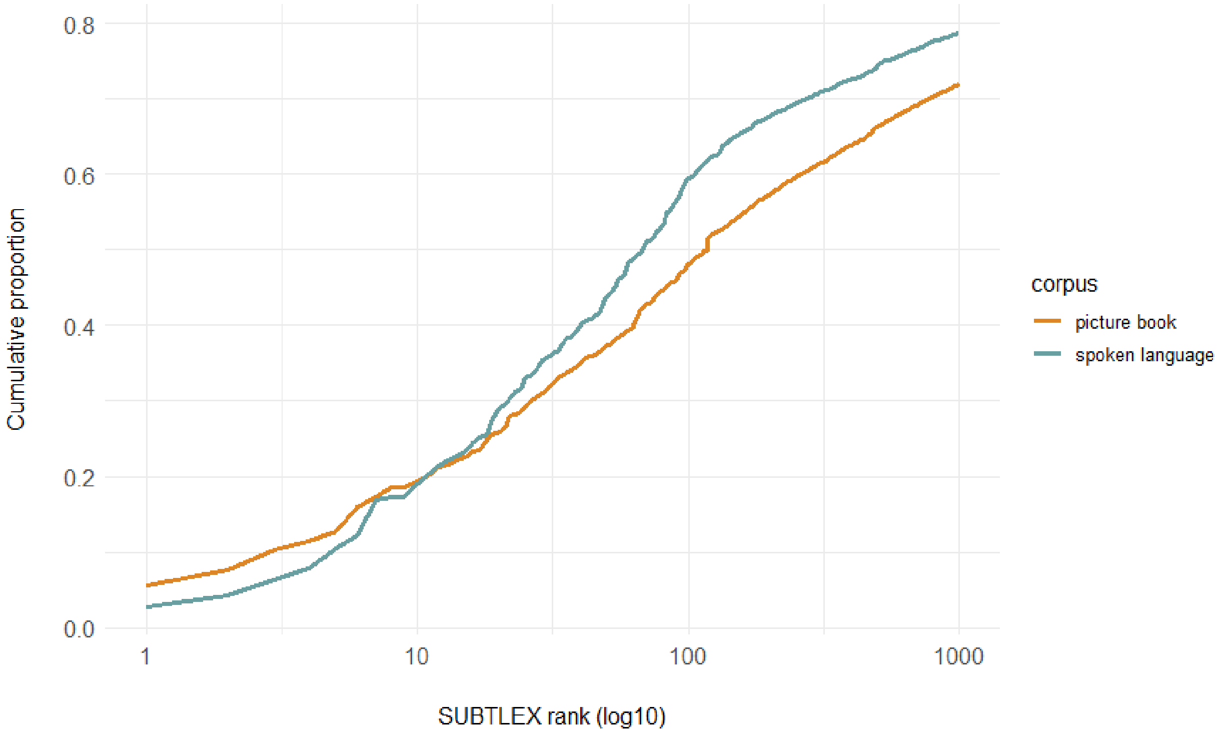


Figure 5. ***Cumulative proportions of total tokens plotted against rank of 1,000 most common words***

Part of Speech Distributions

Figure 6 shows frequency of occurrence (per million words) of each of the major lexical categories across the two corpora. Adjectives, conjunctions and coordinators, determiners, nouns, and prepositions all occurred with greater relative frequency in the picture book corpus compared to the spoken language corpus. Only pronouns were more frequent in spoken language, along with items classed as ‘miscellaneous’, which included proper nouns, auxiliary and modal verbs, and communicators (e.g., *ah*).

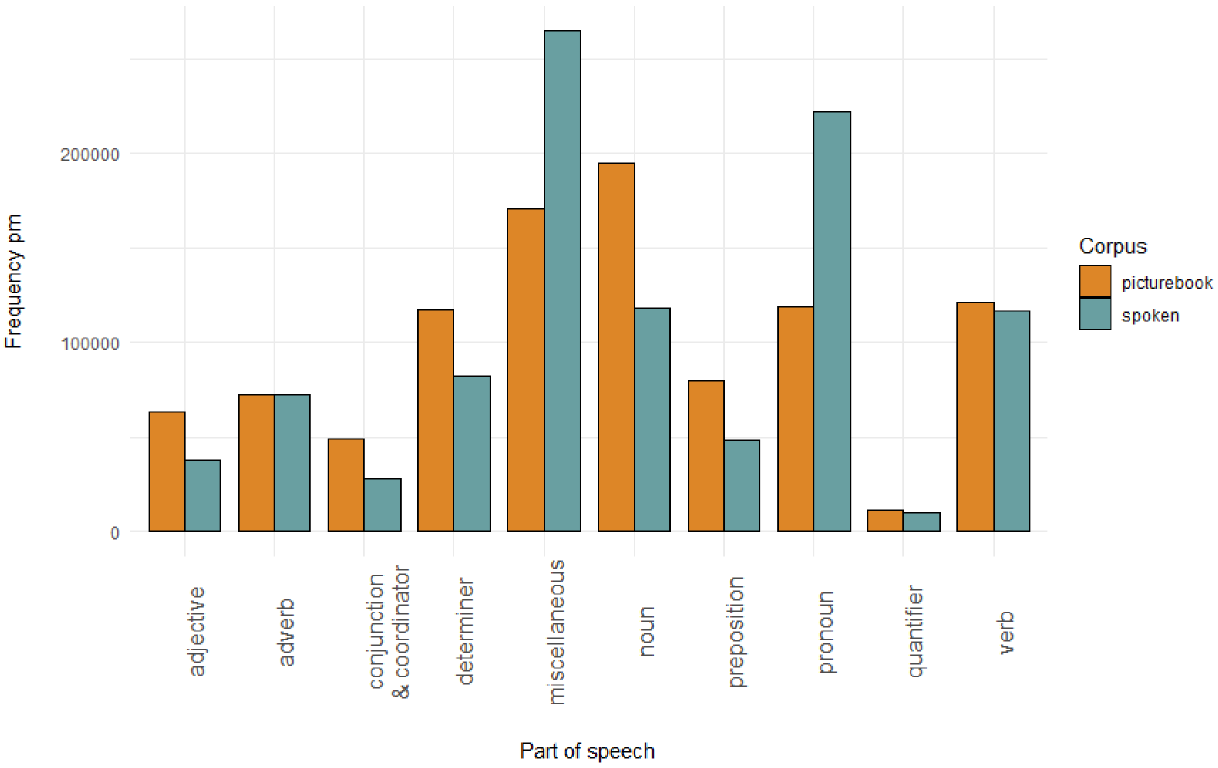


Figure 6. *Part of speech distributions (frequency per million words) across picture book and spoken language corpora*

We conducted further analyses to examine the distributions of different types of pronoun and determiner across the two corpora. Figure 7 indicates that differences in pronoun frequency across picture books and child-directed speech are driven mostly by the large number of personal (*you*), demonstrative (*this*), and interrogative (*what*) pronouns in speech relative to books. While determiners are more frequent overall in books compared to speech, this is particularly the case for articles (*the*) and possessives (*her*), whereas demonstrative determiners (*these*), just as demonstrative pronouns, show the opposite trend.

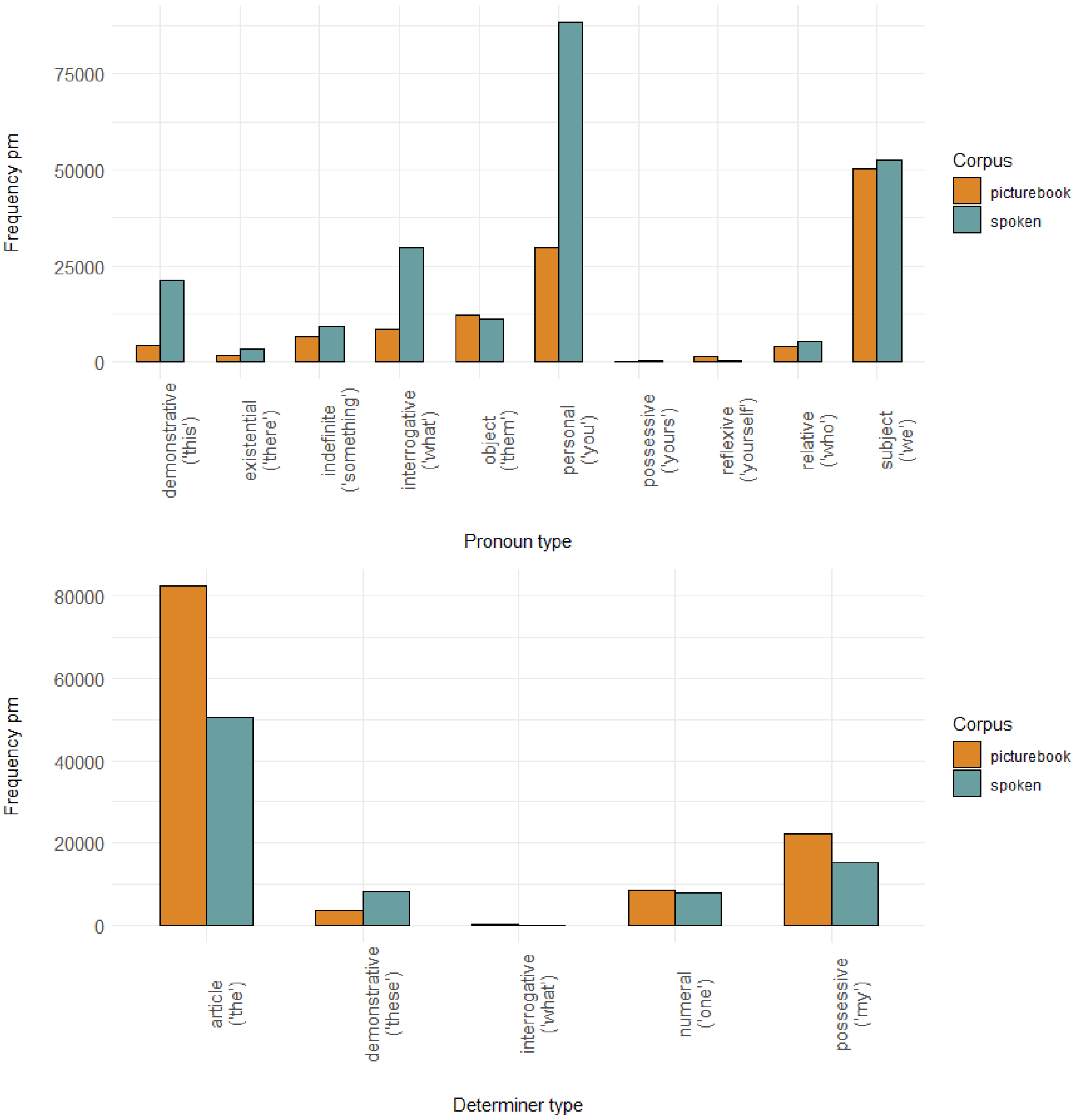


Figure 7. *Pronoun (upper panel) and determiner (lower panel) distributions across picture book and spoken language corpora with examples from each category*

Word Length

Figure 8 shows phoneme count distributions across corpora. We set a maximum cut-off of 10 phonemes for the purposes of plotting the data, given the very small proportion of words that exceeded these values. The distributions indicate a higher proportion of longer words (four or more phonemes) in the picture book corpus, and a higher proportion of shorter words (three or fewer phonemes) in the spoken language corpus.

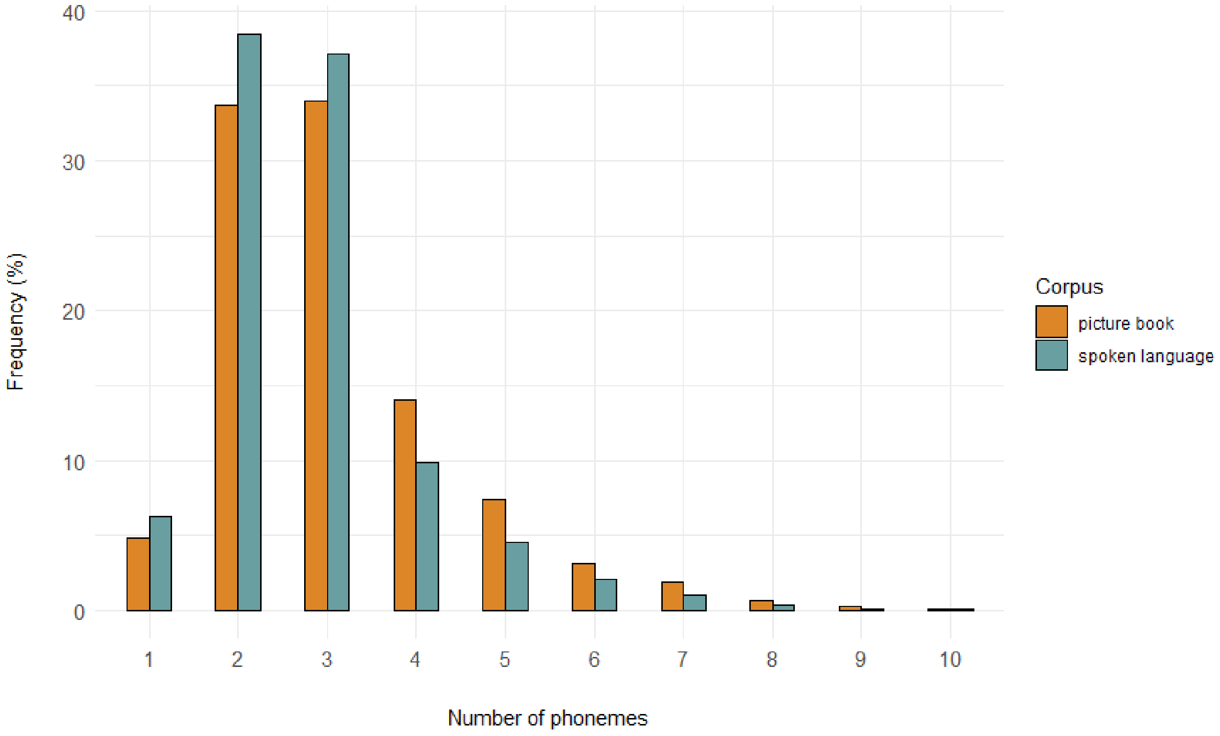


Figure 8. ***Phoneme count distributions across picture book and spoken language corpora***

Morphological Complexity

For each text or conversation, we calculated the percentage of morphologically complex lemma tokens (plotted in Figure 9). Plotting the full dataset indicated a number of outlier texts and conversations containing a high proportion of morphologically complex words (these were typically very short language samples). These were removed by excluding any individual text or conversation that exceeded three standard deviations from the mean for that corpus (corresponding to 0.63% of the texts in the picture book corpus and 0.43% of the conversations in the spoken language corpus). Removing these outliers did not alter the pattern of findings. Welch's Two Sample T-test confirmed that texts in the picture book corpus (*M* = 6.61; *SD* = 3.19) contained a significantly higher percentage of morphologically complex words than conversations in the spoken language corpus (*M* = 4.31;

$SD = 1.09$: $t(161.68) = 9.03$, $p < .0001$).

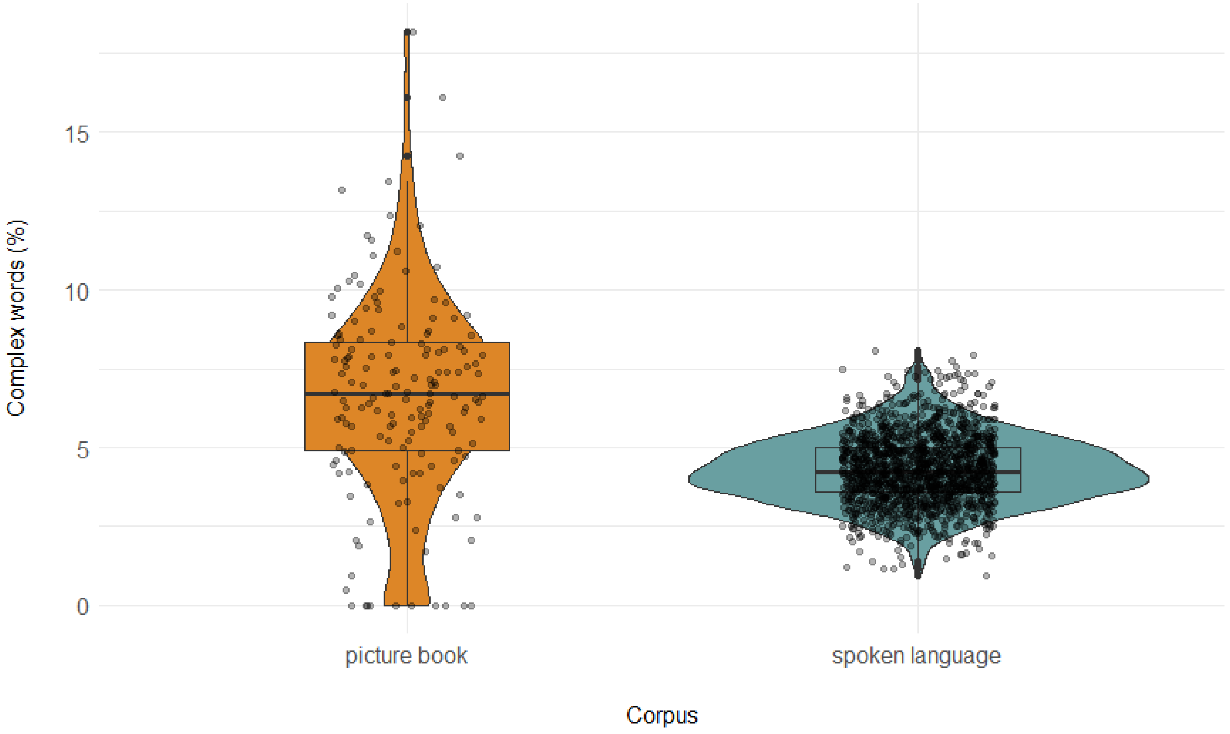


Figure 9. ***Percentage of words in each text (picture book corpus) or conversation (spoken language corpus) comprising two or more morphemes***

To further explore the composition of morphologically complex words across the picture book and spoken language corpora, we calculated the percentage of complex words accounted for by derivations and compounds. Figure 10 indicates that most morphologically complex words across the two corpora were derivations (e.g., *teacher*), followed by compounds (e.g., *football*), whereas derived compounds (e.g., *footballer*) were comparatively rare. The relative contribution of each word type to overall morphological complexity was very similar across the picture books and child-directed speech.

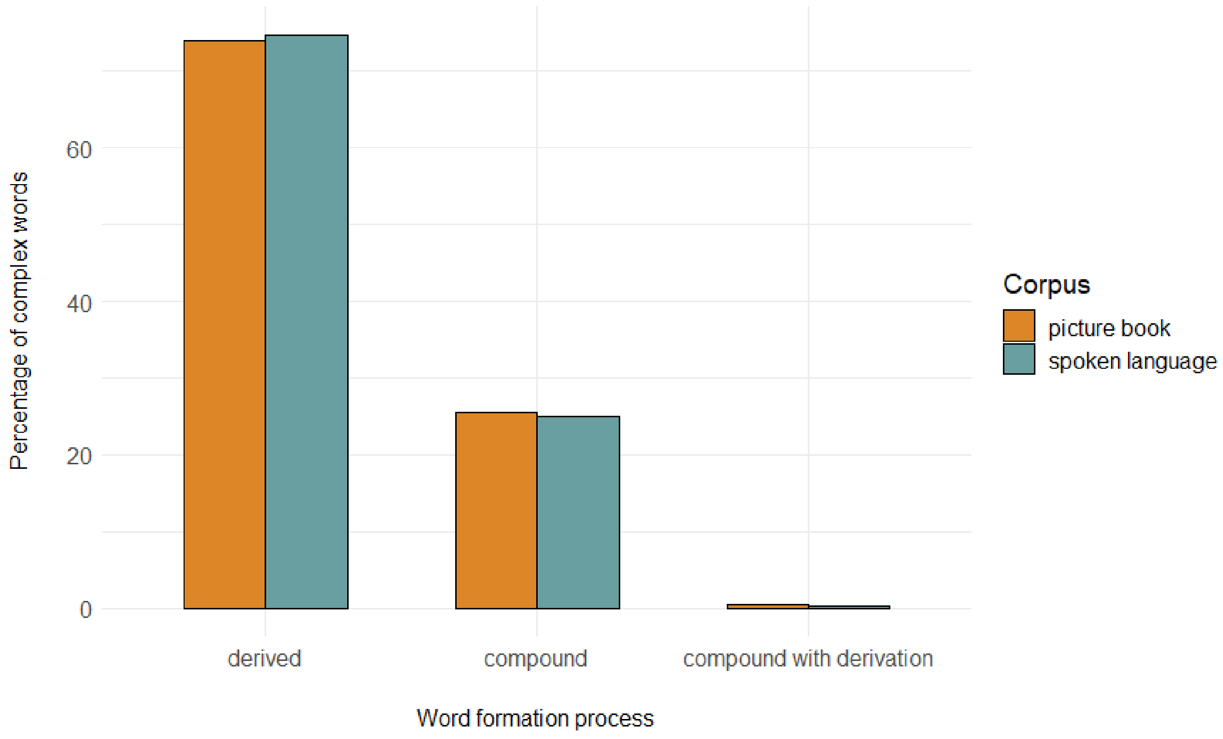


Figure 10. *Percentage of total complex words in each corpus classed as derived, compound, and compounds with derivation*

(ii) Keyword Analysis

Age of Acquisition

Age of acquisition ratings were available for 462 of the 500 book+ words (*M* keyness score = 4.84, *SD* = 2.04), and 451 of the book– words (*M* keyness score = 0.65, *SD* = 0.21). Figure 11 shows distributions, box plots and data points for age of acquisition ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 6.17; *SD* = 1.57) had a significantly higher mean age of acquisition rating than the book– words (*M* = 5.38; *SD* = 1.77): *t*(892.63) = 7.11, $p <$ .0001).

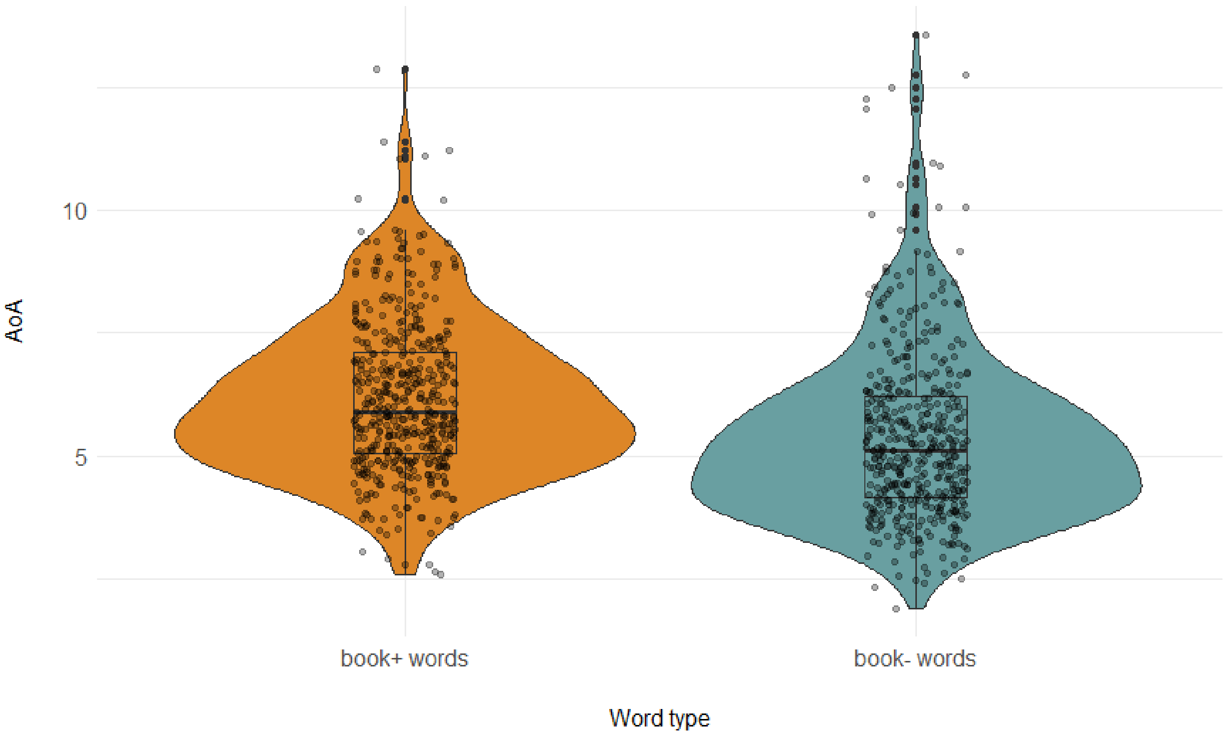


Figure 11. *Age of acquisition ratings for the 500 words with the highest (book+) and lowest (book–) keyness scores*

Concreteness

Concreteness ratings were available for 491 of the book+ words (*M* keyness score = 4.82, *SD* = 2.00), and 469 of the book– words (*M* keyness score = 0.64, *SD* = 0.22). Figure 12 shows distributions, box plots and data points for concreteness ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 3.27; *SD* = 0.98) are lower in concreteness than the book– words (*M* = 3.77; *SD* = 1.20): *t*(901.59) = -6.99, *p* < .0001).

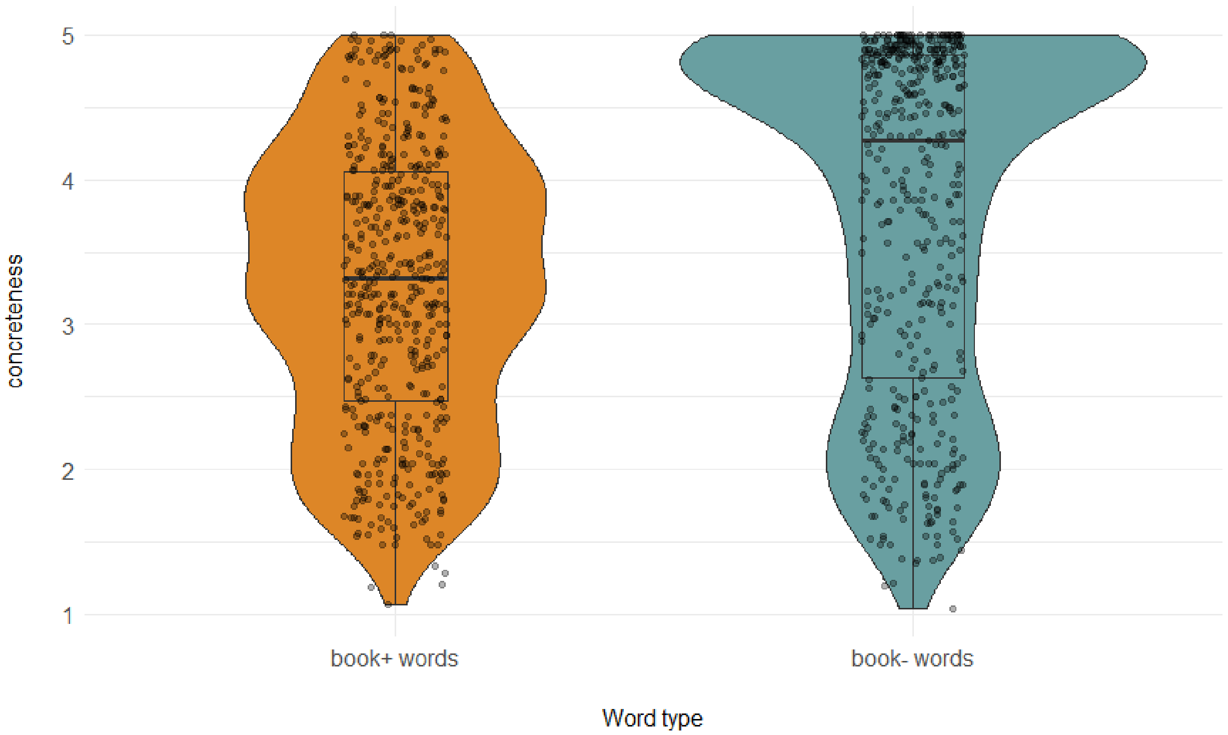


Figure 12. *Concreteness ratings (max = 5) for the 500 words with the highest (book+) and lowest (book–) keyness scores*

Arousal

Arousal ratings were available for 389 of the book+ words (*M* keyness score = 4.82, *SD* = 2.06), and 365 of the book– words (*M* keyness score = 0.67, *SD* = 0.20). Figure 13 shows distributions, box plots and data points for arousal ratings for each set of words. Welch's Two Sample T-test indicated that the book+ words (*M* = 4.30; *SD* = 0.98) had a significantly higher arousal rating than the book– words (*M* = 3.98; *SD* = 0.83): *t*(743.75) = 4.78, *p* < .0001).

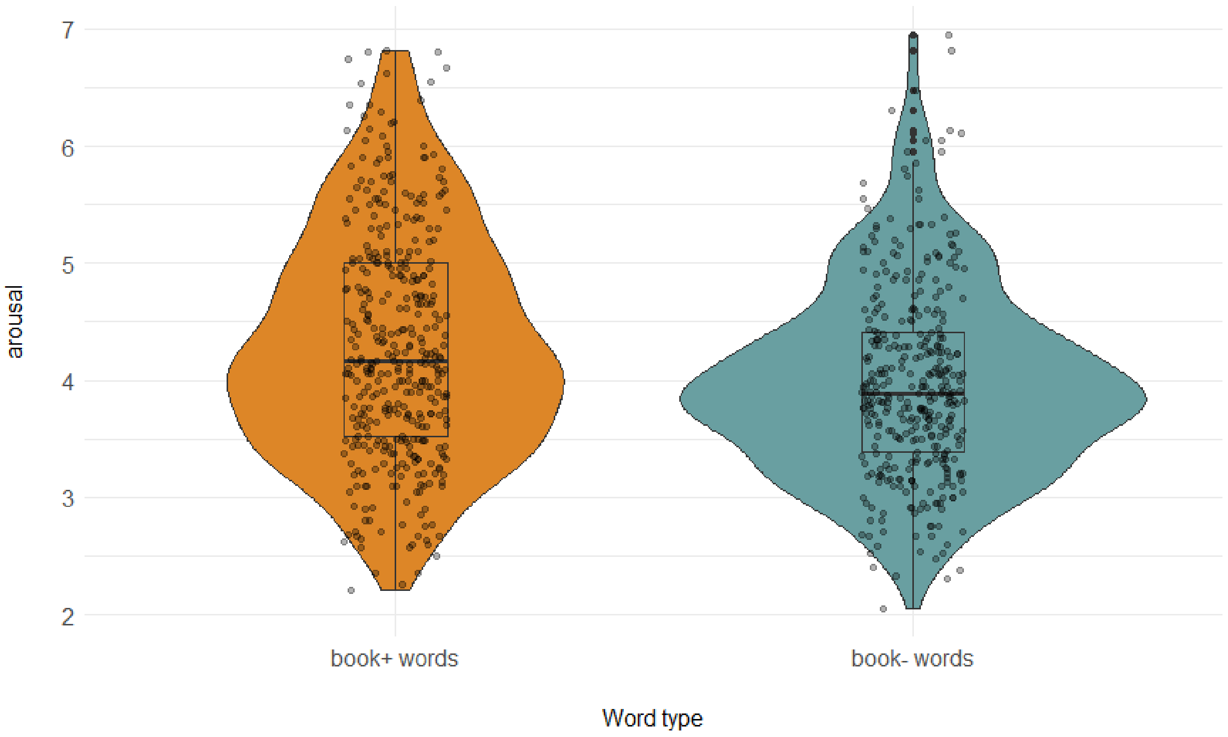


Figure 13. *Arousal ratings (max = 9) for the 500 words with the highest (book+) and lowest (book–) keyness scores*

Valence

Valence ratings were available for the same words included in the analysis of arousal. Figure 14 shows distributions, box plots and data points for centred valence ratings for each set of words. Welch's Two Sample T-test indicated that there was no significant difference in the extremity of valence ratings between book+ words ($M = 1.21$; $SD = 0.82$) and book– words ($M = 1.15$; $SD = 0.70$): $t(745.79) = 1.04$, $p = 0.297$).

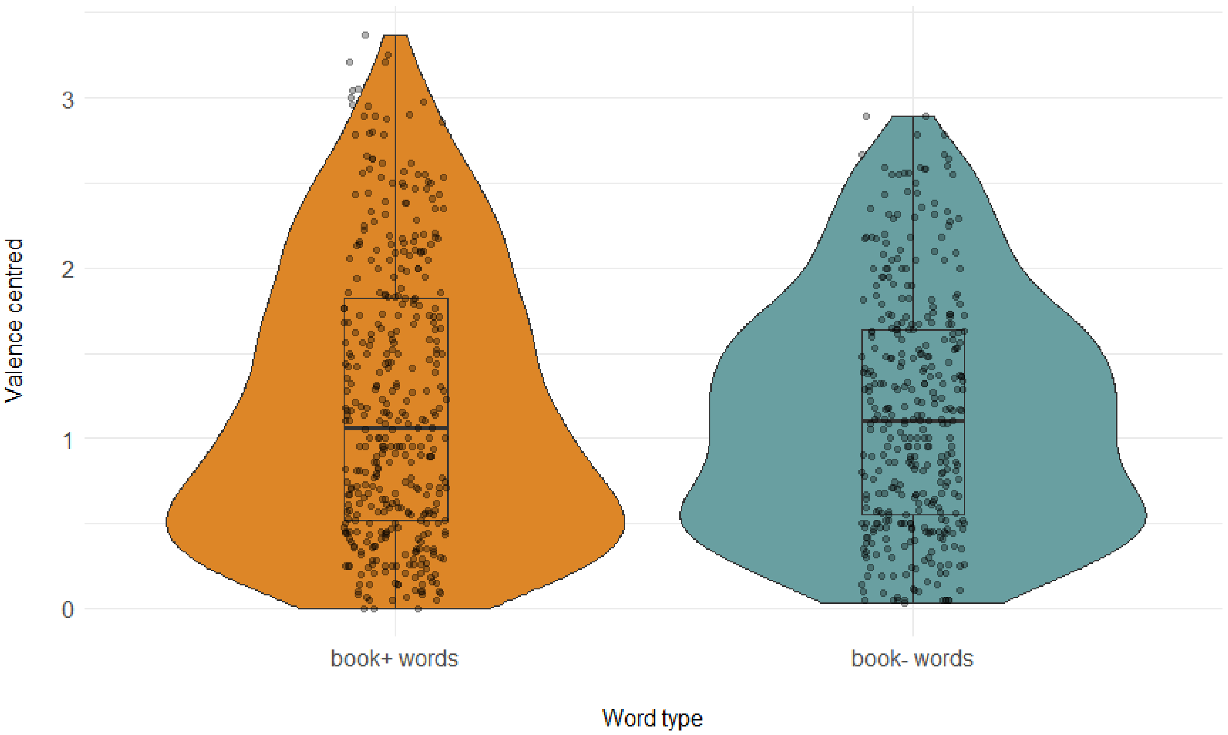


Figure 14. *Centred (from the point of neutrality, see Method) valence ratings for the 500 words with the highest (book+) and lowest (book–) keyness scores*

Discussion

Our aim was to both replicate and build on previous work documenting differences in lexical richness across children's books and child-directed speech (Hayes, 1988; Massaro, 2015; Montag et al., 2015). In line with previous findings, we found that the words used in children's books are typically more diverse, more sophisticated, and lexically denser than those children hear via conversation. We extended these analyses by documenting the structural and lexical properties of these words. We found differences in part of speech distributions, with adjectives and nouns occurring more frequently in books, and pronouns more frequently in child-directed speech. The words in children's books were typically longer and were more likely to be morphologically complex, although the proportion of complex words that were formed through derivation or compounding was similar across the two corpora. Finally, we identified the words most representative of the books in our sample and found these had a higher age of acquisition, were more abstract, and rated higher in arousal than words more common to child-directed speech. We discuss each of these findings in turn and consider the implications for children's exposure to book language and language learning.

Following Montag et al. (2015), we compared lexical diversity in the picture book and spoken language corpora using type-token ratio curves. Our calculations were based on a different sample of child-directed speech, and a new and larger corpus of children's books, yet our analyses clearly replicated their finding that picture books contain a greater number of unique word types than the spoken language corpus at any given sample size. Further, the curves representing type-token ratios showed a steeper trajectory for book language relative to spoken language. This suggests that increasing the amount of book language that children hear has a bigger impact on the number of unique words they are exposed to than an equivalent increase in child-directed speech. Diversity in the linguistic input is considered key to language learning (e.g., Johns et al., 2016). More specifically, some research suggests that lexical diversity in child-directed speech predicts children's vocabulary development over and above the quantity of language they hear (Hsu et al., 2017; Rowe, 2012), a finding backed by computational modelling separating the effects of quantity and diversity (Jones & Rowland, 2017). While caregiver talk may involve frequent repetitions of words and phrases in the context of regular routines, the words in books draw on a broader range of vocabulary sampled from a diverse set of topics. Not only do books provide children with access to these words, but they also provide a more contextually diverse environment for learning of individual words. Greater lexical diversity in the input means that a given word is more likely to co-occur with a broader range of other words, such that children have opportunities to develop semantic associations between them. Words that occur in more diverse contexts are acquired earlier in development, and show a processing advantage in older children and adults (Hills, 2013; Hills et al., 2010; Hsiao & Nation, 2018; Johns et al., 2016).

Our analysis of lexical diversity corroborates previous research showing that children encounter a broader range of vocabulary in books compared to an equivalent-sized input of child-directed speech. Turning to the types of words that children experience via books compared to conversation, our analyses of lexical density and lexical sophistication indicate that a higher proportion of the words in books are meaning-bearing words, and that they tend to occur less frequently in the language overall. This is important given that word frequencies are highly skewed, with only a small number of words occurring very frequently (predominantly function words) and the majority of words forming the long tail of the distribution (Piantadosi, 2014). Child-directed speech samples disproportionately from the higher end of this frequency spectrum. This is unsurprising because, unlike written language, speech is generated in the moment, and therefore word choice is biased towards those words in a speaker's lexicon that are most readily accessible (Navarrete et al., 2006). Similarly, because spoken communication incorporates extra-linguistic information, the variety, choice, and density of content words play a less crucial role in communicating meaning than they do in texts. This suggests that children's books are a particularly rich source of exposure to the types of words that children encounter rarely, if ever, in everyday conversation. While we focused on language directed primarily at pre-schoolers, children may have limited opportunity to access more advanced word types through speech alone, even once they reach school age: although caregivers draw on a more diverse vocabulary when speaking to older children, the types of words they choose come from the same part of the frequency distribution as the words used with younger children (Hayes & Ahrens, 1988). This evidence from older children reinforces book language as a critical source of lexical input.

Differences also emerged in part of speech distributions across the picture book and spoken language corpora. Our analysis revealed that among the major part of speech categories, nouns, adjectives, determiners, prepositions and conjunctions occur with greater relative frequency in books compared to child-directed speech, whereas pronouns are almost twice as common in speech compared to books. The balance of nouns and pronouns in a language sample is typically a trade-off, given that they perform a similar grammatical function (Hudson, 1994). In most comparisons of written and spoken language, nouns are found to occur more frequently in texts than in speech, whereas the reverse is true for pronouns (Rayson et al., 2001). This pattern is particularly characteristic of informational or academic texts, where nominalisations are a common feature and occur more frequently than in fiction (Biber et al., 1998), but our findings indicate that the same is true even for fiction targeted at pre-school children. In books, explicit reference is important for comprehension: characters and objects do not exist in the immediate context and cannot be experienced directly. In child-directed speech, the focus of communication is more interpersonal and takes place within a shared context such that pronouns are often an adequate substitute for nouns. The breakdown of pronoun types indicates that differences in frequency were particularly stark in relation to demonstrative (e.g., *that's the wrong one*), interrogative (*what did I say?)* and personal (*you'll get stuck*) pronouns, all of which reflect a more involved and interactive style and reference entities within the immediate physical environment.

Adjectives were also more characteristic of books than of child-directed speech. Again, this finding aligns with comparisons of written and spoken language more broadly: given that adjectives modify nouns, a greater proportion of nouns in a text is likely to be accompanied by a similar rise in adjectives (Biber, 1988; Mair et al., 2002; Rayson et al., 2001). Nevertheless, in relation to children's learning, acquisition of adjectives plays a key role in the development of a sophisticated lexicon. Adjectives form the basis of descriptions (e.g., *the fluffy cat*) and contrastive relations (e.g., *big truck vs. little truck*), and provide linguistic labels for sensory perceptions, values, and emotions (e.g., *she is cold; he is good; I feel happy*). The meanings of adjectives also tend to vary according to context. For example, *a big rat* differs in size to *a big building* – such terms are relative rather than absolute (Davies et al., 2020). Therefore, experiencing an adjective in combination with a more diverse set of nouns may facilitate a more robust and flexible representation of that word (Blackwell, 2005). This contextual dependency also suggests that children need some basic knowledge of the nouns being modified by a given adjective before they can develop mastery of the adjective itself. Unsurprisingly, children learn adjectives at a slower rate than they do other open word classes, particularly nouns (Caselli et al., 1995; Gasser & Smith, 1998; Sandhofer & Smith, 2007). Storybooks may be a particularly rich source of input for acquisition of adjectives, given that they occur more frequently than in speech, and they also provide more varied contexts through which semantic representations of adjectives can be accumulated and refined.

Our keyword analysis revealed the words that were most unique to the books in our corpus, and a second set of words that occurred in the books, but were relatively more frequent in child-directed speech. We found that the words most representative of children's books are typically acquired later in development according to age of acquisition norms, and are more abstract and more emotionally arousing than the words more common in child-directed speech. However, we found no difference between the two sets of words in relation to whether the emotions they evoked were strongly positive or negative. These findings corroborate our analysis of lexical sophistication, showing that the words in books are more advanced not only in terms of their frequency of occurrence in English overall, but also in relation to the stage of development that children usually acquire them. This has implications for children's language learning. Words that are acquired earlier in development tend to be well-connected to other words in the lexicon, whereas later words have fewer connections (Hills et al., 2009; Steyvers & Tenenbaum, 2005). According to one theory of how children expand their semantic network, the order in which children acquire new words reflects the connectivity of those words to other words in the learning environment (Hills et al., 2009). The words children hear in child-directed speech have a lower age of acquisition on average, and are more likely to be well-integrated in children's semantic networks. Access to books, on the other hand, provides an environment in which children can build semantic associations and develop connections between words that they may not otherwise encounter until later in development.

These words will also typically be more abstract. Concreteness is an important predictor of lexical processing in adults, with words higher in concreteness showing an advantage over abstract words (e.g., Binder et al., 2005), and abstract words tend to be acquired later in development (Ponari et al., 2018). One explanation is that concrete words (e.g., *apple*) refer to concepts that encode direct sensory experiences, and these imaginal representations are activated alongside verbal information during processing and retrieval. By contrast, abstract words (e.g., *validity*) rely more heavily on semantic information encoded linguistically, and the absence of support from perceptual memory means that these words are processed less efficiently (Paivio, 1971, 2013). The concreteness effect has also been accounted for by differences in contextual availability: abstract words are more challenging because they have weaker connections to associated contextual information, which makes it more difficult for an individual to activate that information when the word is encountered in isolation (Schwanenflugel, 1991). Underpinning both accounts is the idea that linguistic experience is key to the acquisition and processing of abstract words. Our analyses suggest that books provide more concentrated access to the types of words that are not supported by direct sensory experience, along with the linguistic and contextual information needed to support learning and consolidation. Acquisition of these words may be supported too by their affective properties. We found that the words in picture books were more emotionally arousing than the words in child-directed speech, although they did not differ on strength of valence ratings. Some theories of embodied semantics propose that emotion may play an important role in the acquisition and processing of abstract words in particular, functioning as an alternate source of experiential information in the absence of sensorimotor input (Kousta et al., 2011; Ponari et al., 2018; Vigliocco et al., 2014, 2018). However, a recent cross-linguistic study based on data from the MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007) found limited evidence that arousal and valence predicted children's comprehension and production of early-acquired words (Braginsky et al., 2019).

Our comparisons of children's picture books and child-directed speech provide clear evidence that books are lexically richer overall, and have a different composition in relation to grammatical class and structural complexity compared to speech. Furthermore, the words children are least likely to encounter via conversation alone are more advanced, more abstract, and more emotionally arousing. Many of the features of 'book language' we have identified are true of written vs. spoken language comparisons more broadly (e.g., Biber, 1988), but it is nevertheless important to document the ways in which these sources of language input differ in relation to children's experiences. Doing so not only highlights the specific lexical structures and properties that may vary across language learning environments, but also reveals that even books designed to be accessible to the youngest children still provide a rich lexical input that is quite different to everyday speech.

In many ways, narrative fiction, particularly for young children, is more akin to oral language than other written genres (e.g., academic texts, newspapers), meaning that our findings are likely to be conservative estimates of the differences between book language and speech. However, it is also important to recognise that the corpus of child-directed speech we used here was predominantly sampled from interactions taking place within home settings, and that this may have limited the range and richness of vocabulary that caregivers used with their children. For example, experiences outside the home (a visit to the zoo, a trip to the beach) may provide greater opportunity for novelty and variety in lexical use, and for talk beyond the 'here and now'. More broadly, while corpus data provides valuable insights into the language structures children have opportunities to experience via books, it cannot speak to the effects of exposure on learning in individuals. Frequency counts alone do not capture the rich, interactive contexts in which language learning takes place (Roy et al., 2015), and nor do they accommodate the wider benefits of shared reading experiences, such as extra-text talk, scaffolding and emotional bonding.

While less lexically rich than book language, child-directed speech nevertheless plays an important part in children's language development. Certain properties of child-directed speech, such as exaggerated intonation patterns and grammatical simplification, have been hypothesised to support early language acquisition (Soderstrom, 2007). Given that the words in books are more advanced, the impact of variation in exposure to book language may relate more closely to the skills that underpin children's emerging literacy. The words that children encounter in picture books are by definition more characteristic of the literary domain. Importantly, experience is key: exposure to picture books via shared reading allows children to start encoding the phonological forms and meanings of more advanced words across different contexts from an early age. Over time, this experience will shape language development and provide a strong foundation to literacy (e.g., Gough & Tunmer, 1986; Perfetti & Hart, 2002). While there are many potential benefits of shared reading for children's development, our findings suggest that one of the key contributions may stem from the language of the books themselves, and specifically the rich and diverse lexical input they offer.

References

Berman, R. A., & Nir, B. (2010). The lexicon in writing–speech-differentiation: Developmental perspectives. *Written Language and Literacy, 13*(2), 183–205. https://doi.org/10.1075/wll.13.2.01ber

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies*. Computational Linguistics, 19*, 219–241.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Binder, J. R., Westbury, C. F., McKiernan, K. A., Possing, E. T., & Medler, D. A. (2005). Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience, 17*(6), 905–917. https://doi.org/10.1162/0898929054021102

Blackwell, A. A. (2005). Acquiring the English adjective lexicon: Relationships with input properties and adjectival semantic typology. *Journal of Child Language, 32*(3), 535–562. https://doi.org/10.1017/S0305000905006938

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

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://biblio.ugent.be/publication/5774089/file/5774125.pdf

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science, 27*(6), 843–873. https://doi.org/10.1016/j.cogsci.2003.06.001

Cameron-Faulkner, T., & Noble, C. (2013). A comparison of book text and Child Directed Speech. *First Language, 33*(3), 268–279. https://doi.org/10.1177/0142723713487613

Carnegie Mellon University. (2014). *The CMU Pronouncing Dictionary.* http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development, 10*(2), 159–199. https://doi.org/10.1016/0885-2014(95)90008-X

Chang, Y. N., & Monaghan, P. (2019). Quantity and Diversity of Preliteracy Language Exposure Both Affect Literacy Development: Evidence from a Computational Model of Reading. *Scientific Studies of Reading, 23*(3), 235–253. https://doi.org/10.1080/10888438.2018.1529177

Clark, E. V. (2020). Conversational Repair and the Acquisition of Language. *Discourse Processes, 57*(5–6), 441–459. https://doi.org/10.1080/0163853X.2020.1719795

Davies, C., Lingwood, J., & Arunachalam, S. (2020). Adjective forms and functions in British English child-directed speech. *Journal of Child Language, 47,* 159–185. https://doi.org/10.1017/S0305000919000242

Demir-Lira, E., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science*, 1–16. https://doi.org/10.1111/desc.12764

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *The MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual (2nd Edition)*. Brookes Publishing.

Gasser, M., & Smith, L. B. (1998). Learning Nouns and Adjectives: A Connectionist Account. *Language and Cognitive Processes, 13*(2–3), 269–306. https://doi.org/10.1080/016909698386537

Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education, 7,* 6–10.

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

Hayes, D. P. (1988). Speaking and writing: Distinct patterns of word choice*. Journal of Memory and Language, 27*, 572–585. https://doi.org/10.1016/0749-596X(88)90027-7

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of "motherese"? *Journal of Child Language, 15*, 395–410.

Healey, P. G. T., de Ruiter, J. P., & Mills, G. J. (2018). Editors' Introduction: Miscommunication. *Topics in Cognitive Science, 10*(2), 264–278. https://doi.org/10.1111/tops.12340

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science, 10*(2), 367–388. https://doi.org/10.1111/tops.12336

Hills, T. (2013). The company that words keep: Comparing the statistical structure of child- Versus adult-Directed language. *Journal of Child Language, 40*(3), 586–604. https://doi.org/10.1017/S0305000912000165

Hills, T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory & Language, 63*(3), 259–273. https://doi.org/10.1038/jid.2014.371

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science, 20*(6), 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x

Hlaváčová, J. (2006). New approach to frequency dictionaries - Czech example. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 373–378.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 17*(5), 1368–1378. https://doi.org/10.1111/j.1467-8721.2008.00596.x

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language, 103*(August), 114–126. https://doi.org/10.1016/j.jml.2018.08.005

Hsu, N., Hadley, P. A., & Rispoli, M. (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of Child Language, 44*(1), 63–86. https://doi.org/10.1017/S0305000915000690

Hudson, R. (1994). About 37% of Word-Tokens are Nouns. *Language, 70*(2), 331–339.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental Psychology, 27*(2), 236–248. https://doi.org/10.1037/0012-1649.27.2.236

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

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning, 63*(SUPPL. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics, 53*, 61–79. https://doi.org/10.7820/vli.v01.1.koizumi

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review, 23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology, 98*, 1–21. https://doi.org/10.1016/j.cogpsych.2017.07.002

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics, 6*(1), 97–133.

Kilgarriff, A. (2009). Simple Maths for Keywords. In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The Representation of Abstract Words: Why Emotion Matters. *Journal of Experimental Psychology: General, 140*(1), 14–34. https://doi.org/10.1037/a0021446

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Laws, J., & Ryder, C. (2014). *MorphoQuantics*. http://morphoquantics.co.uk/

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates.

Mair, C., Hundt, M., Leech, G. N., & Smith, N. (2002). Short Term Diachronic Shifts in Part-of-Speech Frequencies: A Comparison of the Tagged LOB and F-LOB Corpora. *International Journal of Corpus Linguistics, 7*(2), 245–264. https://doi.org/10.2307/3722566

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.

Massaro, D. W. (2015). Two different communication genres and implications for vocabulary development and learning to read. *Journal of Literacy Research, 47*(4), 505–527. https://doi.org/10.1177/1086296X15627528

Massaro, D. W. (2017). Reading aloud to children: Benefits and implications for acquiring literacy before schooling begins. *The American Journal of Psychology, 130*(1), 63–72.

Montag, J. L. (2019). Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language, 39*(5), 527–546. https://doi.org/10.1177/0142723719849996

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science, 26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

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

Navarrete, E., Basagni, B., Alario, F. X., & Costa, A. (2006). Does word frequency affect lexical selection in speech production? *Quarterly Journal of Experimental Psychology, 59*(10), 1681–1690. https://doi.org/10.1080/17470210600750558

Paivio, A. (1971). *Imagery and Verbal Processes*. Holt, Rinehart and Winston.

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al. (2011). *Journal of Experimental Psychology: General, 142*(1), 282–287. https://doi.org/10.1037/a0027004

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child Development, 76*(4), 763–782. https://doi.org/10.1111/1467-8624.00498-i1

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of Functional Literacy* (pp. 189–213). John Benjamins.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.1007/978-3-662-46024-5_8

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science, 21*(2), 1–12. https://doi.org/10.1111/desc.12549

R Developement Core Team. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Article 3.4.3*. https://www.r-project.org/

Rayson, P., Wilson, A., & Leech, G. (2001). Grammatical word class variation within the British National Corpus Sampler. *Language and Computers, 36*(1), 295–306. https://doi.org/10.1163/9789004334113_020

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57*(3), 348–379. https://doi.org/10.1016/j.jml.2007.03.002

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language, 35*(1), 185–205. https://doi.org/10.1017/S0305000907008343

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech vocabulary development. *Child Development, 83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

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 of the United States of America, 112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., & Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods, 50*(4), 1568–1580. https://doi.org/10.3758/s13428-017-0981-8

Sandhofer, C., & Smith, L. B. (2007). Learning Adjectives in the Real World: How Learning Nouns Impedes Learning Adjectives. *Language Learning and Development, 3*(3), 233–267. https://doi.org/10.1080/15475440701360465

Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The Psychology of Word Meanings* (pp. 223–248). Lawrence Erlbaum Associates.

Snow, C. (2010). Academic language and the challenge of reading for learning about science. *Science, 328*, 450–452.

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

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78. https://doi.org/10.1207/s15516709cog2901_3

Strömqvist, S., Johansson, V., Kriz, S., Ragnarsdóttir, H., Aisenman, R., & Ravid, D. (2002). Toward a cross-linguistic comparison of lexical quanta in speech and writing. *Written Language & Literacy, 5*(1), 45–68. https://doi.org/10.1075/wll.5.1.03str

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren & J. L. M. Trim (Eds.), Applications of linguistics. *Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969* (pp. 443–452). Cambridge University Press.

van Heuven, W. J. B. Van, 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. https://doi.org/10.1080/17470218.2013.850521

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex, 24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025

Vigliocco, G., Ponari, M., & Norbury, C. (2018). Learning and Processing Abstract Words and Concepts: Insights From Typical and Atypical Development. *Topics in Cognitive Science, 10*(3), 533–549. https://doi.org/10.1111/tops.12347

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191–1207. http://macsphere.mcmaster.ca/handle/11375/16227

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weizman, Z. O., & Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*(2), 265–279. https://doi.org/10.1037/0012-1649.37.2.265

Data, Code and Materials Availability Statement

The corpora used in this project are available on the Open Science Framework along with our analysis scripts and Supplementary Materials at https://osf.io/zta29/. Note that for copyright reasons, word order within each text in the picture book corpus has been randomised.

Authorship and Contributorship Statement

ND was involved in conceptualization of the research, led on project management, data analysis, and data curation, and wrote the first draft of the manuscript. YH was involved in conceptualization of the research, provided advice on data analysis, and contributed to reviewing and editing the draft manuscript. AT took the lead on corpus processing and data analysis for some of the measures, and contributed to reviewing and editing the draft manuscript. NB contributed to the acquisition of data, corpus processing, and reviewing and editing the draft manuscript. KN contributed to the conceptualization of the research, and was responsible for funding acquisition, provision of resources, supervision, and reviewing and editing the draft manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This research was funded by The Nuffield Foundation (reference number: EDO 43392). The authors have no conflicts of interest to declare. We would like to thank Nilo Pedrazzini and William Thurwell for their contributions to data processing, and Eleanor Holton, Emma Jackson, Georgia Sandars, Rebecca Williams and Tsvetana Myagkova for their assistance in creating and transcribing the picture book corpus.

Appendices

Appendix A: List of Titles in the Picture Book Corpus

Title	Author	Target age range (years)
A Dog with Nice Ears	Lauren Child	3 to 7
A Great Big Cuddle	Michael Rosen	2 to 7
A Little Bit Brave	Nicola Kinnear	2 to 6
A Squash and a Squeeze	Julia Donaldson	3+
Aliens Love Underpants	Claire Freedman	3+
All the Colours I See	Allegra Agliardi	5+
Along Came A Different	Tom McLaughlin	3+
Animal Stories for 5 year olds	Helen Paiba	5 to 9
Barking for Bagels	Michael Rosen	6+
Bedtime Stories for 5 year olds	Helen Paiba	5 to 9
Brown Bear, Brown Bear, What Do You See?	Bill Jnr Martin	2+
But Excuse Me That is My Book	Lauren Child	4+
Colin and Lee: Carrot and Pea	Morag Hood	3+
Cyril and Pat	Emily Gravett	3 to 7
Dave the Lonely Monster	Anna Kemp	2+
Dear Zoo	Rod Campbell	2+
Dinosaur Roar!	Paul Stickland & Henrietta Stickland	1 to 5
Dogger	Shirley Hughes	2+
Dogs Don't Do Ballet	Anna Kemp & Sara Ogilvie	3+
Duck, Death, and the Tulip	Wolf Erlbruch	4 to 8
Each Peach Pear Plum	Allan Ahlberg & Janet Ahlberg	0+
Elmer	David McKee	3+
FArTHER	Grahame Baker-Smith	7+
Fat Frog	Ruth Miskin	5 to 7
Five Minutes Peace	Jill Murphy	3 to 5
Fox & Goldfish	Nils Pieters	3+
Fox's Socks	Julia Donaldson	1+
Franklin's Flying Bookshop	Jen Campbell	6 to 8
Funny Stories for 5 Year Olds	Helen Paiba	5 to 9
George's Marvellous Medicine	Roald Dahl	7+
Get up!	Ruth Miskin	5 to 7
Giraffe in the Bath and Other Tales	Russell Punter & Lesley Sims	3+
Gracie la Roo Goes to School	Marsha Qualey	6+
Gracie la Roo Sets Sail	Marsha Qualey	5+
Grandad's Island	Benji Davies	5+
Granpa	John Burningham	5 to 7
Guess How Much I Love You	Sam McBratney	2+

Hairy Maclary from Donaldson's Dairy	Lynley Dodd	2+
Hampstead the Hamster	Michael Rosen	5+
Heidi	Johanna Spyri	6+
Hide and Seek	NA	3 to 7
Hide-and-Seek Pig	Julia Donaldson & Axel Scheffler	1+
Hippo has a Hat	Julia Donaldson	0 to 3
Horrid Henry and the Secret Club	Francesca Simon	6 to 11
Horrid Henry tricks the Tooth Fairy	Francesca Simon	7 to 10
Horrid Henry: Ghosts and Ghouls	Francesca Simon	7 to 9
Horrid Henry's Halloween Horrors	Francesca Simon	6 to 11
How to be a Lion	Ed Vere	3+
Hubert Horatio How to Raise your Grown-ups	Lauren Child	7 to 11
I Can Hop	Ruth Miskin	5 to 7
I Want My Hat Back	Jon Klassen	6+
If All the World Were...	Joseph Coelho & Allison Colpoys	0 to 6
In the Bath	Ruth Miskin	5 to 7
Into the Forest	Anthony Browne	8+
Is it a Mermaid?	Candy Gourlay	3 to 7
John Brown, Rose and the Midnight Cat	Jenny Wagner	2 to 4
Joy	Corrinne Averiss	3 to 6
Kitchen Disco	Clare Foges & Al Murphy	5+
Little Beauty	Anthony Browne	2+
Looking for Atlantis	Colin Thompson	8+
Lost and Found	Oliver Jeffers	3+
Loved To Bits	Teresa Heapy & Katie Cleminson	3 to 6
Magical Stories for 5 year olds	Helen Paiba	5 to 9
Me and my Fear	Francesca Sanna	3 to 7
Michael Rosen's Sad Book	Michael Rosen	6+
Mog the Forgetful Cat	Judith Kerr	2+
Monkey Puzzle	Julia Donaldson	3 to 8
Mr Men: Chinese New Year	Adam Hargreaves	3+
Murray the Race Horse	Gavin Puckett	7 to 9
My Father's Arms are a Boat	Stein Erik Lunde	4+
Nice Work for the Cat and the King	Nick Sharratt	6 to 9
Night-Time Cat	Julia Tedd	7
Nip and Chip	Ruth Miskin	5 to 7
No-Bot	Sue Hendra & Paul Linnet	3+
Nog in the Fog	Ruth Miskin	5 to 7
Odd Dog Out	Rob Biddulph	3+
of Thee I sing	Barack Obama	4+
Oi Cat!	Kes Gray	1 to 5

Oi Dog!	Kes & Claire Gray	3+
Oi Frog!	Kes Gray	3+
Oi Goat!	Kes Gray	3+
Owl Babies	Martin Waddell & Patrick Benson	3+
Pants	Giles Andreae	2 to 3
Peace at Last	Jill Murphy	3+
Peck Peck Peck	Lucy Cousins	3 to 5
Peppa goes to London	Lauren Holowaty	3+
Peppa meets Father Christmas	Lauren Holowaty	2 to 6
Peppa the Mermaid	Lauren Holowaty	2 to 6
Peppa's Magical Unicorn	Lauren Holowaty	3+
Princess Mirror-Belle and the Flying Horse	Julia Donaldson	7 to 11
Princess Mirror-Belle and the Sea Monster's Cave	Julia Donaldson	7 to 11
Rabbit & Bear Attack of the Snack	Julian Gough	5 to 7
Rabbit & Bear The Pest in the Nest	Julian Gough	5 to 7
Rabbityness	Jo Empson	5+
Raccoon on the Moon	Russell Punter	3+
Rag the Rat	Ruth Miskin	5 to 7
Red Ned	Ruth Miskin	5 to 7
Room on the Broom	Julia Donaldson	6+
Rosie's Walk	Pat Hutchins	0+
Ruby Red Shoes Goes to London	Kate Knapp	4+
Ruby's Worry	Tom Percival	5+
Run, Run, Run!	Ruth Miskin	5 to 7
Sharing a Shell	Julia Donaldson	2+
Sophie Johnson Unicorn Expert	Morag Hood	3+
Squishy McFluff the Invisible Cat: Seaside Rescue!	Pip Jones	5+
Stardust	Jeanne Willis	5+
Stick Man	Julia Donaldson	6+
Sun Hat Fun	Ruth Miskin	5 to 7
Superworm	Julia Donaldson	2 to 7
Sweep	Louise Greig & Julia Sarda	3+
That's Not my Puppy...	Fiona Watt	0+
That's Not my Unicorn…	Fiona Watt	0+
The Bad-Tempered Ladybird	Eric Carle	2+
The BFG	Roald Dahl	6+
The Building Boy	Ross Montgomery	4+
The Bumblebear	Nadia Shireen	4+
The Cat in the Hat	Dr Seuss	5+
The Day the Crayons Quit	Drew Daywalt & Oliver Jeffers	3 to 7
The Day War Came	Nicola Davies	5+

The Detective Dog	Julia Donaldson	3 to 7
The Flat Rabbit	Bardur Oskarsson	4 to 6
The Gift	Carol Ann Duffy	7+
The Gruffalo	Julia Donaldson	3 to 7
The Gruffalo's Child	Julia Donaldson	3+
The Heart and the Bottle	Oliver Jeffers	6+
The Highway Rat	Julia Donaldson	2 to 6
The Jolly Christmas Postman	Janet Ahlberg & Allan Ahlberg	3 to 5
The Jolly Postman or Other People's Letters	Janet Ahlberg & Allan Ahlberg	3 to 5
The Last Chip: The Story of a Very Hungry Pigeon	Duncan Beedie	3+
The Lion Inside	Rachel Bright	3+
The Marvellous Moon Map	Teresa Heapy & David Litchfield	3 to 7
The Memory Tree	Britta Teckentrup	3 to 5
The Owl who was Afraid of the Dark	Jill Tomlinson	5+
The Paper Dolls	Julia Donaldson	3+
The Pond	Nicola Davies	5 to 7
The Scar	Charlotte Moundlic	5+
The Smartest Giant in Town	Julia Donaldson	4 to 7
The Snail and the Whale	Julia Donaldson	2 to 4
The Storm Whale	Benji Davies	3+
The Storm Whale in Winter	Benji Davies	1+
The Tiger Who Came to Tea	Judith Kerr	2+
The Twits	Roald Dahl	7 to 9
The Ugly Five	Julia Donaldson	2 to 6
The Very Hungry Caterpillar	Eric Carle	0+
The Wonky Donkey	Craig Smith	2 to 6
Tiddler	Julia Donaldson	5 to 11
Tug, tug	Ruth Miskin	5 to 7
Very little Cinderella	Teresa Heapy & Sue Heap	4 to 6
We're Going on a Bear Hunt	Michael Rosen	6+
What Happens Next	Shinsuke Yoshitake	8+
What is Poo?	Katie Daynes	0 to 5
Whatever Next!	Jill Murphy	3 to 5
When Sadness Comes to Call	Eva Eland	3 to 8
Where the Wild Things Are	Maurice Sendak	2+
Where's Spot?	Eric Hill	0+
Willy and the Cloud	Anthony Browne	3 to 7
Willy the Wimp	Anthony Browne	7+
Witchfairy	Brigitte Minne	4+
Zog	Julia Donaldson	2 to 7
Zog and the Flying Doctors	Julia Donaldson	2 to 6

Note. The full corpus is available as .csv files containing word tokens (randomised within each document) on the Open Science Framework project page (https://osf.io/zta29/)

Appendix B: Summary of CHILDES Corpora in the Spoken Language Corpus

Corpus	Child age range	n	Reference
Belfast	2;0-4;5	8	Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. New York: Oxford University Press.
Gathercole/Burns	3;0-6;4	12	Gathercole, V. (1986). The acquisition of the present perfect: explaining differences in the speech of Scottish and American children. *Journal of Child Language*, *13*, 537–560
Howe	1;6-1;8 (session 1) 1;11-2;1 (session 2)	16	Howe, C. (1981). *Acquiring language in a conversational context.* New York: Academic Press.
Korman	6-16 weeks	6	Korman, M., & Lewis, C. (2001). Mothers' and fathers' speech to their infants: Explorations of the complexities of context. In M. Almgren, A. Barreña, M.-J. Ezeizabarrena, I. Idiazaabal, & B. MacWhinney (Eds.), *Research on child language acquisition* (pp. 431-453). Somerville, MA: Cascadilla Press
Lara	1;9-3;3	1	Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1-21. doi:10.1016/j.cogpsych.2017.07.002.
Manchester	1;8-3;0	12	Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.
MPI-EVA Manchester	1;8-3;2	4	Lieven, E., Salomo, D. & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20* (3), 481-508.
Nuffield	0;11	76	McGillion, M., Pine, J. M., Herbert, J. S., & Matthews, D. (2017). A randomised controlled trial to test the effect of promot-

			ing caregiver contingent talk on language development in infants from diverse socioeconomic status backgrounds. *Journal of Child Psychology and Psychiatry, 58* (10), 1122-1131
Tommerdahl	2;6-3;6	23	Tommerdahl, J. and Kilpatrick, C. (2014). The Reliability of Morphological Analyses in Language Samples. *Journal of Language Testing, 31* (1), 3-18.
Wells	1;6-5;0	32	Wells, C. G. (1981). *Learning through interaction: The study of language development.* Cambridge, UK: Cambridge University Press.

Note. n = number of children in sample. The full corpus is available as .csv files on the Open Science Framework project page (https://osf.io/zta29/)

Appendix C: List of the 50 Book+ Words with Highest Keyness Score and 50 Book— Words with Lowest Keyness Score

Word	Picture book corpus frequency per million	Spoken language corpus frequency per million	Keyness score	Word set
stare	195.31	2.74	16.12	book
voice	293.6	9.96	15.21	book
begin	401.56	18.76	14.31	book
horrid	274.42	10.03	14.2	book
suddenly	252.01	9.74	13.27	book
father	202.7	6.08	13.22	book
everyone	351.18	18.07	12.87	book
yell	130.8	1.48	12.26	book
world	275.73	13.64	12.09	book
giant	290.58	15.5	11.79	book
deep	201.58	8.09	11.7	book
gasp	121.37	1.3	11.62	book
whisper	188.5	7.48	11.36	book
dad	328.98	21.91	10.62	book
leap	129.11	3.17	10.57	book
sigh	114.5	1.91	10.46	book
perfect	168.61	7.45	10.24	book
enormous	116.03	2.47	10.11	book
reply	106.59	2.3	9.48	book
thought	361.99	29.84	9.34	book
shriek	86.13	0.33	9.3	book
mutter	87.35	0.48	9.29	book
large	141.09	6.94	8.92	book
cheer	98.01	2.46	8.67	book
shout	526.77	52.41	8.6	book
dream	175.13	11.63	8.56	book
each	383.27	36.42	8.47	book
towards	141.98	8.12	8.39	book
cave	99.49	3.12	8.35	book
silence	74.84	0.27	8.26	book
sight	106.82	4.52	8.05	book
howl	75.96	0.69	8.04	book
mother	271.95	25.67	7.9	book
ground	231.7	20.58	7.9	book
against	118.92	6.34	7.89	book
breath	105.48	4.67	7.87	book
parent	82.73	1.85	7.83	book
human	73.42	0.7	7.8	book
slowly	157.61	11.54	7.78	book

evening	108.98	5.43	7.71	book
smile	348.68	36.89	7.65	book
hate	99.93	4.38	7.65	book
most	278.81	28.2	7.56	book
street	128.92	8.39	7.56	book
himself	300.22	31.11	7.55	book
peer	66.64	0.31	7.44	book
scream	167.75	14.21	7.34	book
add	114.5	7.37	7.17	book
notice	186.62	17.45	7.16	book
gaze	62.4	0.27	7.05	book
toy	73.3	257.58	0.31	spoken
where	735.23	2459.97	0.3	spoken
train	63.85	237.14	0.3	spoken
because	496.21	1716.97	0.29	spoken
tidy	22.96	102.79	0.29	spoken
cuddle	7.45	49.88	0.29	spoken
hey	67.96	263.26	0.29	spoken
bye	15.41	79.69	0.28	spoken
ooh	24.83	115.13	0.28	spoken
here	701.04	2567.41	0.28	spoken
toilet	8.96	58.76	0.28	spoken
well	820.1	3061.13	0.27	spoken
put	762.26	2919.67	0.26	spoken
tissue	6.86	54.18	0.26	spoken
brick	18.21	99.05	0.26	spoken
yours	36.3	169.13	0.26	spoken
trouser	17.96	99.15	0.26	spoken
do	5333.39	20871.79	0.26	spoken
right	820.27	3235.08	0.26	spoken
giraffe	10.22	69.44	0.25	spoken
oops	5.37	51.39	0.25	spoken
doll	32.06	160.44	0.25	spoken
we	1391.55	5698.67	0.25	spoken
today	110.42	482.4	0.24	spoken
what	2794.66	11487.89	0.24	spoken
yum	12.12	81.21	0.24	spoken
naughty	37.44	187.53	0.24	spoken
yesterday	27.44	150.82	0.23	spoken
car	90.86	425.88	0.23	spoken
oy	3.25	50.83	0.22	spoken
whee	6.69	67.96	0.21	spoken
you	7427	34770.54	0.21	spoken

want	712.68	3370.12	0.21	spoken
penguin	4.18	57.24	0.21	spoken
yes	515.37	2786.8	0.19	spoken
nursery	3.3	64.39	0.18	spoken
poorly	3.45	66.25	0.18	spoken
shall	204.17	1366.71	0.16	spoken
careful	35.75	303.12	0.15	spoken
mm	6.21	106.22	0.14	spoken
jigsaw	5.45	104.19	0.14	spoken
wee	16.98	246.49	0.11	spoken
hm	46.59	547.67	0.1	spoken
oh	828.94	8781.37	0.1	spoken
whoops	4.35	170.3	0.08	spoken
okay	89.43	1581.96	0.06	spoken
pardon	18.23	469.11	0.06	spoken
darling	20.93	629.48	0.05	spoken
alright	4.49	519.47	0.03	spoken
yeah	28.24	2554.14	0.01	spoken

Note. Frequency columns show average reduced frequencies per million prior to the addition of the constant (10)

License

Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories

Benjamin E. deMayo
Princeton University, USA

Danielle Kellier
University of Pennsylvania, USA

Mika Braginsky
Massachusetts Institute of Technology, USA

Christina Bergmann
Cielke Hendriks
Caroline F. Rowland
Max Planck Institute for Psycholinguistics, The Netherlands

Michael C. Frank
Virginia A. Marchman
Stanford University, USA

Abstract: Understanding the mechanisms that drive variation in children's language acquisition requires large, population-representative datasets of children's word learning across development. Parent report measures such as the MacArthur-Bates Communicative Development Inventories (CDI) are commonly used to collect such data, but the traditional paper-based forms make the curation of large datasets logistically challenging. Many CDI datasets are thus gathered using convenience samples, often recruited from communities in proximity to major research institutions. Here, we introduce Web-CDI, a web-based tool which allows researchers to collect CDI data online. Web-CDI contains functionality to collect and manage longitudinal data, share links to test administrations, and download vocabulary scores. To date, over 3,500 valid Web-CDI administrations have been completed. General trends found in past norming studies of the CDI (e.g., Feldman et al., 2000) are present in data collected from Web-CDI: scores of children's productive vocabulary grow with age, female children show a slightly faster rate of vocabulary growth, and participants with higher levels of educational attainment report slightly higher vocabulary production scores than those with lower levels of education attainment. We also report results from an effort to oversample non-white, lower-education participants via online recruitment (N = 243). These data showed similar age, sex, and primary caregiver education trends to the full Web-CDI sample, but this effort resulted in a high exclusion rate. We conclude by discussing implications and challenges for the collection of large, population-representative datasets.

Keywords: vocabulary development, parent report

Corresponding author(s): Benjamin deMayo, Department of Psychology, Princeton University, Princeton, New Jersey, USA. Email: bdemayo@princeton.edu.

ORCID ID(s): https://orcid.org/0000-0001-8723-6419

Citation: deMayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C. F., Frank, M. C., & Marchman, V. A. (2021). Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories. *Language Development Research*, 1(1), p 55-98. DOI: 10.34758/kr8e-w591

Introduction

Children vary tremendously in their vocabulary development (Fenson et al., 1994; Frank, Braginsky, Yurovsky, & Marchman, 2021). Characterizing this variability is central to understanding the mechanisms that drive early language acquisition, yet capturing this variation in broad, diverse samples of children has been a significant challenge for cognitive scientists for decades. The MacArthur-Bates Communicative Development Inventories (MB-CDI, or CDI for short) are a set of commonly used parent report instruments for assessing vocabulary development in early childhood (Fenson et al., 2007) that were introduced in part to create a cost-effective method for measuring variability across individuals.

In this paper, we introduce a web-based tool, Web-CDI, which was developed to address the need for collecting CDI data in an online format. Web-CDI allows researchers to increase the convenience of CDI administration, further decrease costs associated with data collection and entry (particularly with item-level data), and access participant samples that have traditionally been difficult to reach in language development research. Our purpose in this paper is twofold: first, we describe Web-CDI as a platform which streamlines the process of collecting CDI data and collates the data in a way that facilitates the creation of large-scale, multisite collaborative datasets. Second, we profile usage of Web-CDI thus far, with a particular focus on broadening the reach of traditional paper-based methods of collecting vocabulary development data.

The Importance of Parent Report Data

Gaining empirical traction on variation in children's early language requires reliable and valid methods for measuring language abilities, especially in early childhood (8 to 30 months). Parent report is a mainstay in this domain. Parents' reports are based on their daily experiences with the child, which are much more extensive than a researcher or clinician can generally obtain. Moreover, they are less likely to be influenced by factors that may mask a child's true ability in the laboratory or clinic (e.g., shyness). One widely used set of parent-report instruments is the MacArthur-Bates Communicative Development Inventories, originally designed for children learning American English (Fenson et al., 2007). The American English CDIs come in several versions, two of which are Words & Gestures (WG) for children 8 to 18 months, focusing on word comprehension and production, as well as gesture use, and Words & Sentences (WS) for children 16 to 30 months, focusing on word production and sentence structure. Both the WG and WS measures come in short forms with vocabulary checklists of approximately 90-100 words (Fenson et al., 2000), and long forms, which contain vocabulary checklists of several hundred items each. (An additional shorter form of the Web-CDI for children 30-37 months, CDI-III, also exists.) Together, the CDI instruments allow for a comprehensive picture of milestones that characterize language development in early childhood. A substantial body of evidence suggests that these instruments are both reliable and valid (e.g., Fenson et al., 1994, 2007), leading to their widespread use in thousands of research studies over the last few decades. Initial large-scale work to establish the normative datasets for the American English CDI not only provided key benchmarks for determining children's progress, but also documented the extensive individual differences that characterize early language learning during this critical period of development (Bates et al., 1994; Fenson et al., 1994). Understanding the origins and consequences of this variability remains an important empirical and theoretical endeavor (e.g., Bates & Goodman, 2001; Bornstein & Putnick, 2012; see also, Frank et al., 2021).

The popularity of CDI instruments has remained strong over the years, leading to extensions of the methodology to alternative formats and cross-language adaptations (Fenson et al., 2000). Many teams around the world have adapted the CDI format to particular languages and communities (Dale, 2015). Importantly, these adaptations are not simply translations of the original form but rather incorporate the specific features of different languages and cultures, since linguistic variability exists even among cultures that share a native language. As an example of this phenomenon, the word "Cheerios" is more common in the United States than it is in the United Kingdom; as a result, it might be expected that caregivers would report children's knowledge of this word in the U.S. and not the U.K., even though English is the most common language in both countries. To date there are more than 100 adaptations for languages around the globe. Moreover, several research groups have developed shorter versions of the CDI forms by randomly sampling items from the full CDI and comparing participants' responses to established norms (Mayor & Mani, 2019) or by developing computer adaptive tests (CATs) that use item response theory or Bayesian approaches to guide the selection of a smaller subset of items to which participants respond (Chai, Lo, & Mayor, 2020; Kachergis et al., 2021; Makransky, Dale, Havmose, & Bleses, 2016).

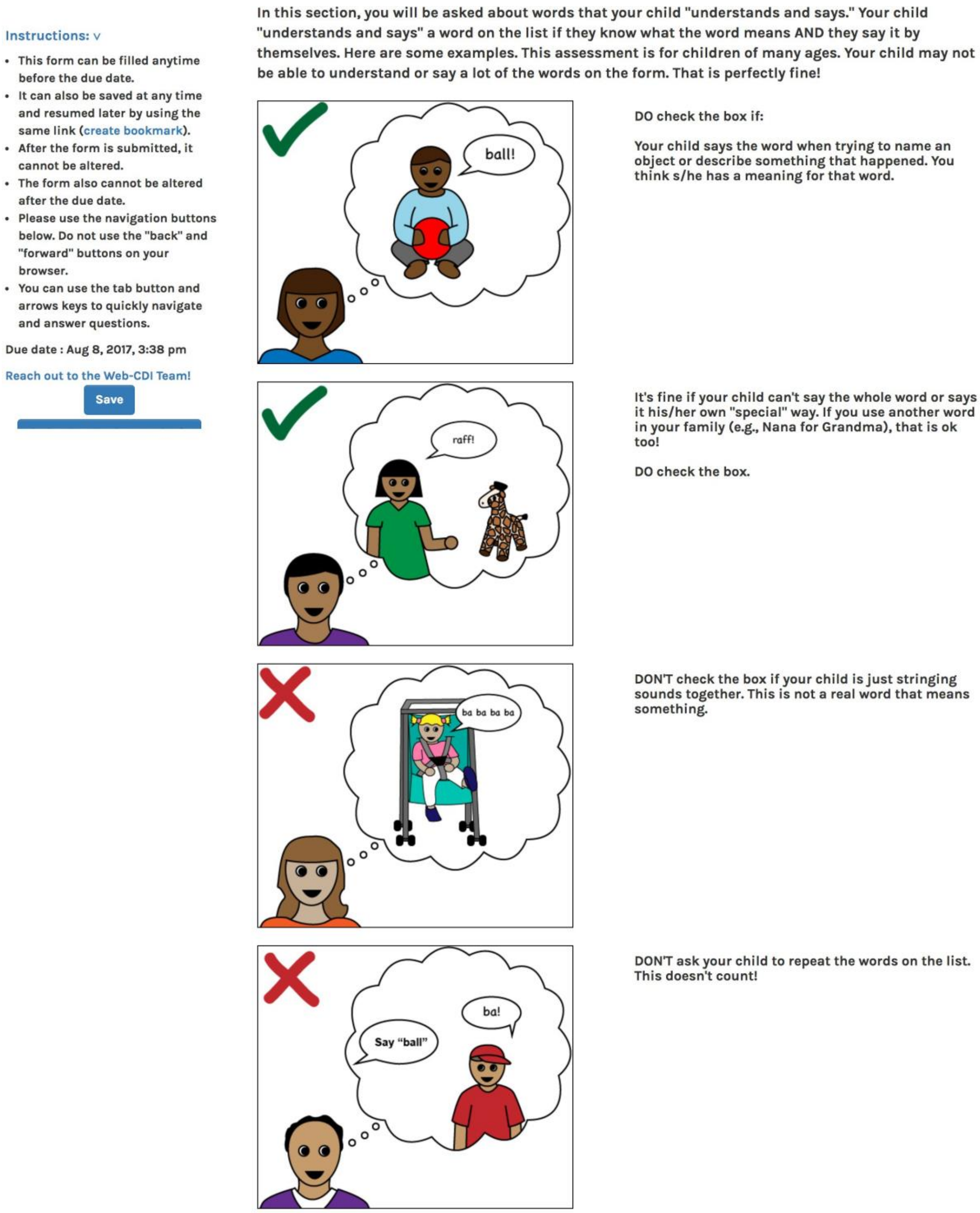
While the reliability and validity of the original CDI instruments are well-established for the American English versions of the forms and several others, most existing norming samples are skewed toward families with more years of formal education and away from non-white groups (Fenson et al., 2007). For example, representation in the American English norming samples is generally restricted to families living on the U.S. east and west coasts. Further, although paper survey administration is a time-tested method, increasingly, researchers and participants would prefer to use an electronic method to administer and fill CDI forms, obviating the need to track (and sometimes mail) paper forms, and the need to key in hundreds of item-wise responses for each child.

Here, we report on our recent efforts to create and distribute a web-based version of the CDIs in order to address some of the limitations of the standard paper versions. Online administration of the CDI is not a novel innovation – a variety of research groups have created purpose-build platforms for administering the CDI in particular languages. For example, Kristoffersen et al. (2013) collected a large normative sample of Norwegian CDIs using a custom online platform. Similarly, the Slovak adaptation of the CDI uses an online administration format (Kapalková & Slanèová, 2007). And many groups have used general purpose survey software such as Qualtrics and Survey Monkey to administer CDIs and variants online (e.g., Caselli, Lieberman, & Pyers, 2020). The innovation of Web-CDI is to provide a comprehensive researcher management interface for the administration of a wide range of CDI forms, allowing researchers to manage longitudinal administrations, download scores, and share links with parents easily, all while satisfying strong guarantees regarding privacy and anonymity. Moreover, a key benefit of a unified data collection and storage system such as Web-CDI is that data from disparate sources are combined into a single repository. This substantially reduces the overhead efforts associated with bringing together data collected by researchers across the world and allows for the analysis of large comparative datasets with the power to detect general trends in vocabulary development that may emerge across languages. Finally, due to an agreement between the CDI Advisory Board and Brookes Publishing, the publisher of the print versions of the CDI suite, Web-CDI is free of charge for those researchers who agree to contribute their data for the renorming of the long form instruments.

Introducing Web-CDI

Web-CDI is a web-based platform for CDI administration and management. Web-CDI allows researchers to communicate with families by sharing URLs (web links that contain individual users' own administration of the Web-CDI) via email or social media, facilitating access to families in areas distant from an academic institution and eliminating costly mailings and laboratory visits. Web-CDI also standardizes electronic administration and scoring of CDI forms across labs and institutions, making possible the aggregation of CDI data for later reuse and comparison across administrations by different labs. Indeed, researchers who use Web-CDI grant the CDI Advisory Board permission to access and analyze the resulting data on an opt-out basis, providing a path towards continual improvement of CDI instruments. Since 2018, more than 3,500 CDIs have been collected by 15 research groups throughout the U.S. who are using Web-CDI, demonstrating the potential for large-scale data collection and aggregation.

Below, we outline how Web-CDI is used. We begin by detailing the consent process and participant experience. Second, we describe the interface that researchers use to collect data using Web-CDI, specifying a number of common use cases for the platform.



Instructions: v

- This form can be filled anytime before the due date.
- It can also be saved at any time and resumed later by using the same link (create bookmark).
- After the form is submitted, it cannot be altered.
- The form also cannot be altered after the due date.
- Please use the navigation buttons below. Do not use the "back" and "forward" buttons on your browser.
- You can use the tab button and arrows keys to quickly navigate and answer questions.

Due date : Aug 8, 2017, 3:38 pm

Reach out to the Web-CDI Team!

In this section, you will be asked about words that your child "understands and says." Your child "understands and says" a word on the list if they know what the word means AND they say it by themselves. Here are some examples. This assessment is for children of many ages. Your child may not be able to understand or say a lot of the words on the form. That is perfectly fine!

DO check the box if:

Your child says the word when trying to name an object or describe something that happened. You think s/he has a meaning for that word.

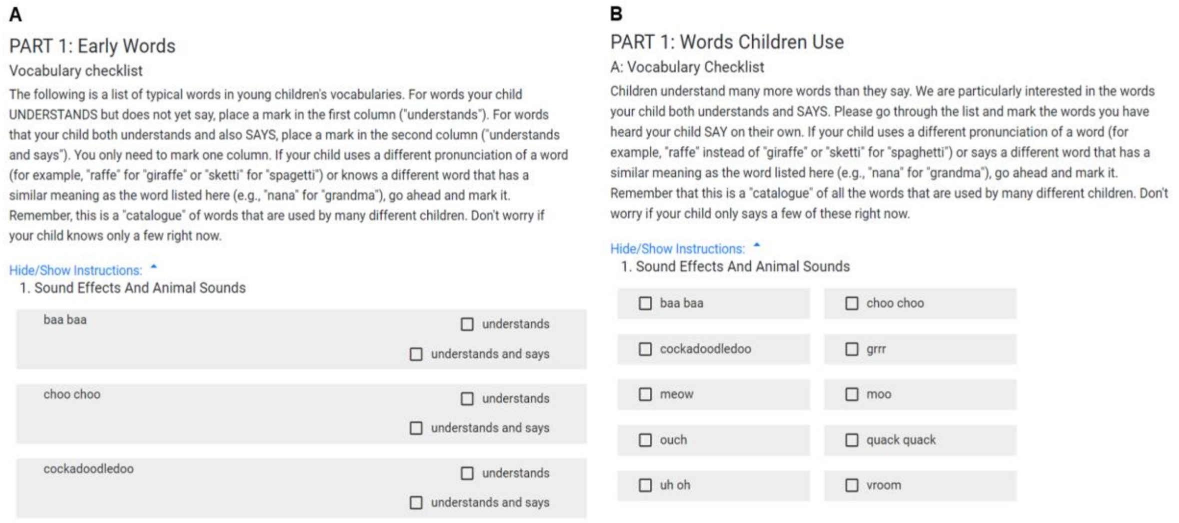
It's fine if your child can't say the whole word or says it his/her own "special" way. If you use another word in your family (e.g., Nana for Grandma), that is ok too!

DO check the box.

DON'T check the box if your child is just stringing sounds together. This is not a real word that means something.

DON'T ask your child to repeat the words on the list. This doesn't count!

Figure 1. ***Pictorial instructions indicating how to mark whether a child "understands and says" a word, from the Web-CDI WS instrument.***



A

PART 1: Early Words

Vocabulary checklist

The following is a list of typical words in young children's vocabularies. For words your child UNDERSTANDS but does not yet say, place a mark in the first column ("understands"). For words that your child both understands and also SAYS, place a mark in the second column ("understands and says"). You only need to mark one column. If your child uses a different pronunciation of a word (for example, "raffe" for "giraffe" or "sketti" for "spagetti") or knows a different word that has a similar meaning as the word listed here (e.g., "nana" for "grandma"), go ahead and mark it. Remember, this is a "catalogue" of words that are used by many different children. Don't worry if your child knows only a few right now.

Hide/Show Instructions: ▴

1. Sound Effects And Animal Sounds

baa baa □ understands □ understands and says

choo choo □ understands □ understands and says

cockadoodledoo □ understands □ understands and says

B

PART 1: Words Children Use

A: Vocabulary Checklist

Children understand many more words than they say. We are particularly interested in the words your child both understands and SAYS. Please go through the list and mark the words you have heard your child SAY on their own. If your child uses a different pronunciation of a word (for example, "raffe" instead of "giraffe" or "sketti" for "spaghetti") or says a different word that has a similar meaning as the word listed here (e.g., "nana" for "grandma"), go ahead and mark it. Remember that this is a "catalogue" of all the words that are used by many different children. Don't worry if your child only says a few of these right now.

Hide/Show Instructions: ▴

1. Sound Effects And Animal Sounds

□ baa baa	□ choo choo
□ cockadoodledoo	□ grrr
□ meow	□ moo
□ ouch	□ quack quack
□ uh oh	□ vroom

Figure 2. ***(A) Sample items from the American English WG form. (B) Sample items from the American English WS form.***

Participant Interface

Participants can complete the Web-CDI on a variety of devices, including personal computers and tablets. Web-CDI can also be administered on a smartphone, although the experience is not ideal for the user due to the length of the survey and the small screen. As Web-CDI moves in the future to incorporate more short forms and computer adaptive test (CAT) formats (e.g., Chai, Lo, & Mayor, 2020; Makransky, Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019), smartphone-responsive design will become a priority.

When a participant clicks a URL shared by a researcher, they are directed to a website presenting their own personal administration of the Web-CDI. In some cases, they may be asked to read and accept a waiver of consent documentation, depending on whether the researcher has chosen to use that feature (see also Researcher Interface below).

Instructions

After completing the first demographics page, participants are provided with detailed instructions that are appropriate for either the WG or WS version (see Figure 1 for an example of the instructions for parents to determine whether their child "understands and says" a word, which is pertinent to both the WG and WS forms). In addition, there are more detailed instructions for completing the vocabulary checklist. Unlike the traditional paper versions, instructions on how to properly choose responses are provided both in written and pictorial form. The pictorial instructions (Figure 1) aim to further increase caregivers' understanding of how to complete the checklist. For example, these instructions clarify that the child's understanding of a word requires them to have some understanding of the object that the word refers to or some aspect of the word's meaning. In addition, caregivers are reassured that "child-like" forms (e.g., "raff" for "giraffe") or family- or dialect-specific forms (e.g., "nana" for "grandma") are acceptable evidence. Lastly, caregivers are reminded that the child should be able to produce the words "on their own" and that imitations are not acceptable. These general "rules of thumb" for completing the form should be familiar to researchers who are distributing the forms to caregivers so they can field any questions that may arise. While this is not possible for certain use-cases (e.g., social media recruitment), these instructions should ideally also be reviewed either in writing (e.g., via email) or verbally (e.g., over the phone), so that these pictured instructions serve merely as a reminder to caregivers when completing the form. Pictured instructions are available for download on the MB-CDI website at http://mb-cdi.stanford.edu/about.html.

Completing the Instrument

The majority of the participant's time is spent completing the main sections of the instruments. As shown in Figure 2, on the American English WG form, the vocabulary checklist portion (396 items) asks caregivers to indicate whether their child "understands" or "understands and says" each word; they can also indicate that their child neither understands nor says the word by leaving the boxes unchecked. Additionally, gesture communication and other early milestones are assessed. In the American English WS form, the vocabulary checklist (680 items) only asks caregivers to indicate which words their child "says." Additional items assess children's production by requesting three of their longest sentences, as well as morphological and syntactic development more broadly. All of these items are broken up across multiple screens for easier navigation through the form.

At the completion of the form, a graph is displayed illustrating how the responses of "understands" or "understands and says" are distributed across the semantic categories on the form. Participants can select to download their own responses. In addition, data from the norming studies are used to estimate the "hardest" [i.e., most advanced based on previous work on age of acquisition of individual words, Frank et al. (2021)] word that the child currently understands or produces. This feedback to caregivers is intended to provide caregivers with a fun "thank you" and intentionally avoids any information which frames their child's progress relative to other children or any normative standard, so as to not give the impression that the Web-CDI is a clinical assessment of the child's development. To further underscore this point, the closing page reminds caregivers that their participation does not constitute a clinical evaluation and that they should contact their pediatrician or primary care physician if they have any concerns about their child's development.

Researcher Interface

One of the main goals of Web-CDI is to provide a unified CDI platform to the child language research community. To that end, researchers request an account by contacting members of the CDI Advisory Board at webcdi-contact@stanford.edu. Once the request is granted, they can design and distribute studies. One rationale for this personalized registration process is that we ask that researchers allow fully anonymized data from their participants to be shared with the CDI Advisory Board, so that it can be added to Wordbank [http://wordbank.stanford.edu/; Frank et al. (2017)] and shared with the broader research community. However, if particular participants indicate in the consent process that they do not want their data to be shared more broadly, then researchers can indicate this in the Web-CDI dashboard to prevent data from specific administrations being contributed to any analyses conducted by the CDI Advisory Board and/or Wordbank. Data currently in Web-CDI, which have not yet been added to the Wordbank repository, will be vetted before being added to ensure that all Web-CDI data in Wordbank are drawn from families with typically-developing children who meet similar inclusion criteria to the ones we describe below in the Dataset 1 section. Additionally, date of form completion will be preserved when adding Web-CDI data into Wordbank, so that researchers can choose to filter out data that may be affected by the particular point in time at which they were collected (e.g., the COVID-19 pandemic, Kartushina et al., 2021).

A study in the context of the Web-CDI system is a set of individual administrations created by a researcher that share certain specifications. Table A1 in the Appendix gives an overview of the customizable features that are available at the study level in Web-CDI. These features are set when creating a study using the "Create Study" tool, and most of the features can be updated continuously during data collection using the "Update Study" tool. While some of these features are only relevant to specific use cases (e.g., longitudinal research and social media data collection, described below), others are relevant to all researchers using Web-CDI.

There are currently several CDI forms available for distribution via Web-CDI, including the English WG and WS forms and forms in other languages (see Cross-linguistic Research below). When creating a study, researchers choose one of the forms that they would like to distribute to participants; only one can be used in a given study. Researchers who wish to send multiple forms to participants simultaneously (e.g., those conducting multilingual research) should create multiple studies, each with a single instrument associated with it.

Researchers can download participant data in two formats. Both formatting options output a comma-separated values file with one row per participant; the full data option includes participant-by-item responses, and allows researchers to explore item-level trends, while the summary data option omits item-level data and only provides summary scores and normative information, including total number of words understood/produced and percentile scores by age in months and sex. Percentile scores are calculated to a single percentile resolution using norms from Fenson et al. (2007).

Below, we outline several possible use cases of Web-CDI, as well the features which may facilitate them from a researcher's perspective.

Individual Recruitment

A first possible workflow using Web-CDI is to send unique study URLs to individual participants. Researchers do so by entering numerical participant IDs or by auto-generating a specified quantity of participant IDs, each with its own unique study URL, using the "Add Participants" tool in the researcher dashboard. New participants can be added on a continual basis so that researchers can adjust the sample size of their study during data collection. Unique links generated for individual participants expire, by default, 14 days after creation, though the number of days before link expiration is adjustable, which may be an important consideration for some researchers, depending on their participant populations and specific project timelines. Workflows that involve generating unique links are most suitable for studies which pair the CDI with other measures, or when researchers contact specific participants from an existing database.

Longitudinal Studies

Web-CDI also facilitates longitudinal study designs in which each participant completes multiple administrations. Researchers wishing to design longitudinal studies can do so by entering a list of meaningful participant IDs using the "Add Participants" tool in the researcher dashboard. If a specific participant ID is added multiple times, Web-CDI will automatically create multiple unique study URLs in the study dashboard that have that ID. In addition, when creating studies, researchers can select whether they would like the demographics information, vocabulary checklist, or no sections at all to be pre-filled when a participant fills out a repeat administration of the instrument. Unless researchers are interested in cumulative vocabulary counts, it is strongly recommended that they do not use the option to pre-fill the vocabulary checklist portion of the instrument in longitudinal administrations as caregivers should complete the instrument at each time point independently. In the case that researchers do choose this option, this is recorded in the Web-CDI database so that, when the data are added to Wordbank, researchers can choose to filter out any pre-filled questionnaires.

Social Media and Survey Vendors

Web-CDI contains several features designed to facilitate data collection from social media recruitment or through third-party crowd-sourcing applications and vendors (e.g., Amazon Mechanical Turk, Prolific). First, rather than creating unique survey links for each participant, researchers can also use a single, anonymous link. When a participant clicks the anonymous link, a new administration with a unique subject ID is created in the study dashboard. Additionally, Web-CDI studies have several customizable features that are geared towards anonymous online data collection. For example, researchers can adjust the minimum amount of time a participant must take to fill out the survey before they are able to submit; with a longer minimum time to completion, researchers can encourage a more thorough completion of the survey. This feature is typically most relevant in research designs in which participants are not vetted by the researcher or those in which there is no direct communication between participants and researchers, as might be the case when recruiting respondents on social media. Responses collected via personal communication with participants show low rates of too-fast responding, mostly removing the need for the minimum time feature. Even in the case of anonymous data collection, however, it is recommended that researchers not raise the minimum completion time higher than 6 minutes, since some caregivers of very young children may theoretically be able to proceed through the measure quickly if their child is not yet verbal. Aside from the minimum time feature, researchers can ask participants to verify that their information is accurate by checking a box at the end of the survey, and can opt to include certain demographic questions at both the beginning and end of the survey, using response consistency on these redundant items as a check of data quality.

Paid Participation

If researchers choose to compensate participants directly through the Web-CDI interface, Web-CDI has built-in functionality to distribute redeemable gift codes when a participant reaches the end of the survey. Web-CDI contains several features to facilitate integration with third-party crowdsourcing applications and survey vendors, should they choose to handle participant compensation through another platform. For example, when creating studies, researchers can enter a URL to which participants are redirected when they reach the end of the survey. In addition, researchers using the behavioral research platform Prolific can configure their study to collect participants' unique Prolific IDs and pre-fill them in the survey.

Cross-linguistic Research

Web-CDI forms are currently available in English (U.S. American and Canadian), Spanish, French (Quebecois), Hebrew, Dutch and Korean. We are looking to add more language forms to the tool, as the paper version of the forms has been adapted into more than 100 different languages and dialects, and further ongoing adaptations have been approved by the MB-CDI board (http://mb-cdi.stanford.edu/adaptations).

System Design

Web-CDI is constructed using open-source software. All of the vocabulary data collected in Web-CDI are stored in a standard MySQL relational database, managed using Django and Python and hosted either by Amazon Web Services or by a European Union (GDPR) compliant server (see below). Individual researchers can download data from their studies through the researcher interface, and Web-CDI administrators have access to the entire aggregate set of data from all studies run with Web-CDI. Website code is available in a GitHub repository at https://github.com/langcog/web-cdi, where interested users can browse, make contributions, and request technical fixes.

Data Privacy and GDPR Compliance

Web-CDI is designed to be compliant with stringent human subjects privacy protections across the world. First, for U.S. users, we have designed Web-CDI based on the United States Department of Health and Human Services “Safe Harbor” standard for collecting protected health information as defined by the Health Insurance Portability and Accountability Act (HIPAA). In particular, participant names are never collected, birth dates are used to calculate age in months (with no decimal information) but never stored, and geographic ZIP codes are trimmed to the first three digits. Because of the architecture of the site, even though participants enter ZIP codes and dates of birth, these are never transmitted in full to the Web-CDI server. Since no identifying information is being collected by the Web-CDI system, this feature ensures that Web-CDI can be used by United States labs without a separate Institutional Review Board agreement between users’ labs and Web-CDI (though of course researchers using the site will need Institutional Review Board approval of their own research projects).[1]

[1] Issues of de-identification and re-identifiability are complex and ever changing. In particular, compliance with DHHS “Safe Harbor” standards does not in fact fully guarantee the impossibility of statistical re-identification in some cases and if potential users have questions, we encourage them to consult with an Institutional Review Board.

In the European Union (EU), research data collection and storage is governed by the Generalized Data Protection Regulation (GDPR) and its local instantiation in the legal system of the member states. Some of the questions on the demographic form contain information that may be considered sensitive (e.g., information about children's developmental disorders), and in some cases, the possibility of linking this sensitive information to participant IDs exists, particularly when researchers draw on local databases that contain full names and addresses for recruitment and contacting. As a result, issues regarding GDPR compliance arise when transferring data outside the EU, namely to Amazon Web Services servers housed in the United States. Following GDPR regulations, these issues would make a data sharing agreement between data collectors and Amazon Web Services necessary. In addition, all administrators who can access the collected data would have to enter such an agreement, which needs updating whenever personnel changes occur.

To overcome these hurdles, and in consultation with data protection officers, we opted to leverage the local technical expertise and infrastructure to set up a sister site housed on GDPR-compliant servers, currently available at http://webcdi.mpi.nl. This site is updated synchronously with the main Web-CDI website to ensure a consistent user experience and access to the latest features and improvements. This site has been used in 135 successful administrations so far and is the main data collection tool for an ongoing norming study in the Netherlands. We are further actively advertising the option to use the European site to other labs who are following GDPR guidelines and are planning adaptations to multiple European languages, where copyright allows.

Current Data Collection

We now turn to an overview of the data collected thus far using Web-CDI. First, we examine the full sample of all of the Web-CDI administrations collected as of autumn 2020 (Dataset 1); we then focus in on a specific subset of Dataset 1 which is comprised of data from recent efforts to oversample non-white, less highly-educated U.S. participants (Dataset 2). Across both datasets, we show that general trends from prior research on vocabulary development are replicated using Web-CDI. Based on this work to date, we then discuss the potential for using Web-CDI to collect vocabulary development data from diverse communities online.

Dataset 1: Full Current Web-CDI Usage

In this section, we provide some preliminary analyses of Dataset 1, which consists of the full sample of American English Web-CDI administrations collected before autumn 2020. At time of writing, researchers from 15 universities in the United States have collected over 5,000 administrations of the American English CDI using Web-CDI since it was launched in late 2017, with 2,868 administrations of the WG form before exclusions and 3,565 administrations of the WS form before exclusions. We excluded participants from the subsequent analyses based on the following set of stringent criteria designed for the creation of future normative datasets. We excluded participants if it was not their first administration of the survey; they were born prematurely or had a birthweight under 5.5 lbs (< 2.5 kg); reported more than 16 hours of exposure to a language other than English per week on average (amounting to approximately > 10% of time during a week that a child hears another language than English); had serious vision impairments, hearing deficits or other developmental disorders or medical issues[2]; were outside of the correct age range for the survey; or spent less time on the survey than a pre-specified timing cut-off. Timing cut-offs were determined by selecting two studies within Dataset 1 that, upon a visual inspection, appeared to contain high-quality responses (i.e., did not contain a disproportionate number of extremely quick responders), and using these to estimate the 5th percentile of completion time by the child's age in months with a quantile regression (following a similar quantile regression method as Bleses, Makransky, Dale, Højen, & Ari, 2016). Thus, for each age on the WG and WS measures, we obtained an estimate of the 5th percentile of completion time and used this estimate as the shortest amount of time participants could spend on the Web-CDI without being excluded from our analyses here.

The exclusion criteria we used were designed to be generally comparable with those used in Fenson et al. (2007), who adopted stringent criteria to establish vocabulary norms that reflect typically developing children's vocabulary trajectories. A complete breakdown of the number of participants excluded on each criterion is in Table 1. Of the completed WG forms, 1,248 were excluded, leading to a final WG sample size of 1,620 administrations, and 1,665 WS administrations were excluded, leading to a final WS sample size of 1,900.

Demographic Distribution and Exclusions

Figure 3 shows the distribution of participant ethnicities in Dataset 1 as compared with previously reported numbers in the published norming study of the paper-based CDI form by Fenson et al. (2007). Several issues pertaining to sample representativeness are appreciable. First, as shown in Figure 3A, white participants comprised nearly three quarters of Dataset 1, which is comparable to U.S. Census estimates in

[2] Exclusions on the basis of child health were decided on a case-by-case basis by author V.M. in consultation with Philip Dale, Donna Thal, and Larry Fenson.

2019 of U.S. residents between the ages of 15 and 34 in 2019; however, Figure 3C shows that, compared with U.S. Census estimates, many more white participants in Dataset 1 were non-Hispanic than is true of the U.S. population in general, indicating that Web-CDI is significantly oversampling white, non-Hispanic individuals (the breakdown of white participants into Hispanic and non-Hispanic is not reported in the 2007 norms). Moreover, few participants identified as Hispanic/Latinx: 6.4% of WG participants and 5.2% of WS participants reported Hispanic or Latinx heritage. The low percentage of Hispanic/Latinx participants was due in part to our exclusion of children with substantial exposure to languages other than English: before exclusions, 8.4% of WG participants were Hispanic/Latinx, and 8.2% of WS participants were Hispanic/Latinx. Finally, representation of Black participants is generally lower in Dataset 1 (3.5%) than in the 2007 norms (10.5%), which is in turn lower than U.S. Census estimates (15.2%). This indicates that both Web-CDI data and existing norming samples tend to substantially underrepresent Black participants.

Table 1. ***Exclusions from Dataset 1: full Web-CDI sample***

Exclusion	WG exclusions	% of full WG sample excluded	WS exclusions	% of full WS sample excluded
Not first administration	163	5.68%	444	12.45%
Premature or low birthweight	37	1.29%	67	1.88%
Multilingual exposure	449	15.66%	492	13.80%
Illnesses/Vision/Hearing	191	6.66%	203	5.69%
Out of age range	88	3.07%	199	5.58%
Completed survey too quickly	319	11.12%	256	7.18%
System error in word tabulation	1	0.03%	4	0.11%
Total exclusions	1248	44%	1665	47%

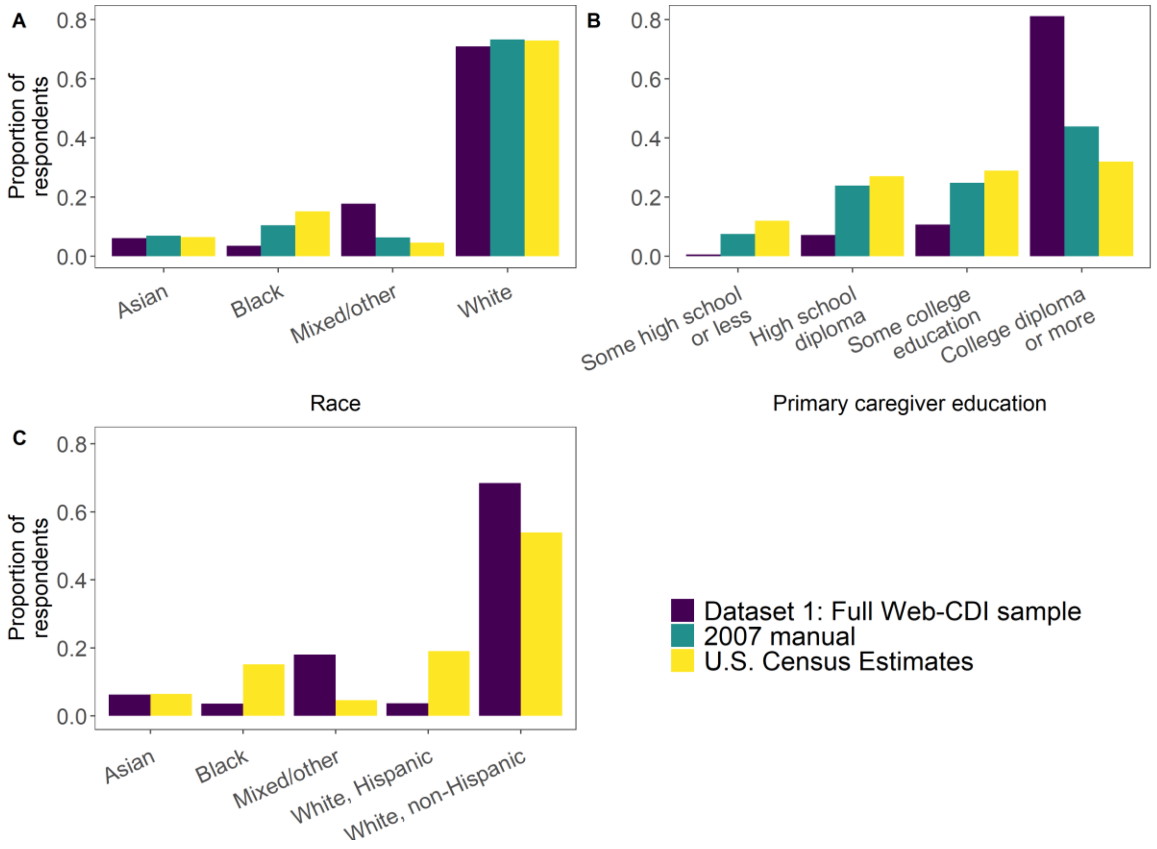


Figure 3. ***Top row: Proportion of respondents plotted by child race (A) and educational level of primary caregiver (B) from full Web-CDI sample (Dataset 1) to date (N = 3,520), compared with norming sample demographics from Fenson (2007) and U.S. Census data (American Community Survey, 2019; National Center for Education Statistics, 2019). Bottom row (C): Participant breakdown by race in Dataset 1 as compared with U.S. Census data, splitting white participants into those who are Hispanic and those who are not.***

Participants' educational attainment level, as measured by the primary caregiver's highest educational level reached[3], was similarly skewed. In Dataset 1, 81.2% of

[3] Maternal education level is a common measure of family socioeconomic status; we probe primary caregiver education level here to accommodate family structures in which child-rearing may not

responses came from families with college-educated primary caregivers compared to 43.8% from the same group in the 2007 norms and 32.0% (Figure 3) of adults 25 and older according to the U.S. National Center for Education Statistics in 2017. Furthermore, fewer than 1% of participants report a primary caregiver education level less than a high school degree, compared to 7% from the same group in the 2007 norms.

The overrepresentation of white, non-Hispanic Americans and those with high levels of education attainment points to a general challenge encountered in vocabulary development research, which we return to when we detail our efforts to recruit more diverse participants. Figure 4 shows that, of the recruitment methods used in Dataset 1, the studies conducted using the platform Prolific (which we detail in the Dataset 2 section) contributed the least to the high proportion of white, non-Hispanic, college educated participants. Respondents not known to be recruited through an online channel or crowdsourcing platform (labelled "Other method" in Figure 4) showed the most overrepresentation of white, college educated participants, suggesting that reliance on university convenience samples may be driving the demographic skewness of Dataset 1 most acutely.

Results: Dataset 1

Although the CDI instruments include survey items intended to measure constructs other than vocabulary size, such as gesture, sentence production, and grammar, we focus exclusively on the vocabulary measures here. We also visualize key analyses from Dataset 1 alongside the analogous analyses on the American English CDI administrations from the Wordbank repository (Frank et al., 2021) that include the relevant demographic information needed to provide a comparison dataset of traditional paper-and-pencil forms. Across both the WG and WS measures, Dataset 1 shows greater reported vocabulary comprehension and production for older children. Moreover, data from both the WG and WS measures in Dataset 1 replicate a subtle but reliable pattern such that female children tend to have slightly larger vocabulary scores than male children across the period of childhood assessed in the CDI forms (Frank et al., 2021), though in these data this difference does not appear until around 18 months (Figure 5).

primarily be the responsibility of the child's mother, but we expect that in the vast majority of cases this corresponds to the child's mother.

On the WG form, respondents' reports of children's vocabulary comprehension and production both increased with children's age (Figure 6). We replicate overall patterns found by Feldman et al. (2000) in that, on both the "Words Understood" measure (in which caregivers indicate which words their child "understands") and the "Words Produced" measure (in which caregivers indicate which words their child "understands and says"), vocabulary scores were slightly negatively correlated with primary caregivers' education level, such that those caregivers without any college education reported higher vocabulary scores on both scales; on the word comprehension scale, this was particularly the case for the youngest infants in the sample. A linear regression model with robust standard errors predicting comprehension scores with children's age and primary caregivers' education level (binned into categories of "High school diploma or less," "Some college education" and "College diploma or more"[4]) as predictors shows main effects of both age ($\beta = 20.05$, $p < 0.001$) and caregiver primary education ($\beta_{\text{highschool}} = 21.86$, $p = 0.05$). Similarly, a linear regression model with robust standard errors predicting production scores by children's age and primary caregivers' education level shows main effects of age ($\beta = 7.60$, $p < 0.001$) and primary caregiver education ($\beta_{\text{highschool}} = 20.46$, $p = 0.008$). These analyses were not preregistered, but generally follow the analytic strategy in Frank et al. (2021); additionally, we fit linear models with robust standard errors to account for heteroskedasticity in the data (Astivia & Zumbo, 2019). Generalized linear model predictions for Web-CDI shown in Figure 6 differ somewhat from those for Wordbank; prediction curves for caregivers of different education attainment levels diverge slightly more in the Web-CDI sample than in the Wordbank sample.

The pattern of results seen in the WG subsample of Dataset 1 is consistent with prior findings indicating that respondents with lower levels of education attainment report higher vocabulary comprehension and production on the WG form (Feldman et al., 2000; Fenson et al., 1994). However, although caregivers with lower levels of education attainment report higher mean levels of vocabulary production and comprehension, median vocabulary scores (which are more robust to outliers) show no clear pattern of difference across primary caregiver education levels (Figure 7). This discrepancy between the regression effects and a group-median analysis suggests that the regression effects described previously are driven in part by differential interpretation of the survey items, such that a few caregivers with lower levels of education attainment are more liberal in reporting their children's production and

[4] "High school diploma or less" corresponds to 12 or fewer years of education; "Some college" corresponds to 13-15 years of education; "College diploma or more" refers to 16 or more years of education.

comprehension vocabulary scores, especially for the youngest children, driving up the mean scores for this demographic group.

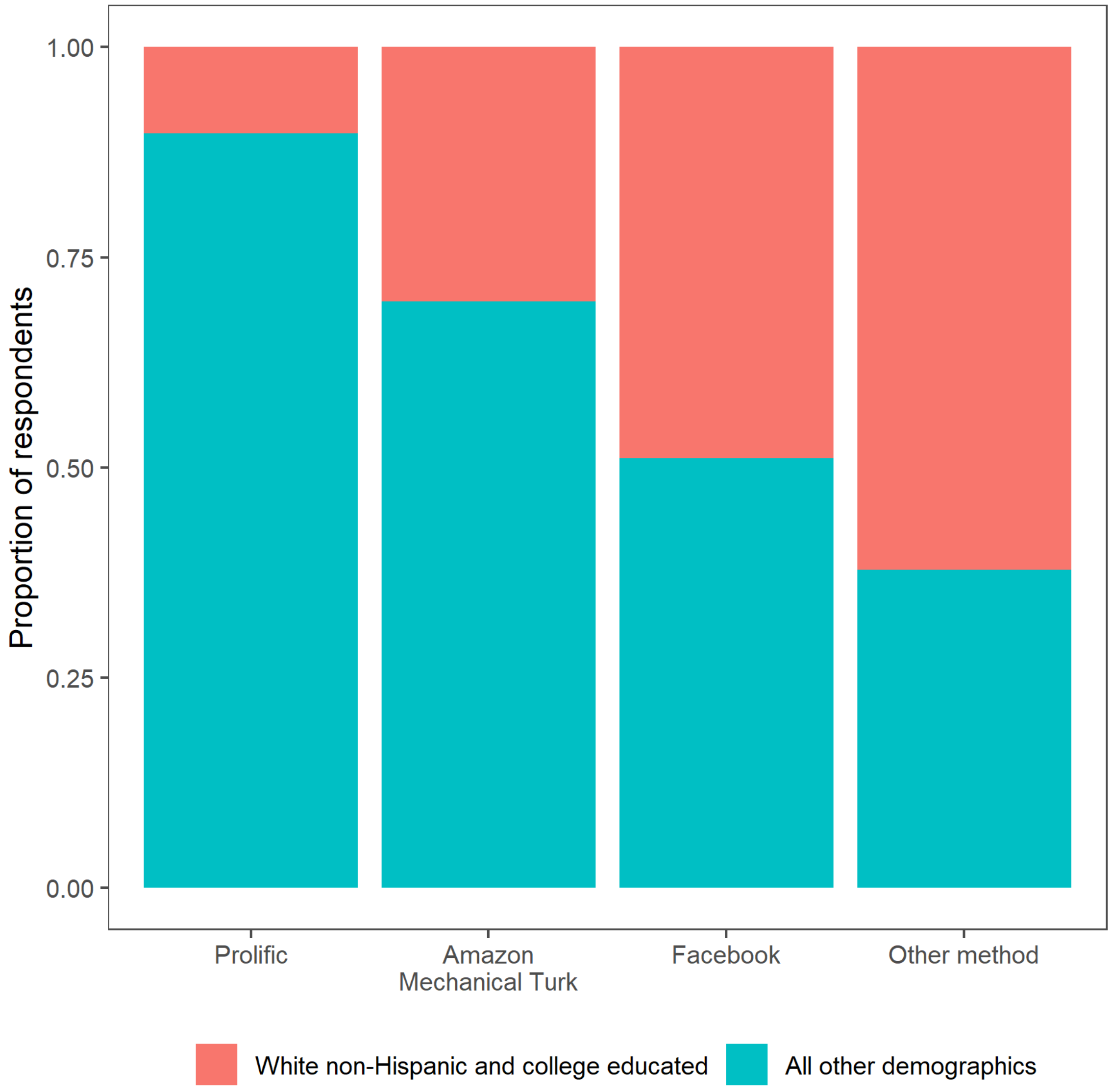


Figure 4. ***Proportion of participants from Dataset 1 who were white, college educated and not Hispanic, plotted by recruitment method.***

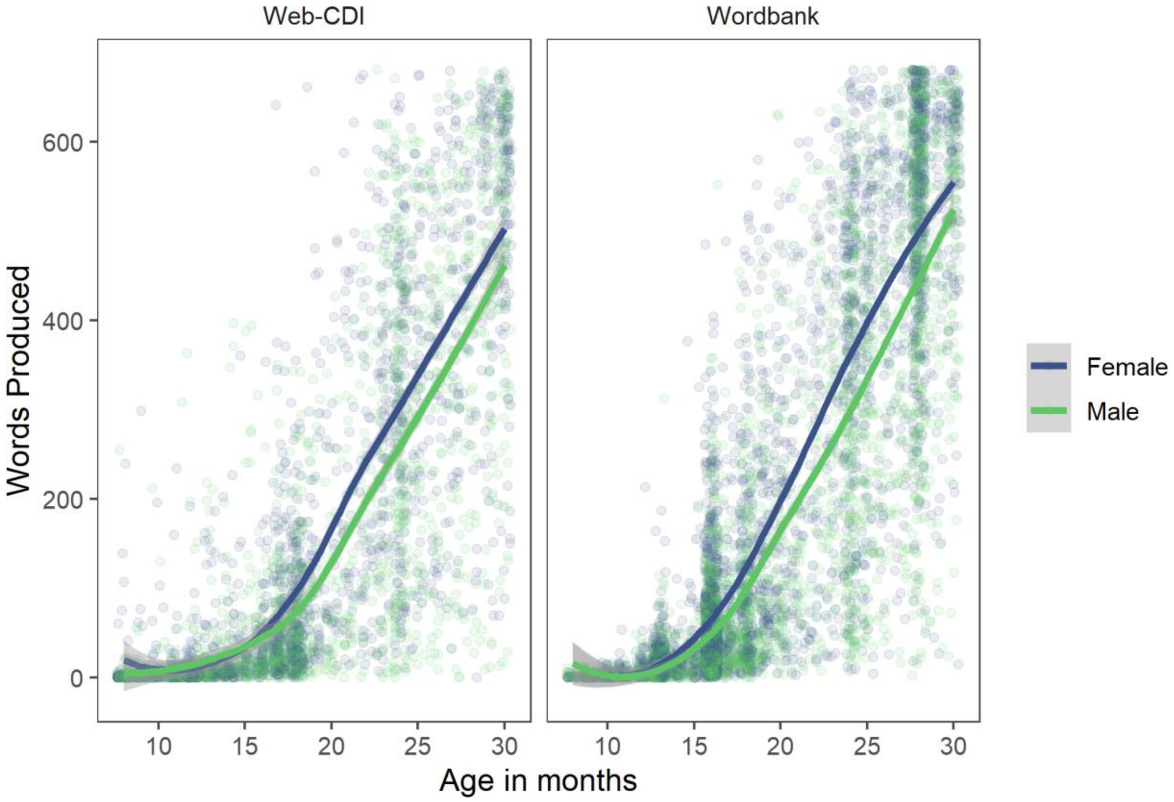


Figure 5. ***Individual children's vocabulary production scores plotted by children's age and sex (both WG and WS). Left panel: Dataset 1 (full sample of Web-CDI administrations, N = 3,510, with 1,673 girls). Right panel: American English CDI administrations in the Wordbank repository (Frank et al., 2021), including only those administrations for which the child's sex was available (N = 6,486, with 3,146 girls). Lines are locally weighted regressions (LOESS) with associated 95% confidence intervals. Children with a different or no reported sex (N = 10) are omitted here.***

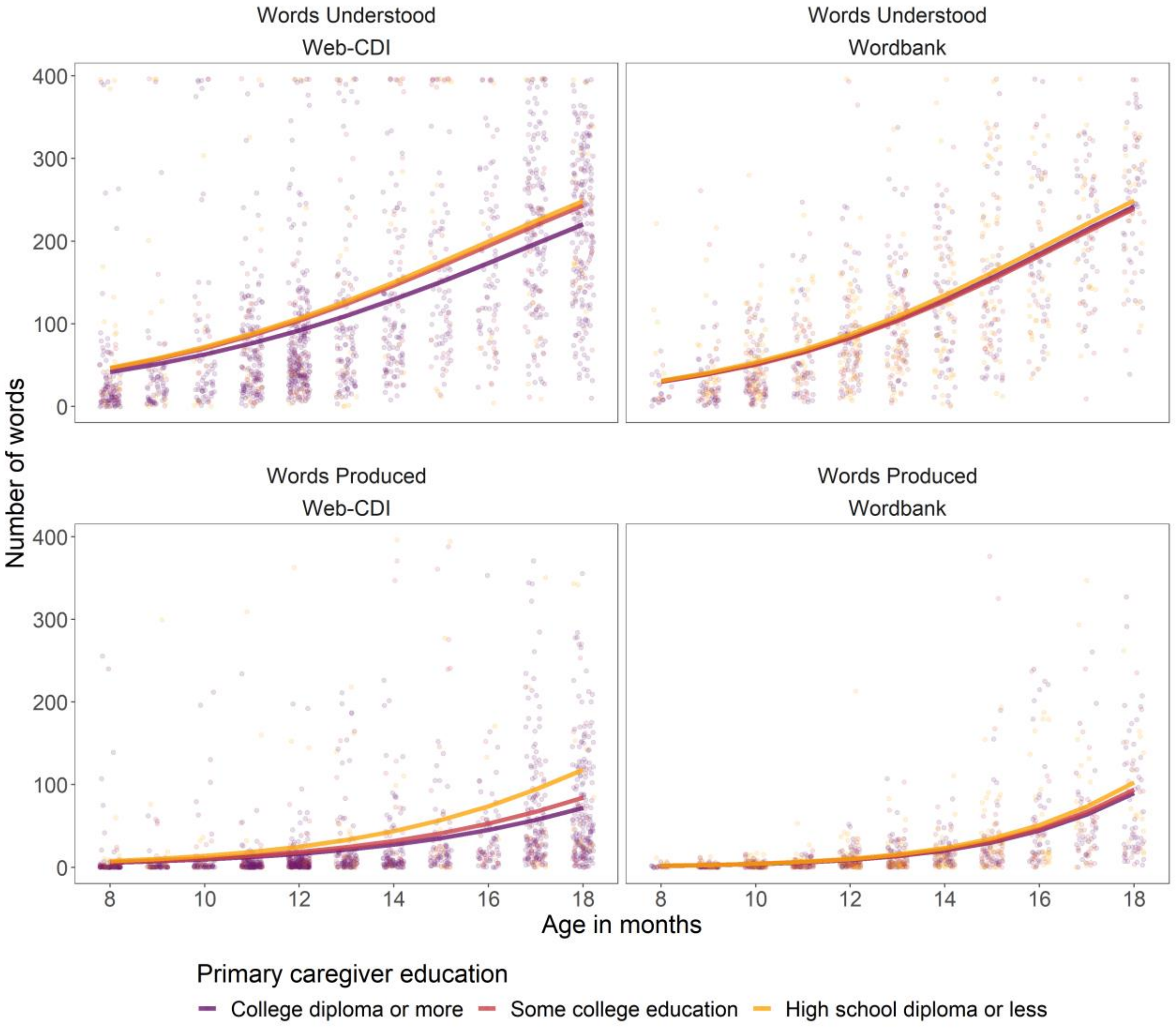


Figure 6. ***Individual children's word production (top panels) and comprehension (bottom panels) scores from Dataset 1 (full Web-CDI sample) plotted by age and primary caregiver's level of education (binned into "High school diploma or less," "Some college education," and "College diploma or more"). Left panels show results from the sample of WG Web-CDI administrations collected as of November 2020 (N = 1,620), and right panels show the subset of American English administrations from Wordbank (Frank et al., 2021) that contain information about caregiver education (N = 1,068) for comparison. Curves show generalized linear model fits.***

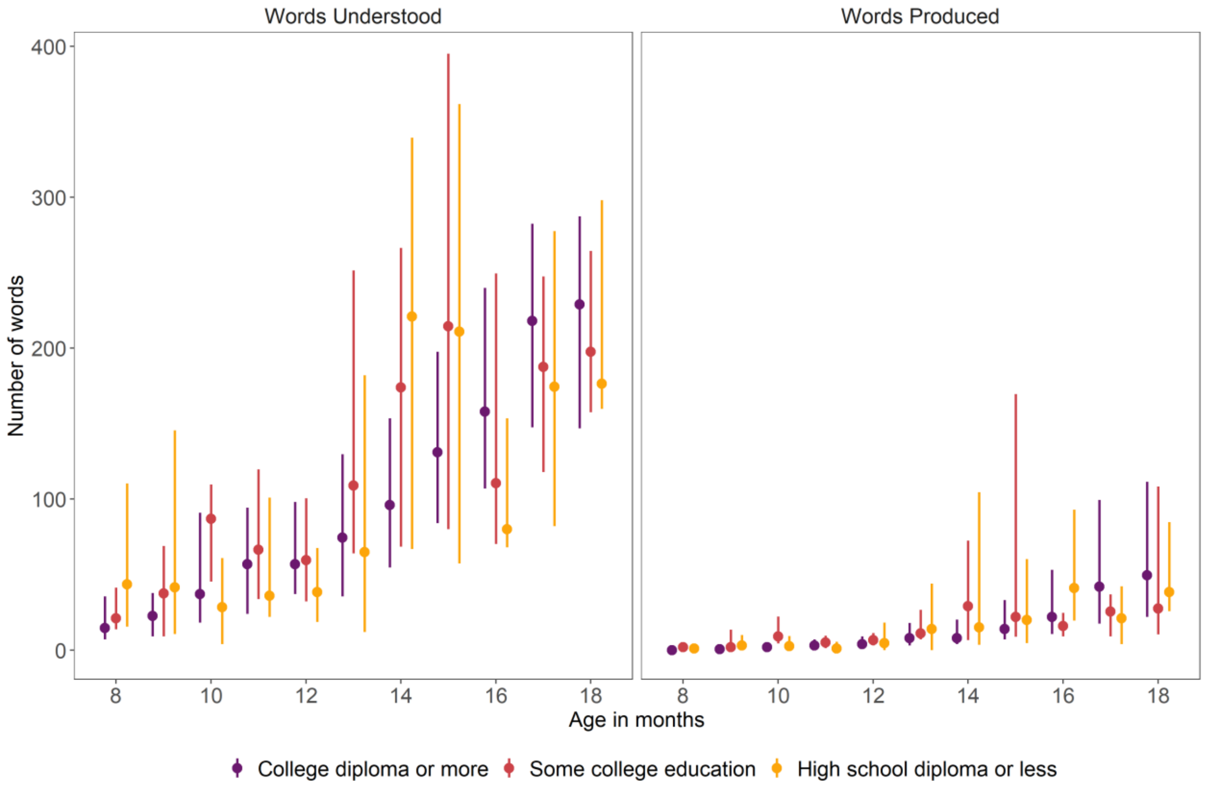


Figure 7. ***Median vocabulary comprehension (left) and production (right) scores from Dataset 1 (full Web-CDI sample) by age and primary caregiver's level of education attainment on the WG form. Lines indicate span between first and third quartiles for each age.***

Vocabulary production scores on the WS form show the expected pattern of increase with children's age in months; in addition, scores replicate the trend reported in Feldman et al. (2000) and Frank et al. (2021) such that primary caregiver education is positively associated with children's reported vocabulary size (Figure 8). Because representation of caregivers without a high school diploma is scarce (N = 6 out of a sample of 1,900), interpretation of the data from this group is constrained. Nevertheless, as shown in Figure 8, a small but clear positive association between primary caregiver education and vocabulary score exists such that college-educated caregivers report higher vocabulary scores than those of any other education level. Notably, this association is not the result of outliers and is still appreciable in median scores (Figure 9), unlike the data from the WG measure shown in Figure 7. The implications from these data converge with previous findings which indicate that parental education levels, often used as a metric of a family's socioeconomic status, are related to children's vocabulary size through early childhood.

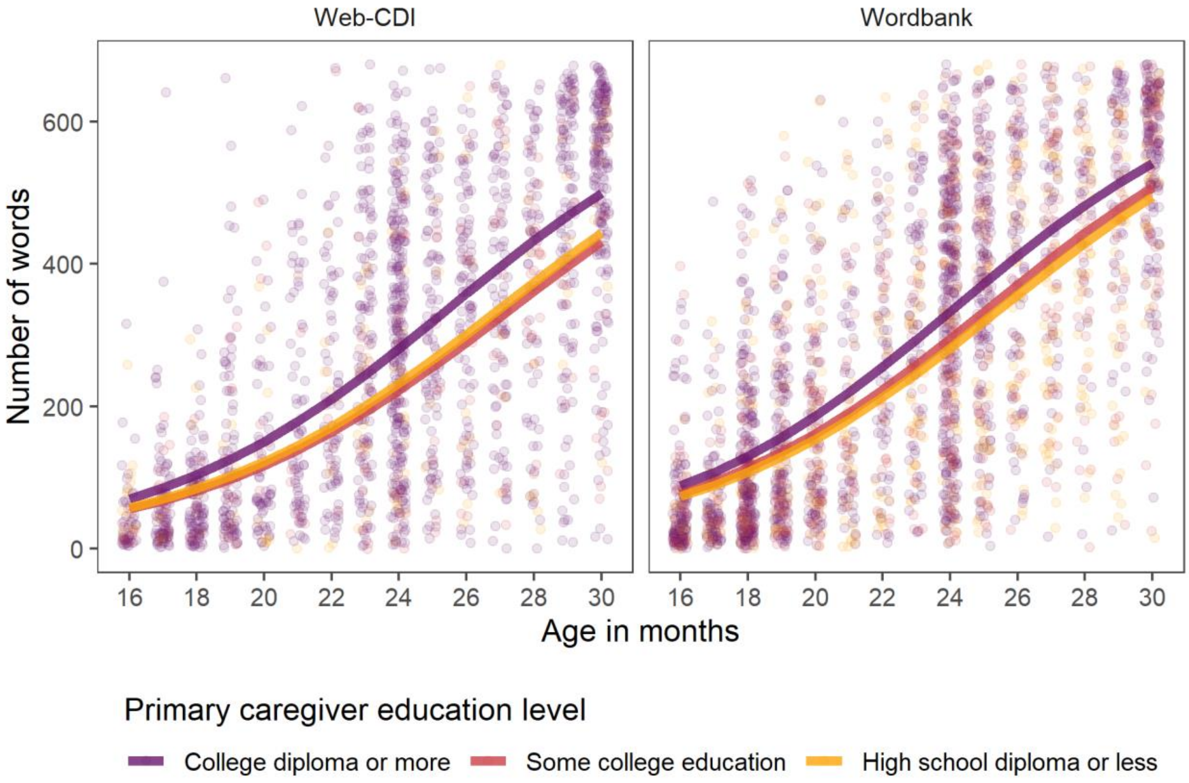


Figure 8. ***Individual children's vocabulary production scores from Dataset 1 (full Web-CDI sample) plotted by children's age and primary caregiver education level as reported in the sample of WS Web-CDI administrations collected as of November 2020 (N = 1,900, left panel) and in the Wordbank repository (N = 2,776, right panel). Curves show generalized linear model fits.***

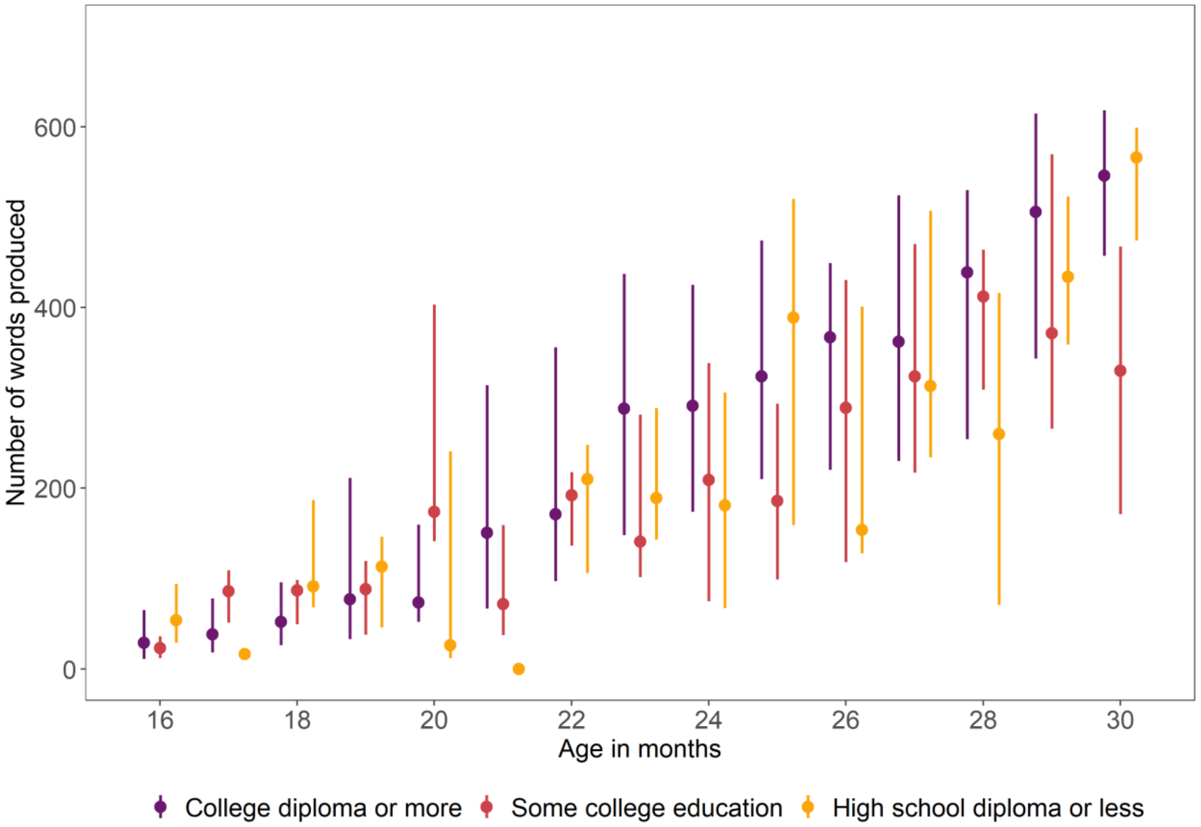


Figure 9. ***Median vocabulary production scores from Dataset 1 (full Web-CDI sample) by age and primary caregiver's level of education attainment on the WS form. Lines indicate span between first and third quartiles for each age.***

Discussion: Dataset 1

In general, the full sample of Web-CDI data after exclusions (Dataset 1) replicates previous norming datasets used with the standard paper-and-pencil form of the MB-CDI. We find that vocabulary scores grow with age and that females hold a slight advantage over males in early vocabulary development. Moreover, Dataset 1 replicates a previously documented relationship between primary caregiver education level and vocabulary scores: on the WG form, primary caregiver education shows a slight negative association with vocabulary scores, whereas the trend is reversed in the WS form. Taken together, these data illustrate that Web-CDI and the standard paper-and-pencil form of the CDI give similar results, and thus that Web-CDI can be used as a valid alternative to the paper format.

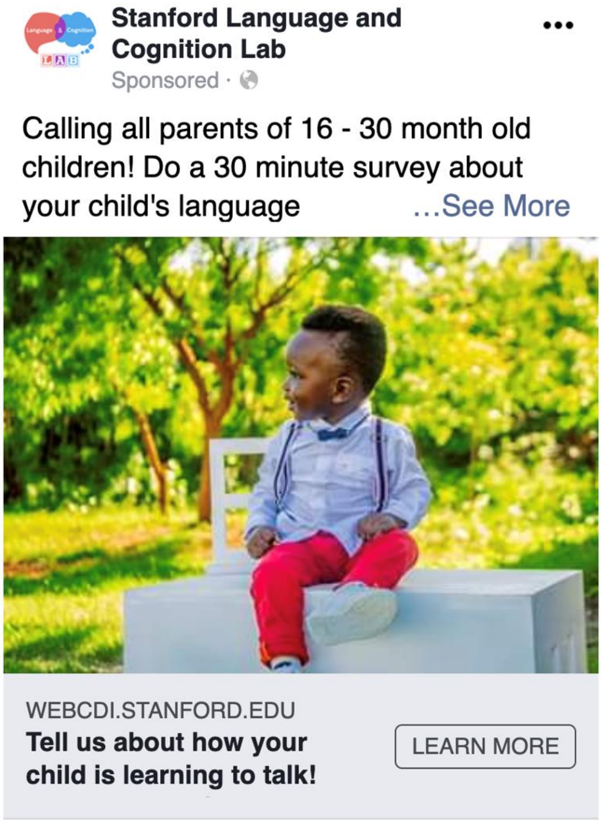
The data discussed above have resulted from efforts by many researchers across the United States whose motivations for using the Web-CDI vary. As a result, they reproduce many of the biases of standard U.S. convenience samples. In the next section, we describe in more detail our recent efforts to use the Web-CDI to collect vocabulary development data from traditionally underrepresented participant populations in the United States, attempting to counteract these trends.

Dataset 2: Using Web-CDI to Collect Data from Diverse U.S.-based Communities

Despite the large sample sizes we achieved in the previous section, Dataset 1 is, if anything, even more biased towards highly-educated and white families than previous datasets collected using the paper-and-pencil form. How can we recruit more diverse samples to remedy this issue? Here, we discuss and analyze Dataset 2, which consists of those administrations from Dataset 1 which were part of recent data-collection efforts (within the past year and a half) that were specifically aimed towards exploring the use of online recruitment as a potential way to collect more diverse participant samples than are typical in the literature. In other words, the following data from Dataset 2 were included in the previous discussion and analysis of Dataset 1, but we examine them separately here to give special attention to the issue of collecting diverse samples online.

Online Data Collection

Online recruitment methods, such as platforms like Amazon Mechanical Turk, Facebook and Prolific, represent one possible route towards assembling a large, diverse sample. These methods allow researchers to depart from their typical geographical recruitment area much more easily than with paper-and-pencil administration. Online recruitment strategies for vocabulary development data collection have been used in the United Kingdom (Alcock, Meints, & Rowland, 2020), but their usage in the U.S. context remains, to our knowledge, rare. In a series of data collection efforts, we used Web-CDI as a tool to explore these different channels of recruitment.



Stanford Language and Cognition Lab

Sponsored ·

Calling all parents of 16 - 30 month old children! Do a 30 minute survey about your child's language …See More

WEBCDI.STANFORD.EDU

Tell us about how your child is learning to talk!

Figure 10. ***Example Facebook advertisement in Phase 1 of recent data collection.***

Dataset 2 consists of data that were collected in two phases. In the first phase, we ran advertisements on Facebook which were aimed at non-white families based on users' geographic locations (e.g., targeting users living in majority-Black cities) or other profile features (e.g., ethnic identification, interest in parenthood-related topics). Advertisements consisted of an image of a child and a caption informing Facebook users of an opportunity to fill out a survey on their child's language development and receive an Amazon gift card (Figure 10). Upon clicking the advertisement, participants were redirected to a unique administration of the Web-CDI; they received $5 upon completing the survey. This open-ended approach to recruitment offered several advantages, namely that a wide variety of potential participants from specific demographic backgrounds can be reached on Facebook. However, we also received many incomplete or otherwise unusable survey administrations, either from Facebook users who clicked the link and decided not to participate, or those who completed the survey in an extremely short period of time (over half of all completed administrations, Table 2).

In the second phase, we used the crowdsourcing survey vendor Prolific (http://prolific.co) in the hopes that some of the challenges encountered with Facebook recruitment would be addressed. Prolific allows researchers to create studies and post them to individuals who are in the platform's participant database, each of whom is assigned a unique alphanumeric "Prolific ID." Importantly, Prolific maintains detailed demographic information about participants, allowing researchers to specify who they would like to complete their studies. Prolific further has a built-in compensation infrastructure that handles monetary payments to participants, eliminating the need to disburse gift cards through Web-CDI.

Table 2. ***Exclusions from Dataset 2: Recent data collection using Facebook and Prolific***

Exclusion	WG exclusions	% of full WG sample excluded	WS exclusions	% of full WS sample excluded
Not first administration	0	0.00%	0	0.00%
Premature or low birthweight	7	2.53%	1	0.33%
Multilingual exposure	18	6.50%	23	7.62%
Illnesses/Vision/Hearing	4	1.44%	4	1.32%
Out of age range	1	0.36%	26	8.61%
Completed survey too quickly	119	42.96%	133	44.04%
System error in word tabulation	0	0.00%	0	0.00%
Total exclusions	149	54%	187	62%

In the particular case of Web-CDI, the demographic information needed to determine whether an individual was eligible to complete our survey (e.g., has a child in the correct age range, lives in a monolingual household, etc.) was more specific than the information that Prolific collects about their participant base. We therefore used a brief pre-screening questionnaire to generate a list of participants who were eligible to participate, and subsequently advertised the Web-CDI survey to those participants. Given that we were interested only in reaching participants in the United States who were not white or who did not have a college diploma, our data collection efforts only yielded a sample that was small (N = 68) but much more thoroughly screened than that which we could obtain on Facebook.

Across both phases (Facebook and Prolific recruitment), we used the same exclusion criteria as in the full Web-CDI sample to screen participants. A complete tally of all excluded participants is shown in Table 2. In both the WG and WS surveys, exclusion rates in Dataset 2 were high, amounting to 58% of participants who completed the survey. The high exclusion rates were notably driven by an accumulation of survey administrations which participants completed more quickly than our time cutoffs allow (Tables A4 and A5). Many of the survey administrations excluded for fast completion also had missing demographic information reported: Among WG participants excluded for too-fast completions, 93% did not report ethnicity, and among WS participants excluded for the same reason, 97% did not report ethnicity. Absence of these data prevents us from drawing conclusions about the origin or demographic profile of administrations that were excluded. After exclusions, full sample size in Dataset 2 was N = 128 WG completions and N = 115 WS completions.

Results: Dataset 2

The results from Dataset 2 show overall similar patterns to the full Web-CDI sample in several regards. Word production scores from both the WG and WS administrations reflect growing productive vocabulary across the second and third years, with a very small sex effect such that female children's vocabularies are higher across age than males' (Figure 11). The relationship between caregivers' reported levels of education and child's vocabulary score is not as clear as it is in the full Web-CDI sample (Figure 12); however, children of college-educated caregivers reported generally higher vocabulary scores across age than did children of caregivers without any college degree. These patterns suggest that our data show similar general patterns to other CDI datasets with other populations (Frank et al., 2021).

Importantly, Dataset 2 showed a substantial improvement in reaching non-white or less highly-educated participants. After exclusions, Dataset 2 has a higher proportion of non-white participants than Dataset 1 (the overall Web-CDI sample) and the norms established by Fenson et al. (2007) (Figure 13). Black participants in particular showed a marked increase in representation, from 10.5% in the 2007 norms to 30.7% in Dataset 2, while the proportion of white participants decreased from 73.3% in the 2007 norms to 50.5% in Dataset 2. Representation on the basis of families' reported primary caregiver education also improved (Figure 13). Participants with only a high school diploma accounted for 33.3% of Dataset 2 as compared to 23.8% in the 2007 norms, and representation of those with a college diploma or more education decreased from 43.8% in the 2007 norms to 36.2% in Dataset 2. Notably, the distribution of Dataset 2 with regard to primary caregiver education level is quite similar to Kristoffersen et al. (2013), who collected a large, nationally-representative sample of CDI responses in Norway and obtained a sample with 30%, 42%, and 24% for participants reporting 12, 14-16, and 16+ years of education, respectively.

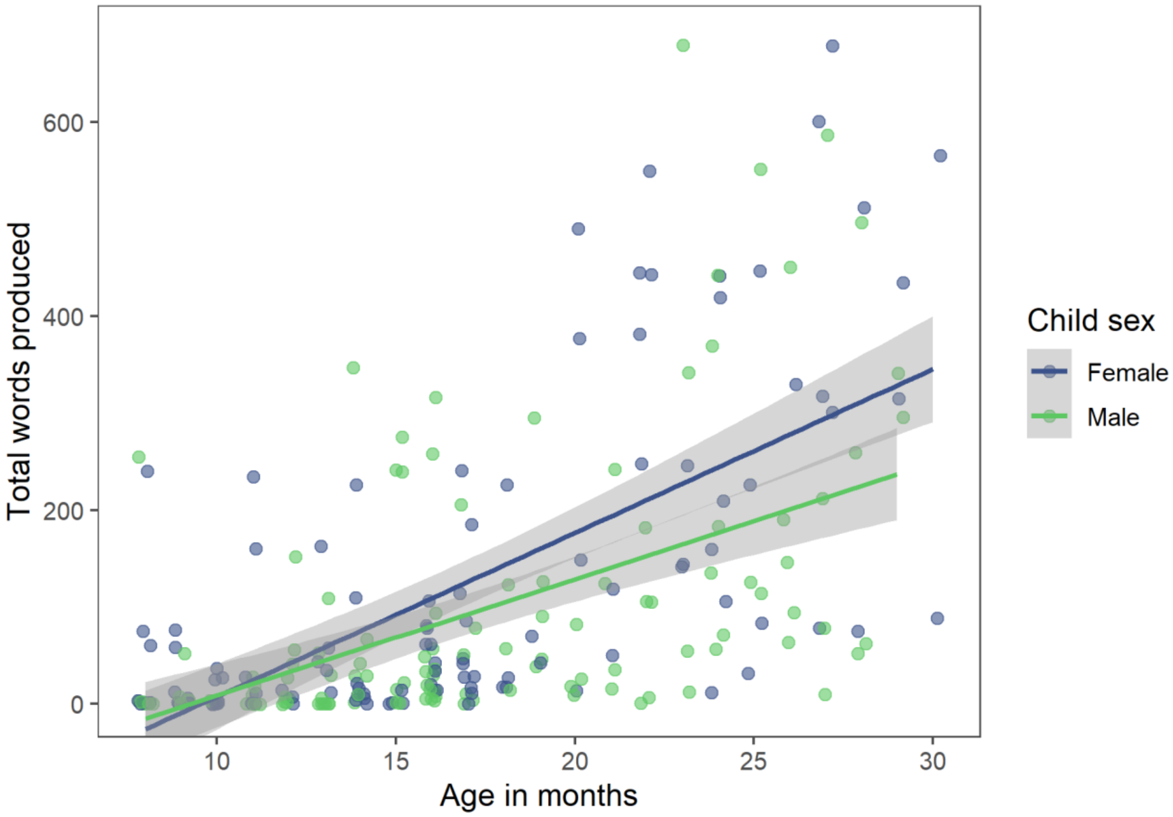


Figure 11. ***Individual children's vocabulary production scores from Dataset 2 (recent data collection efforts) plotted by children's age and sex (both WG and WS, N = 240, with 114 girls). Lines are best linear fits with associated 95% confidence intervals. Children with a different or no reported sex (N = 3) are omitted here.***

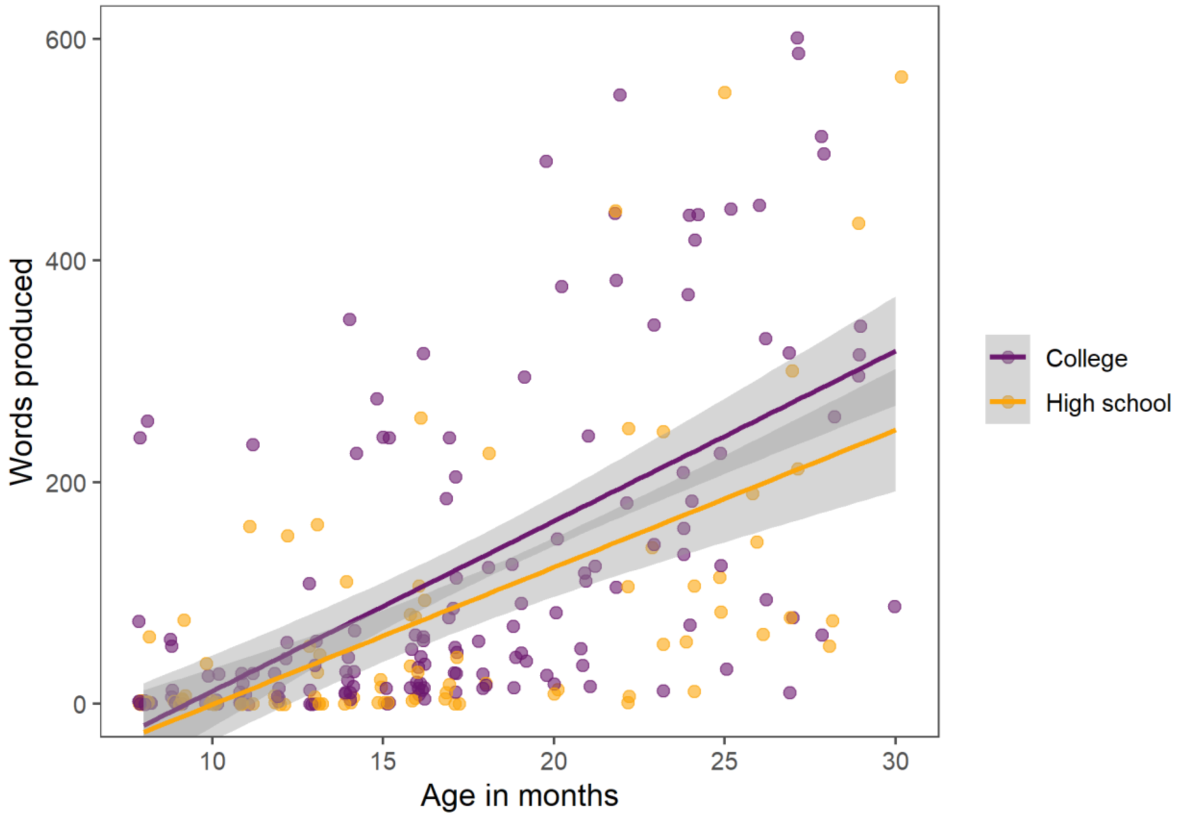


Figure 12. ***Individual children's vocabulary production scores from Dataset 2 (recent data collection efforts) plotted by age and level of primary caregiver education, binned into those with a high school diploma or less education and those with some college education or a college diploma (N = 243). Lines show best linear fits and associated 95% confidence intervals.***

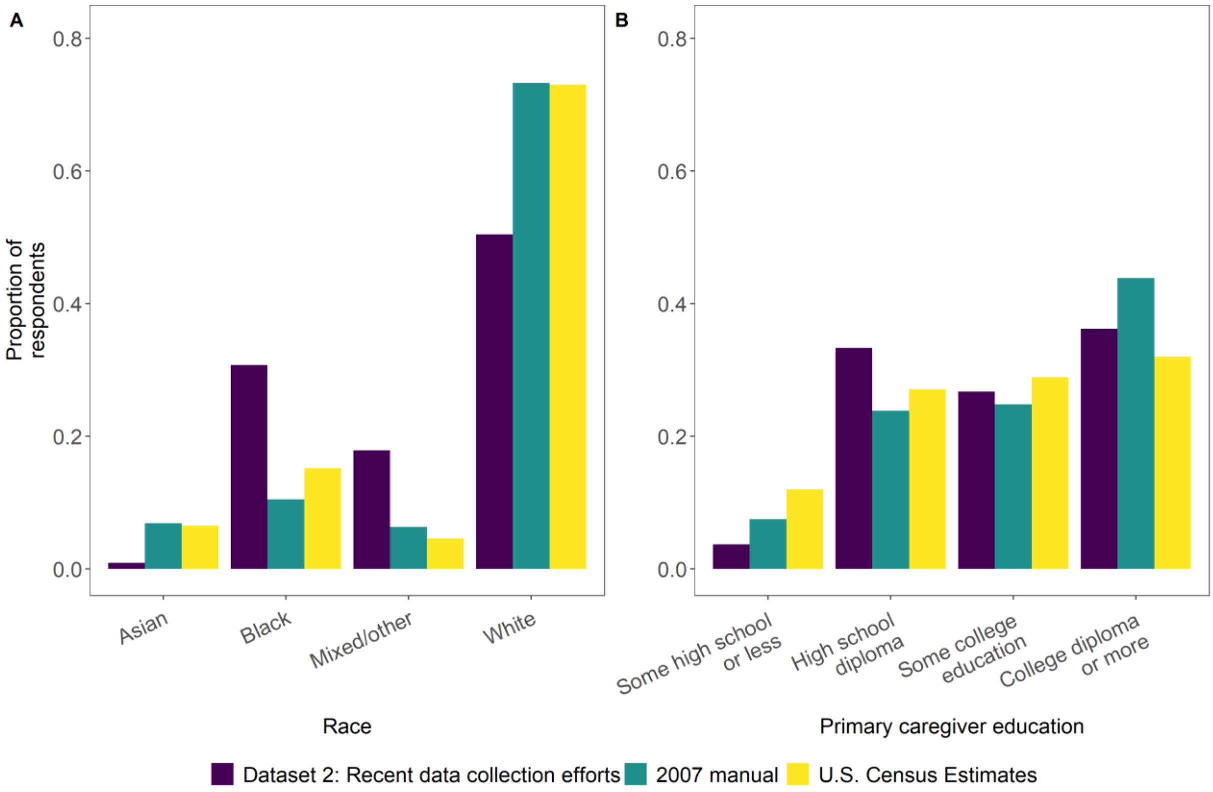


Figure 13. ***Proportion of respondents plotted by child race (A) and educational level of primary caregiver (B) from Dataset 2, recent data collection efforts aimed towards oversampling non-white, less highly-educated families (N = 243), compared with norming sample demographics from Fenson (2007). Latinx participants can be of any race and are thus not represented as a separate category here.***

Discussion: Dataset 2

The results from Dataset 2 indicate that Web-CDI could be a promising platform to collect vocabulary development data in non-white populations and communities with lower levels of education attainment when paired with online recruitment methods that yield legitimate, representative participant samples. At the same time, however, these data convey clear limitations of our approach. Perhaps most conspicuously, more than half of completed administrations in this sample had to be excluded, in many cases because the information provided by participants appeared rushed or incomplete: over 40% of administrations were completed in a shorter amount of time than that allowed by our cut-off criteria (Tables A4 and A5), and of these quick completions, well over 90% were missing demographic information that is rarely missing in other administrations of the form. Determining the precise reasons for the high exclusion rate, and how (if at all) this (self-)selection may bias data reflecting demographic trends in vocabulary development, requires a more thorough assessment of who is submitting hastily-completed forms. Such an assessment is beyond the scope of the current study. However, all respondents who got to the end of the form were compensated regardless of how thoroughly they completed it, creating the possibility that some participants who clicked the anonymous link may not have been members of the population of interest, but rather were other individuals motivated by compensation. To the extent that participants moved through the form quickly because they found the length burdensome, a transition to short forms, including computer adaptive ones (e.g., Chai, Lo, & Mayor, 2020; Kachergis et al., 2021; Makransky, Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019), would potentially increase data quality and completion rates substantially.

Additionally, the exclusion rates described previously provide information only on those participants who did, at some point, submit a completed form, but many individuals clicked the advertisement link and did not subsequently continue on to complete the form. Without an in-depth exploration of who is clicking the link and why they might choose not to continue, we cannot draw conclusions about the representativeness of the sample in Dataset 2 with regard to the communities we would like to include in our research. As such, a more thorough understanding of how users from different communities respond to various recruitment and sampling methods is needed in future work in order to draw conclusions about demographic trends above and beyond those already established in the literature.

Participants in Dataset 2 were recruited through a targeted post on social media, a technique that is considerably more anonymous than recruitment strategies which entail face-to-face or extended contact between researchers and community members. Online recruitment methods may not be suitable for all communities, especially when researchers ask participants to report potentially sensitive information about the health, developmental progress, ethnicity and geographic location of their children (even when such information is stored anonymously). Our goal here was to assess whether general trends in past literature could be recovered using such an online strategy, but future research should take into account that other more personal methods of recruitment, such as direct community outreach or liaison contacts, may improve participants' experiences and their willingness to engage with the study. Furthermore, despite the many invalid responses we received in this study, it may nevertheless be possible to use social media to recruit interested participants using a more rigorously-vetted approach. For example, participants could respond to an ad to be entered into a database and be sent study links later, rather than receiving a study link immediately after seeing the ad.

An additional limitation of Dataset 2 is that it only examines vocabulary development in monolingual children. While understanding that the performance of standard measurement tools like the CDI among multilinguals is of immense import to the field of vocabulary development research (Gonzalez et al., in prep; Floccia et al., 2018; De Houwer, 2019), we focused in Dataset 2 only on vocabulary development in monolingual children, because collecting data from multilingual populations introduces additional methodological considerations (e.g., how to measure exposures in each language) that are not the focus of our work here. However, it will be imperative in future to collect large-scale datasets of vocabulary data in bilingual children, both to better calibrate standard tools such as the CDI, as well as to reduce the bias towards monolingual families in the existing literature on measuring vocabulary development.

Finally, a significant limitation of the data collection process in Dataset 2 is that many people in the population of interest - particularly lower-income families - do not have reliable internet access. Having participants complete the Web-CDI on a mobile device may alleviate some of the issues caused by differential access to Wi-Fi, since the vast majority of American adults own a smartphone (Pew Research Center, 2019). Accordingly, improving Web-CDI's user experience on mobile platforms will be an important step towards ensuring that caregivers across the socioeconomic spectrum can easily complete the survey. For smartphone users on pay-as-you-go plans, who may be reluctant to use phone data to complete a study, a possible solution could be compensating participants for the amount of "internet time" they incurred completing the form.

General Discussion and Conclusions

In this paper, we have presented Web-CDI, a comprehensive online interface for researchers to measure children's vocabulary by administering the MacArthur-Bates Communicative Development Inventories family of parent-report instruments. Web-CDI provides a convenient researcher management interface, built-in data privacy protections, and a variety of features designed to make both longitudinal and social-media sampling easy. To date, over 3,500 valid administrations of the WG and WS forms have been collected on Web-CDI from more than a dozen researchers in the United States after applying strict exclusion criteria derived from previous norming studies (Fenson et al., 1994, 2007). Our analysis of Dataset 1 shows that demographic trends from previous work using the paper-and-pencil CDI form are replicated in data gleaned from Web-CDI, suggesting that the Web-CDI is a valid alternative to the paper form and captures similar results.

Many research laboratories, not only in the United States but around the world, collect vocabulary development data using the MacArthur-Bates CDI in its original or adapted form. With traditional paper-based forms, combining insights from various research groups can prove challenging, as each group may have slightly different ways of formatting and managing data from CDI forms. By contrast, if all of these groups' data come to be stored in a single repository with a consistent database structure, data from disparate sources can easily be collated and analyzed in a uniform fashion. As such, a centralized repository such as Web-CDI provides a streamlined data-aggregation pipeline that facilitates cross-lab collaborations, multisite research projects, and the curation of large datasets that provide more power to characterize the vast individual differences present in children's vocabulary development.

Beyond the goal of simply getting more data, we hope that Web-CDI can advance efforts to expand the reach of language development research past convenience samples into diverse communities. A key question in the field of vocabulary development concerns the mechanisms through which sociodemographic variables, such as race, ethnicity, income, and education are linked to group differences in vocabulary outcomes. Large, population-representative samples of vocabulary development data are needed to understand these mechanisms, but research to date (including the full sample of Web-CDI administrations) has often oversampled non-Hispanic white participants and those with advanced levels of education.

We explored the use of Web-CDI as part of a potential strategy to collect data from non-white and less highly-educated communities in two phases (Dataset 2). Several overall patterns emerged which we expected: vocabulary scores grew with age, providing a basic validity check of the Web-CDI measure; females held a slight advantage in word learning over males; and children of caregivers with a college education showed slightly higher vocabulary scores. Nonetheless, the insights from these data, while aligned with past norming studies, are necessarily constrained by several features of our method.

Limitations of our method notwithstanding, a transition to web-based data collection streamlines the process by which historically underrepresented populations can be reached in child language research. In particular, recruitment methods involving community partners, such as parenting groups, childcare centers and early education providers, are simplified substantially if leaders in these organizations can distribute a web survey to their members that is easy to fill out, as compared with paper forms, which typically present logistical hurdles for distribution and collection. Additionally, we hope that Web-CDI can serve as an accessible, free, and easy to use resource for researchers already doing extensive work with underrepresented groups.

Web-based data collection can capture useful information about vocabulary development from diverse communities, but future research will need to examine which sampling methods can yield accurate, population-representative data that can advance our understanding of the link between sociodemographic variation and variation in language outcomes.

References

Alcock, K., Meints, K., & Rowland, C. (2020). *The UK Communicative Development Inventories: Words and Gestures.* J&R Press.

Astivia, O. L. O., & Zumbo, B. D. (2019). Heteroskedasticity in multiple regression analysis: What it is, how to detect it and how to solve it with applications in R and SPSS. *Practical Assessment, Research, and Evaluation,* 24(1), 1.

Aust, F., & Barth, M. (2020). Papaja: Create APA manuscripts with R Markdown.

Bates, E., & Goodman, J. C. (2001). On the inseparability of grammar and the lexicon: Evidence from Acquisition. *Essential Readings in Developmental Psychology. Oxford,* 134-162.

Bates, E., Marchman, V. A., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language,* 21(01), 85–123.

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2020). Estimatr: Fast estimators for design-based inference.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics,* 37 (6), 1461–1476.

Bornstein, M. H., & Putnick, D. L. (2012). Stability of language in childhood: A multiage, multidomain, multimeasure, and multisource study. *Developmental Psychology,* 48(2), 477.

Braginsky, M. (2020). Wordbankr: Accessing the Wordbank database. Retrieved from https://CRAN.R-project.org/package=Wordbankr.

Broman, K. W. (2020). Broman: Karl Broman's R code.

Caselli, N. K., Lieberman, A. M., & Pyers, J. E. (2020). The ASL-CDI 2.0: An updated, normed adaptation of the MacArthur Bates Communicative Development Inventory for American Sign Language. *Behavior Research Methods,* 1–14.

Chai, J. H., Lo, C. H., & Mayor, J. (2020). A bayesian-inspired item response theory–based framework to produce very short versions of MacArthur–Bates Communicative Development Inventories. *Journal of Speech, Language, and Hearing Research,* 63(10), 3488–3500.

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). Xtable: Export tables to LaTeX or HTML. Retrieved from https://CRAN.R-project.org/package=xtable.

Dale, P. S. (2015). Adaptations, Not Translations! Retrieved from http://mb-cdi.stanford.edu/Translations2015.pdf.

De Houwer, A. (2019). Equitable evaluation of bilingual children's language knowledge using the CDI: It really matters who you ask. *Journal of Monolingual and Bilingual Speech*, 1(1), 32–54.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development*, 71(2), 310–322.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates Communicative Development Inventories.* Brookes Publishing Company.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5).

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics*, 21(1), 95–116.

Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., ... others. (2018). Vocabulary of 2-year-olds learning English and an additional language: Norms and effects of linguistic distance. *Monographs of the Society for Research in Child Development*, 83(1), 1–135.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... others. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project.* MIT Press.

Henry, L., & Wickham, H. (2020). Purrr: Functional programming tools.

Hester, J., & Wickham, H. (2020). Fs: Cross-platform file system operations based on ’libuv’.

Kachergis, G., Marchman, V. A., Dale, P., Mehta, H., Mankewitz, J., & Frank, M. C. (2021, April 7-9). *An online computerized adaptive test (CAT) of children’s vocabulary development in English and Mexican Spanish*. Poster presented at the Biennial Meeting of the Society for Research in Child Development, virtual conference.

Kapalková, S., & Slančová, D. (2006, May). Adaptation of CDI to the Slovak language. In *Proceedings from the first European network meeting on the communicative development inventories* (pp. 24-28).

Kartushina, N., Mani, N., Aktan-Erciyes, A. S. L. I., Alaslani, K., Aldrich, N. J., Almohammadi, A., ... Mayor, J. (2021). COVID-19 first lockdown as a unique window into language acquisition: What you do (with your child) matters. PsyArXiv. https://doi.org/10.31234/osf.io/5ejwu

Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E. A., & Henriksen, L. Y. (2013). The use of the internet in collecting CDI data–an example from Norway. *Journal of Child Language*, 40(03), 567–585.

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the MacArthur–Bates Communicative Development Inventory: Words & Sentences (CDI: WS). *Journal of Speech, Language, and Hearing Research*, 59(2), 281–289.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur–Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, 51(5), 2248–2255.

Müller, K. (2017). Here: A simpler way to find your files.

Müller, K., & Wickham, H. (2020). Tibble: Simple data frames.

Pew Research Center. (2021). *Mobile fact sheet.* Retrieved from https://www.pewresearch.org/internet/fact-sheet/mobile/.

R Core Team. (2020). R: A language and environment for statistical computing. Journal of Open Source Software (Vol. 4, p. 1686). Vienna, Austria: R Foundation for Statistical Computing; Springer-Verlag New York.

Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). Digest of education statistics 2017, NCES 2018-070. National Center for Education Statistics.

U.S. Census Bureau. (2011). *Table NC-EST2019-ASR6H: 2019 Population Estimates by Age, Sex, Race and Hispanic Origin*. Retrieved from https://www.census.gov/newsroom/press-kits/2020/population-estimates-detailed.html.

U.S. Department of Education, National Center for Education Statistics. (2018). *Table 104.40: Percentage of persons 18 to 24 years old and age 25 and over, by educational attainment, race/ethnicity, and selected racial/ethnic subgroups: 2010 and 2017.* In U.S. Department of Education, National Center for Education Statistics (Ed.), *Digest of Education Statistics* (2018 ed.). Retrieved from https://nces.ed.gov/programs/digest/d18/tables/dt18_104.40.asp?referer=raceindica.asp.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis.

Wickham, H. (2019). Stringr: Simple, consistent wrappers for common string operations.

Wickham, H. (2020a). Forcats: Tools for working with categorical variables (factors).

Wickham, H. (2020b). Tidyr: Tidy messy data.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2020). Dplyr: A grammar of data manipulation.

Wickham, H., & Hester, J. (2020). Readr: Read rectangular text data.

Wickham, H., & Seidel, D. (2020). Scales: Scale functions for visualization.

Wilke, C. O. (2020). Cowplot: Streamlined plot theme and plot annotations for ’ggplot2’.

Zhu, H. (2020). kableExtra: Construct complex table with 'kable' and pipe syntax.

Data, Code and Materials Availability Statement

- Open data: All data analyzed in this work are available on the Open Science Framework at https://osf.io/nmdq4/.

- Code: All code for this work is avaiable on the Open Science Framework at https://osf.io/nmdq4/.
- Materials: All code and materials for the Web-CDI are openly available at https://github.com/langcog/web-cdi. If readers wish to view the Web-CDI interface in full from the participants' or researchers' perspectives, they are encouraged to contact webcdi-contact@stanford.edu.

Ethics Statement

Data collected in the United States for this project are anonymized according to guidelines set forth by the United States Department of Health and Human Services. Data collection at Stanford University was approved by the Stanford Institutional Review Board (IRB), protocol 20398.

Authorship and Contributorship Statement

All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

- Conceptualization: Benjamin deMayo, Danielle Kellier, Mika Braginsky, Christina Bergmann, Caroline Rowland, Michael Frank and Virginia Marchman.
- Data Curation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.
- Formal Analysis: Benjamin deMayo.
- Funding Acquisition: Caroline Rowland and Michael Frank.
- Investigation: Benjamin deMayo, Danielle Kellier and Virginia Marchman.
- Methodology: Benjamin deMayo, Danielle Kellier, Michael Frank and Virginia Marchman.
- Project Administration: Caroline Rowland, Michael Frank and Virginia Marchman.
- Software: Danielle Kellier, Mika Braginsky, Christina Bergmann and Cielke Hendriks.
- Supervision: Christina Bergmann, Caroline Rowland, Michael Frank and Virginia Marchman.
- Visualization: Benjamin deMayo.
- Writing - Original Draft Preparation: Benjamin deMayo, Michael Frank and Virginia Marchman.
- Writing - Review & Editing: Benjamin deMayo, Danielle Kellier, Mika Braginsky, Christina Bergmann, Cielke Hendriks, Caroline Rowland, Michael Frank and Virginia Marchman.

Acknowledgements

We thank Larry Fenson, Philip Dale, and Donna Thal for their assistance and helpful feedback preparing this manuscript.

Appendix

Table A1 ***Settings customizable by researchers when creating new studies to be run on the Web-CDI platform***

Setting	Default value	Notes
Study name	none	
Instrument	none	
Age range for study	none	Defaults based on instrument selected.
Number of days before study expiration	14	Must be between 1 and 28 days.
Measurement units for birth weight	Pounds and ounces	Weight can also be measured in kilograms (kg).
Minimum time (minutes) a parent must take to complete the study	6	
Waiver of documentation	blank	Can be filled in by researchers to include a Waiver of Documentation for the participant to approve before proceeding to the experiment.
Pre-fill data for longitudinal participants?	No, do not populate any part of the form	Researchers can choose to pre-fill the background information and the vocabulary checklist.
Would you like to pay subjects in the form of Amazon gift cards?	No	If checked, researchers can enter gift codes to distribute to participants once they have completed the survey.

Table A1 (continued) ***Settings customizable by researchers when creating new studies to be run on the Web-CDI platform***

Setting	Default value	Notes
Do you plan on collecting only anonymous data in this study? (e.g., posting ads on social media, mass emails, etc.)	No	If checked, researchers can set a limit for the maximum number of participants, as well as select an option that asks participants to verify that the information entered is accurate.
Would you like to show participants graphs of their data after completion?	Yes	
Would you like participants to be able to share their Web-CDI results via Facebook?	No	
Would you like participants to answer the confirmation questions?	No	Asks redundant demographic questions to serve as attention checks.
Provide redirect button at completion of study?	No	Used to redirect users to external site after form completion.
Capture the Prolific ID for the participant?	No	For integration with Prolific.
Allow participant to print their responses at end of Study?	No	
End message	Standard end-of-study message	Can be changed to customize end-of-study message.

Table A2. ***Regression output for WG comprehension measure***

term	estimate	standard error	statistic	*p* value	conf low	conf high	df
Intercept	122.275	2.427	50.381	0.000	117.515	127.035	1610
Age	20.050	0.767	26.127	0.000	18.545	21.556	1610
Caregiver education: Some college	17.445	8.179	2.133	0.033	1.403	33.487	1610
Caregiver education: High school or less	21.862	10.935	1.999	0.046	0.413	43.311	1610
Age * Caregiver education: Some college	-1.991	2.261	-0.881	0.379	-6.425	2.443	1610
Age * Caregiver education: High school or less	-6.604	3.159	-2.091	0.037	-12.800	-0.408	1610

Table A3. ***Regression output for WG production measure***

term	estimate	standard error	statistic	*p* value	conf low	conf high	df
Intercept	29.771	1.332	22.358	0.000	27.159	32.382	1610
Age	7.599	0.498	15.264	0.000	6.622	8.575	1610
Caregiver education: Some college	5.640	4.919	1.147	0.252	-4.009	15.289	1610
Caregiver education: High school or less	20.455	7.693	2.659	0.008	5.366	35.545	1610
Age * Caregiver education: Some college	-1.357	1.327	-1.022	0.307	-3.960	1.247	1610
Age * Caregiver education: High school or less	-0.121	2.095	-0.058	0.954	-4.229	3.988	1610

Table A4. ***Minimum time to completion, WG measure***

Age in months	Minimum time to completion (minutes)
8	3.496
9	4.057
10	4.619
11	5.181
12	5.743
13	6.305
14	6.867
15	7.429
16	7.991
17	8.553
18	9.115

Table A5. ***Minimum time to completion, WS measure***

Age in months	Minimum time to completion (minutes)
16	8.129
17	8.613
18	9.097
19	9.581
20	10.065
21	10.550
22	11.034
23	11.518
24	12.002
25	12.486
26	12.970
27	13.455
28	13.939
29	14.423
30	14.907

License

Passive sentence reversal errors in autism: Replicating Ambridge, Bidgood, and Thomas (2020)

Samuel Jones
Lancaster University, UK
ESRC International Centre for Language and Communicative Development (LuCiD)

Madeline Dooley
University of Liverpool, UK

Ben Ambridge
University of Liverpool, UK
ESRC International Centre for Language and Communicative Development (LuCiD)

Abstract: Ambridge, Bidgood, and Thomas (2020) conducted an elicitation-production task in which children with and without autism described animations following priming with passive sentences. The authors reported that children with autism were more likely than IQ-matched children without autism to make reversal errors, for instance describing a scene in which the character Wendy surprised the character Bob by saying *Wendy was surprised by Bob*. We set out to test whether this effect replicated in a new sample of children with and without autism (N = 26) and present a cumulative analysis in which data from the original study and the replication are pooled (N = 56). The main effect reported by Ambridge et al. (2020) replicated: While children with and without autism produced a similar number of passive responses in general, the responses of children with autism were significantly more likely to include reversal errors. Despite age-appropriate knowledge of constituent order in passive syntax, thematic role assignment is impaired among some children with autism.

Keywords: autism; syntax; priming; passives; language disorder

Corresponding author(s): Correspondence concerning this article should be addressed to Sam Jones, Department of Psychology, Lancaster University, Lancaster, United Kingdom, LA1 4YF. Email: sam.jones@lancaster.ac.uk. Telephone: +44 (0) 1524 593698.

ORCID ID(s): Samuel Jones: https://orcid.org/0000-0002-8870-3223; Ben Ambridge: https://orcid.org/0000-0003-2389-8477

Citation: Jones, S. D., Dooley, M., & Ambridge, B. (2021). Passive sentence reversal errors in autism: Replicating Ambridge, Bidgood, and Thomas (2020). *Language Development Research*, 1(1), 99-122. DOI: 10.34842/g1zk-3715

Introduction

Language development in autism

Approximately 1% of English-learning children are affected by autism, defined as persistent deficits in social interaction and communication, and restricted and repetitive patterns of behaviour, interests, or activities (Baron-Cohen et al., 2009). The language abilities of children with autism vary widely. Some children have little or no language, while others have advanced language skills and may appear pedantic or verbose. Although, as a group, children with autism tend to use shorter and grammatically simpler sentences than children without autism (Eigsti, Bennetto, & Dadlani, 2007), the acquisition of morphosyntax and word order appear relatively standard among affected children (Tek, Mesite, Fein, & Naigles, 2014). Semantic-pragmatic and narrative development are, in contrast, key areas of difficulty for children with autism, who often show poor understanding of metaphorical and figurative language, poor inferencing skills, difficulty resolving semantic ambiguities (e.g. homographs), and pronoun reversals in which the speaker mistakenly uses *you* in self-reference and *I* to refer to the listener (Naigles & Tek, 2017; see Norbury, 2015, for review).

Target study: Ambridge, Bidgood, and Thomas (2020)

The aim of the current work was to replicate a study that attempted to separate out syntactic and semantic-pragmatic factors contributing to language deficits in a group of children with autism. It is important to be clear at the outset exactly what we mean by syntactic versus semantic-pragmatic factors. Here, we adopt the definition set out in the study that is the target of our replication (Ambridge et al, 2020: 185).

A widely held view in the literature is that, despite broader linguistic and communicative difficulties, 'pure syntax' is relatively spared in children with autism, i.e., spared relative to the broader cognitive deficits that accompany this condition. On this view, which might be summarized in the phrase 'form is easy, meaning is hard' (Naigles, 2002; Naigles & Tek, 2017), syntax itself is spared, and the communicative difficulties that are experienced by children with autism are caused by impairments in other areas of language, such as vocabulary, semantics, socio-pragmatics and narrative (e.g., Tager-Flusberg, Lord & Paul, 1997; Jordan, 1993). To be clear, there is evidence that even children without autism find certain semantic or pragmatic aspects of language more difficult than purely syntactic or structural ones (Naigles, 2002); the claim is, then, that this is even more true for children with autism.

Ambridge et al. (2020) investigated the ability of children aged 6-9 years, with and without autism, to accurately describe an animation using primed passive sentences such as *Bob was surprised/chased/pulled by Wendy*. These authors argue that the (English) passive is a particularly useful test case for separating out syntactic and semantic or pragmatic impairments, since it exemplifies standard syntactic representations and relations – e.g. [SUBJECT] [BE] [VERB] ([PP]) – yet is unusual in terms of its semantics and pragmatics, reversing the [AGENT][PATIENT] word order of actives, and treating the [PATIENT] rather than the [AGENT] as topical.

In the target study of Ambridge et al. (2020), one experimenter described an example video animation using a model passive sentence before a second experimenter provided the participant with a cue verb with which to describe a novel animation. For example, given the verb *surprise* and an animation in which the character Wendy surprised the character Bob, a successfully primed response of *Bob was surprised by Wendy* was coded as a correct passive. Correct actives, meanwhile, were coded when the child produced responses such as *Wendy surprised Bob*; that is, when there was little evidence of a priming effect and the child defaulted to the more frequent active form. Of central interest in the Ambridge et al. (2020) study – and in the current replication – was the rate of reversal errors children made, in which a passive response exhibited an error in thematic role assignment. For instance, in response to the animation in which Wendy surprised Bob – that is, Wendy is the [AGENT] and Bob is the [PATIENT] – the child produced the passive *Wendy was surprised by Bob*; mis-assigning Wendy as [PATIENT] and Bob as [AGENT]. Given that many children affected by autism have difficulty with the referential, inferential, and narrative building aspects of language, it was hypothesised that this group would produce a higher rate of passive reversal errors than IQ-matched children without autism.

Ambridge et al. (2020) note that prior work testing passive sentence comprehension among children with autism reports mixed results. For instance, Tager-Flusberg (1981) reported that children with autism (aged *M* = 8;1) were no more likely than younger (*M* = 3;10) IQ-matched controls to mis-comprehend passive structures, as evidenced in an act-out task. In contrast, Paul, Fisher, and Cohen (1988), who used stimuli matched to those used by Tager-Flusberg (1981), reported evidence that children with autism do make more reversal errors than IQ-matched controls. Ambridge et al. (2020) was the first production study to look at reversal errors in children with autism, with prior production studies in this area excluding reversal errors from analyses (e.g. Allen, Haywood, Rajendran, & Branigan, 2011). Ambridge et al. (2020) report a modest though reliable pattern of higher reversal errors among children with autism relative to IQ-matched peers. These results were interpreted as further evidence that semantics, pragmatics, and narrative, rather than 'pure syntax', constitute key areas of language difficulty for children affected by autism (though other interpretations are possible; a point to which we return at length in the *Discussion*).

Why replicate?

The value of the Ambridge et al. (2020) study is that it investigates a specific grammatical structure – the passive. Work testing the processing and comprehension of specific grammatical structures among children with autism is lacking (Norbury, 2014), and this is unfortunate because such work can provide a basis for developing finely targeted games or activities to be used in programmes of language support. In the case of the passive, for instance, if children with autism do indeed have a good command of the syntax of this construction, but not of its semantic-pragmatic aspects, interventions based around this construction should focus on the latter, not the former. For example, a narrative-based intervention which emphasizes how the current discourse topic (e.g. *Have you heard the news about YouTube?...)* makes a natural passive PATIENT-SUBJECT (*...It got bought by Google*; Pullum, 2014: 64) is likely to be more useful than one focussed directly on syntax, such as a task encouraging children to produce passive sentences when describing pictures with no discourse context.

Nevertheless, one limitation of Ambridge et al. (2020) – and indeed all prior studies into passive sentence processing and comprehension in children with autism – is that as a hard-to-reach population, language development studies involving children with autism often have small sample sizes. For instance, Ambridge et al. (2020) tested 15 children with autism, while Tager-Flusberg (1981) tested 18 children, and Paul et al. (1988) tested just six children. For this reason, while the use of a paradigm sensitive enough to identify specific deficits in the processing of a defined linguistic structure among children with autism is welcomed, without further replication many readers may be unconvinced by this effect, especially given its small magnitude.

The current study

The purpose of the current study was, therefore, to test whether the findings of Ambridge et al. (2020) – i.e., higher rates of reversal error among children with autism than among IQ-matched peers – replicated in a new sample of children. In approaching this project, we faced similar resourcing constraints, and tested a similar number of children (*N* = 26, *n* = 13 with autism). However, re-using the original stimuli and procedure enabled us to produce – in addition to our own replication analysis – a cumulative analysis of the pooled data involving 28 participants per group (*N* = 56). Cumulative analysis should be distinguished from questionable research practices such as optional stopping or *p*-hacking, in which researchers covertly gather data up to the point at which their hypothesis is superficially confirmed, or add or remove specific data points in order to retrieve a *p*-value below the standard .05 alpha level. In contrast, we explicitly label 'original', 'replication', and 'pooled' data throughout this analysis, and all of our data and code is made publicly available via an online repository: https://osf.io/c2pjd/.

In both the present replication and the original study, we used syntactic (or 'structural') priming purely as a method for eliciting passives. The phenomenon of syntactic priming itself is not under investigation, and we remain agnostic with regard to the question of whether priming constitutes a particularly useful window into children's learning and representation of structural knowledge. Since the passive is a highly infrequent and marked construction, it is likely that most children would have produced very few, if any, passives, had we run the study as a simple elicited-production task with no priming element.

Method

Participants

Thirteen children aged 6 to 9 years (*M* = 8;0) with autism were recruited from specialist schools in North West England. Entry to these schools was based on a prior diagnosis of autism and an extensive battery of screening assessments, resembling that shown in Appendix A of Ambridge et al. (2020). In the current study, we took the additional precaution of screening children independently using the Lifetime version of the Social Communicative Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003). The SCQ comprises 40 items to which caregivers are required to provide yes or no responses. Responses are then tallied to determine the child's SCQ score. A child with an SCQ score of 15 or over is likely to be on the autistic spectrum. Children in the autism group of the current replication study had SCQ scores ranging from 19 to 29 (*M* = 22.85), providing independent validation of diagnosis and experimental group identity. Thirteen children without autism aged four to six (*M* = 5;3) were recruited from mainstream English pre-schools and schools. By-participant demographics and SCQ and IQ scores are presented in the Appendix.

Following Ambridge et al. (2020), children with and without autism were IQ-matched using the short version of the Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV; Wechsler, 2012). IQ scores were used to match the with-autism and without-autism groups, and for use as a control predictor in the statistical analyses, but were not used to define cut-offs for either group. The results of this administration are shown in Table 1, alongside corresponding administration results from Ambridge et al. (2020). Visual inspection of this data indicates reasonable similarity both across studies and between experimental and control groups within studies. Where there are discrepancies between groups, these are attributable to children with autism outperforming children without autism, meaning that matching may be considered conservative. For instance, in both the original study and in the replication, children with autism scored numerically higher on the object assembly subset of the WPPSI-IV, while in the replication a numerical advantage for picture memory was also recorded among children with autism.

It is important to note that, since the children with autism were considerably older than the IQ-matched children without autism (i.e., a mean age of 8;0, as opposed to 5;3), it would not be accurate to refer to the former group as 'children with autism but without intellectual disabilities' (previously termed 'high functioning autism'; though see Alvares, Bebbington, Clearly, Evans, Glasson, et al., 2020). However, since all of the children with autism were able to complete a relatively complex verbal task, any intellectual disabilities present for children in this group were relatively minor. It would also not, therefore, be appropriate to generalize the findings from the present study (or that of Ambridge et al, 2020, which was conducted with similar participant groups) to children with autism with greater intellectual disabilities.

Table 1. Mean (and standard deviation) scores for the Wechsler Preschool and Primary Scale of Intelligence, Fourth Edition (WPPSI-IV), across seven subsets. The unit-weighted composite mean is also shown. WA = Without Autism; ASC = autism spectrum condition.

	Ambridge, Bidgood, and Thomas (2020) (*N* = 30)		Jones, Dooley, and Ambridge (2020) (*N* = 26)	
	ASC (*n* = 15)	WA (*n* = 15)	ASC (*n* = 13)	WA (*n* = 13)
Receptive vocabulary	22.98 (3.36)	23 (2.7)	22.41 (3.04)	21.62 (1.82)
Block design	19.7 (7.7)	20.48 (2.69)	20.99 (3.22)	19.32 (2.2)
Picture memory	20.37 (5.33)	15.34 (4.28)	20.35 (2.73)	18.62 (2.53)
Information	19.19 (3.66)	21.02 (2.29)	20.63 (3.23)	19.64 (2.68)
Object assembly	28.38 (6.7)	22.45 (7.88)	25.22 (2.89)	19.74 (2.48)
Zoo locations	10.42 (2.53)	11.13 (0.99)	12.19 (1.47)	12.03 (2.02)
Picture naming	17.95 (2.71)	17.45 (2.5)	17.43 (2.33)	17.57 (1.5)
Unit-weighted composite mean	19.86 (2.97)	18.7 (2.18)	19.89 (1.92)	18.36 (1.24)

Scaled, unit-weighted composite means of WPPSI-IV scores were included as control variables in the hierarchical Bayesian models used throughout this study (referred to as ‘IQ’ in Ambridge et al., 2020). These composite means were calculated by summing the scaled scores for each subset for each child and then dividing by the number of subsets (i.e., seven). The value shown at the bottom of Table 1 was, in contrast, calculated by summing the raw (i.e., not scaled) mean scores across children and dividing by the number of children. Note that these mean scores in the replication align well with those in the original article (e.g., for the experimental group, *M* = 19.89 versus *M* = 19.86). While Ambridge et al. (2020) used independent t-tests to check for equivalence between groups – and reported no statistically significant differences on the basis of the data shown in Table 1 – we avoided this analysis given concerns regarding the use of inferential methods to test for so-called nuisance effects (Sassenhagen & Alday, 2016). Readers interested in formally testing for equivalence between groups may use our R code to do so.

Procedure and scoring

The procedure and scoring used in this study were identical to those used in Ambridge et al. (2020). The participant and two experimenters sat in a quiet room in front of a computer screen and played a bingo-style game designed to engage and sustain the participant's attention. One experimenter acted as adjudicator, and first passed a prime verb card (Table 2) to the second experimenter. The second experimenter then used the specified prime verb in a passive sentence to describe a short animation played on the computer screen. After this, a target verb card (Table 3) was passed to the participant, who was required to use the specified verb to describe a novel animation. After each trial, the adjudicator, who was not able to see the computer screen, looked into a tub and, if one was available, retrieved a bingo point card corresponding to the description made. The game was engineered to ensure that the participant always finished with more bingo points than the experimenter.

Inspection of Tables 2 and 3, which shows age-of-acquisition and (where available) familiarity ratings (Bird, Franklin & Howards, 2001; Kuperman, Stadthagen-Gonzalez & Brysbaert, 2012), suggests that the majority of prime and target verbs would be known by even the youngest children who participated in the present study. Note that in the original study target verbs (Table 3) were split into three semantic classes; agent-patient, experiencer-theme, and theme-experiencer. This manipulation was included to test whether children with or without autism found verbs of a particular semantic class easier to use in the task described. Prior research suggests, for instance, that children without autism have particular difficulty processing experiencer-theme verbs, such as *forget, love,* and *remember* (e.g. Ambridge, Bidgood, Pine, Rowland, & Freudenthal, 2016; see Ambridge et al., 2020, p. 4, for overview). Ambridge et al. (2020) report identifying this effect among children both with and without autism. However, due to focussed theoretical interest in the rate of reversal errors in the current replication, the verb type manipulation does not form part of the current analysis or write up, where we instead home in on the main effects of response type by group (though see R code for additional analyses).

Table 2. Twenty-four prime verbs, with available age-of-acquisition and imageability ratings from Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) and Bird, Franklin, and Howards (2001).

Verb	AOA (Kuperman et al., 2012): Years	AOA (Bird et al., 2001): 1-7 scale*	Imageability (Bird et al., 2001): 1-7 scale
Avoid	8.50	4.22	3.40
Bite	3.58		
Call	4.74	2.54	4.21
Carry	5.16		
Chase	5.53	2.82	5.29
Cut	4.43		
Dress	4.05	2.31	
Drop	3.26	2.31	
Eat	2.78	1.67	
Follow	5.11	2.91	
Help	3.65	2.69	4.05
Hit	4.75	2.30	
Hold	4.67		
Hug	2.58	2.45	
Kick	4.06	2.43	
Kiss	3.61		
Lead	6.76		
Pat	5.07	2.42	
Pull	4.79		
Push	4.26	2.39	
Shake	5.26	2.84	
Squash	6.94		
Teach	4.67	3.04	
Wash	4.00	1.95	5.84

* 1 = 0-2 years; 2 = 3-4 years; 3 = 4-5years; 4 = 6-7 years; 5 = 9-10 years; 6 = 11-12years; 7 = 13 years or older.

Table 3. Thirty-six target verbs, with available age-of-acquisition and imageability ratings from Kuperman, Stadthagen-Gonzalez and Brysbaert (2012) and Bird, Franklin and Howards (2001).

Verb	AOA (Kuperman et al., 2012): Years	AOA (Bird et al., 2001): 1-7 scale*	Imageability (Bird et al., 2001): 1-7 scale
Amaze	7.50	3.83	4.92
Annoy	7.22	3.11	4.57
Bite	3.58		
Bother	6.50	3.36	3.52
Carry	5.16		
Chase	5.53	2.82	5.29
Dress	4.05	2.31	
Forget	4.78	3.25	3.36
Frighten	8.83	2.86	
Hate	5.53	3.33	3.95
Hear	3.80	2.53	
Hit	4.75	2.30	
Hug	2.58	2.45	
Ignore	6.74	4.30	
Impress	10.17		
Kick	4.06	2.43	
Know	4.50	2.75	
Like	3.69	2.49	3.32
Love	5.17	2.51	5.03
Pat	5.07	2.42	
Please	3.48		
Pull	4.79		
Push	4.26	2.39	
Remember	5.63	3.27	3.91
Scare	4.22		
See	3.06	2.39	
Shock	7.53	4.13	4.13
Smell	4.22	2.41	
Squash	6.94		

Table 3 continued. Thirty-six target verbs, with available age-of-acquisition and imageability ratings from Kuperman, Stadthagen-Gonzalez and Brysbaert (2012) and Bird, Franklin and Howards (2001).

Verb	AOA (Kuperman et al., 2012): Years	AOA (Bird et al., 2001): 1-7 scale*	Imageability (Bird et al., 2001): 1-7 scale
Surprise	5.47		
Tease	5.11		
Understand	6.17	3.94	3.40
Upset	5.26	3.29	4.16
Wash	4.00	1.95	5.84
Watch	4.33		
Worry	6.61	4.15	4.76

* 1 = 0-2 years; 2 = 3-4 years; 3 = 4-5years; 4 = 6-7 years; 5 = 9-10 years; 6 = 11-12years; 7 = 13 years or older.

Children's responses were coded using the regime described in Ambridge et al. (2020; pp. 6–7), and touched on in the introduction to the current study. Given an animation in which Wendy scared Bob, for instance, a response of *Bob was scared by Wendy* was coded as a 'correct passive'; a response of *Wendy was scared by Bob* was coded as an 'incorrect passive' (i.e., a reversal error – the response of primary interest); a response of *Wendy scared Bob* was coded as a 'correct active'; and responses such as *scared Bob* were coded as 'other use of target verb'. Responses outside of these four categories were excluded from the analysis. Given only two incorrect active responses among participants, this category was excluded from all statistical analyses, as it was excluded in the original study.

Statistical analysis

A series of maximal Bayesian hierarchical models were fitted using the brms package in R (Bürkner, 2018; R Core Team, 2016). In each model, response type (i.e., correct active, correct passive, incorrect passive, and other verb) was predicted by group (i.e., non-autism, autism) and WPPSI score, with target sentence (i.e., verb) and participant as grouping variables. In brms syntax:

Model = brm(formula = Response ~ Group + WPPSI +
(1 + Group + WPSSI | *Target sentence) +*
(1 | *Participant)*

One additional model was fitted with identical fixed and random effects and total passive responses (i.e., correct plus incorrect passives) as the dependent variable. The purpose of this model was to determine whether groups produce similar rates of passive response overall. Following Ambridge et al. (2020), we set a conservative prior of 2.77 on beta (β; see R code for detailed model specification, and p. 7 of the original study for the justification of prior). These models were fitted not only to our replication data (*N* = 26) but also to the original data (*N* = 30) and to the pooled data (*N* = 56; i.e., the original and replication data combined). Each model fitted well, as indicated by rhat values uniformly at one and credible posterior predictive visualisation checks (see brms package documentation for details of diagnostics; Bürkner, 2018). We believe the model specifications used here and indeed in the original target study to be well justified. However, researchers keen to test different configurations of, for instance, prior or random or fixed effects are invited to do so using our data and code. Note that we switched coding of non-autism and autism groups relative to the original study.

Across the analyses presented in this paper, then, children without autism form the baseline group, rather than children with autism. This allows readers to see more clearly the associations between a diagnosis of autism and the likelihood of giving a response of a certain type. Note also that we do not follow Ambridge et al.'s (2020) approach of calculating *pMCMC* values, or 'Bayesian *p*-values', but rather report a combination of proportional odds and 90% highest density intervals (HDIs), i.e., the most credible 90% span of the posterior distribution. This broadly follows the approach outlined in McElreath (2016; though McElreath uses narrower 89% HDIs – the choice is arbitrary), which we believe to provide an intuitive method of communicating results and propagating uncertainty in the data. Readers who disagree are welcome to calculate *pMCMC* values or conduct alternative analyses (e.g. using 89% or 95% HDIs) using our data and R code.

The results that follow can be interpreted in the following way. A HDI bound above zero (e.g. 0.2 to 0.5) suggests a positive association between variables (e.g. a diagnosis of autism and higher rates of reversal error). A HDI bound below zero (e.g. -0.8 to -0.3) suggests a negative association between variables (e.g. a diagnosis of autism and lower rates of accurate passive responses). And a HDI spanning zero (e.g. -0.3 to 0.4) suggests no linear relationship between predictors is plausible (i.e., no difference between children with and without autism with respect to a particular response).

Results

A by-participant summary of the results can be found in the Appendix. As this table shows, at least one reversed passive was produced by 6/13 children with autism and 2/13 children without autism. Descriptive statistics of task performance are presented for reference in Table 4. Importantly, the correct passive and incorrect passive columns of Table 4 provide evidence of a priming effect. Every animation in the task could have been described accurately using the cue verb in an active sentence. However, children both with and without autism appeared to be primed to some extent by the experimenter's example sentence and used passive syntax to describe animations in 31.45% of trials overall (i.e., 256 passives out of 814 total responses), despite passive syntax being low frequency in everyday speech.

Table 4. Performance (mean, with standard deviation in brackets) in the original study (Ambridge, Bidgood, & Thomas, 2020; ABT) and current replication study (Jones, Dooley, & Ambridge, 2020; JDA) by group (WA= Without Autism; ASC = autism spectrum condition).

Study	Group	Correct active	Incorrect active	Correct passive	Incorrect passive	Other verb
ABT	WA	3.13 (1.67)	0.09 (0.29)	0.91 (1.2)	0.18 (0.44)	0.33 (0.56)
ABT	ASC	2.2 (1.82)	0.07 (0.25)	0.91 (1.33)	0.77 (1.29)	0.55 (0.79)
JDA	WA	3.58 (1.89)	0.0 (0.0)	2.1 (1.97)	0.02 (0.16)	0.08 (0.27)
JDA	ASC	2.62 (1.57)	0.05 (0.32)	1.13 (1.54)	0.33 (0.62)	0.38 (0.63)

Prior to our main analysis, we tested whether groups were similarly likely to produce passive sentences overall, i.e., correct passives and reversal errors combined. The results of this analysis are presented in Table 5, which shows estimates and 90% HDIs for the original ($N = 30$), replication ($N = 26$), and pooled ($N = 56$) data. In the original study it was reported on the basis of descriptive statistics (i.e., no model was fitted) that children with autism were more likely to produce passive sentences than children without autism. While the Bayesian analysis of the original data implies this effect (estimate = 0.11), we note that the 90% HDI for this estimate crosses zero (HDI = -0.05, 0.25), indicating that the true effect may be practically null. In the replication data, the estimate suggests children with autism were in contrast less likely to produce passive sentences than children without autism (estimate = -0.13), however the 90% HDI for this estimate again suggests that the effect may not be substantial (HDI = -0.29, 0.01). Analysis of the pooled data indicates the absence of any group effect on the production of passive sentences (estimate = 0.02, HDI = -0.08, 0.13). Overall, then, children with and without autism were equally likely to respond using passive syntax. Children with autism produced passives in 131 out of 374 responses (i.e., 35.03%), while children without autism produced passives in 134 out of 440 responses (i.e., 30.45%).

Table 5. Estimates and 90% highest density intervals (HDI) for the association between a diagnosis of autism and a passive response.

Data	Estimate	90% HDI
Original	0.11	-0.05, 0.25
Replication	-0.13	-0.29, 0.01
Pooled	0.02	-0.08, 0.13

We then looked at rates of reversal error. Analyses of the original (*N* = 30), replication (*N* = 26), and pooled (*N* = 56) data indicate that children with autism were more likely to make reversal errors than children without autism (Table 6; pooled HDI = 1.06, 4.21). Overall, 47 out of 374 responses made by children with autism contained reversal errors (i.e., 12.57%), while just 9 out of 440 responses made by children without autism contained reversal errors (i.e., 2.05%). In the pooled analysis, the beta coefficient for the association between a diagnosis of autism and the production of a reversal error was = 2.59. Exponentiating this estimate shows that, while groups produced a comparable number of passives in general (Table 5), the proportional odds of a child with autism mis-assigning thematic roles and producing a reversal error were approximately thirteen times (13.33) higher than the odds of a child without autism doing likewise.

Table 6. Estimates and 90% highest density intervals (HDI) for the association between a diagnosis of autism and reversal errors.

Study	Estimate	90% HDI
Original	2.11	-0.52, 4.50
Replication	2.67	-0.39, 6.03
Pooled	2.59	1.06, 4.21

Next, we looked at whether children with autism were more or less likely than children without autism to respond using correct actives (Table 7). Analysis re-confirmed that in the original study children with autism were less likely than children without autism to produce correct actives (HDI = -1.99, -0.12). However, replication and data pooling indicate a density interval spanning zero (pooled HDI = -1.20, 0.01). Overall, then, it is not clear that children without autism produced substantially more correct active responses than children with autism. The number of correct active responses made by children in each group was high. Overall, 199 out of 374 responses made by children with autism were correct actives (i.e., 53.21%), while 284 out of 440 responses made by children without autism were correct actives (i.e., 64.55%).

Table 7. Estimates and 90% highest density intervals (HDI) for the association between a diagnosis of autism and correct actives.

Study	Estimate	90% HDI
Original	-1.05	-1.99, -0.12
Replication	0.12	-0.75, 0.98
Pooled	-0.58	-1.20, 0.01

Finally, we looked at whether children with autism were more or less likely than children without autism to respond with an alternative use of the target verb. For instance, responding *Wendy pulling Bob* where the target passive sentence was *Bob was pulled by Wendy*[1]. The results of these analyses are shown in Table 8. Estimates and HDIs indicate that children with autism were consistently more likely than children without autism to use the target verb in a response other than the correct active or a passive.

Table 8. Estimates and 90% highest density intervals (HDI) for the association between a diagnosis of autism and other uses of the target verb.

Study	Estimate	90% HDI
Original	1.60	-0.11, 3.37
Replication	2.08	-1.11, 5.07
Pooled	2.47	0.88, 4.06

In the pooled data, 39 out of 374 responses made by children with autism involved an alternative use of the target verb (i.e., 10.43%), while 18 out of 440 responses made by children without autism involved an alternative use of the target verb (i.e., 4.09%). We note that many of these responses were reasonable. For instance, the response of *Homer was annoying Marge* instead of the expected target *Marge was annoyed by Homer*; the response of *Wendy was letting Bob pat her* instead of *Wendy was patted by Bob*; and the response of *Marge is carrying Homer* instead of *Homer was carried by Marge*.

[1] As these examples show, this response category includes both grammatical and ungrammatical uses of the target verb. Of the 17 responses in this category, ten were fully grammatical, two (both produced by children without autism) included a past-tense overgeneralization error (*bited,* in both cases), and five were unclear. These were all cases such as *Marge remembering Homer* which is ungrammatical as a standalone sentence, but which could be acceptable as a response to an implicit question such as *What can you see in this video?*.

Discussion

The language of children with autism varies dramatically, from children who have little or no language to children who have advanced language skills and may appear pedantic or verbose (Norbury, 2014). While as a group children with autism often use shorter and grammatically simpler sentences than children without autism (Eigsti et al., 2007), it has been argued that the main areas of language difficulty for children with autism are semantics, pragmatics, and narrative, rather than 'pure syntax' (Naigles & Tek, 2017). The current study aimed to tease apart these effects through a replication of work by Ambridge et al. (2020). These authors asked 30 children aged 6-9 years, with and without autism, to describe a series of animations using a cue verb, primed by the experimenter to use passive syntax. The response of primary interest was the rate of reversal errors, in which passive syntax is used accurately but thematic roles are mis-assigned (e.g. the child describes an animation in which Wendy [AGENT] surprises Bob [PATIENT] with the phrase *Wendy* [PATIENT] *was surprised by Bob* [AGENT]). Ambridge et al. (2020) report a higher rate of reversal errors among children with autism than among children without autism.

We set out to test whether this effect replicated in a new sample of children with and without autism (*N* = 26) and presented a cumulative analysis in which data from the original study and the replication were pooled (*N* = 56). Analysis indicated that the main effect reported by Ambridge et al. (2020) replicated in this new sample of children. Table 5 of the current study shows that children with autism were in general as likely as children without autism to produce passive sentences. However, the groups differed substantially in the rate of reversal errors they made, with children with autism approximately thirteen times more likely than children without autism to make an error in thematic role assignment, for instance describing a scene in which Wendy surprised Bob using the phrase *Wendy was surprised by Bob* (Table 6). Results corroborate Ambridge et al.'s (2020) conclusion that despite age-appropriate knowledge of (at some level) constituent order in passive syntax, the ability of certain children with autism to map syntax to thematic roles is impaired.

Embedding the cue verb in an accurate passive sentence was clearly challenging for children both with and without autism, due to their young age and the high complexity and low frequency of this syntactic structure. This was reflected in the high rate of 'default' active responses made by children with and without autism (i.e., 53.21% and 64.55% respectively; see Table 7), and the relatively high rate of alternative responses made by children with autism (i.e., 10.43%; see Table 8). The real challenge, of course, is to explain why children with autism produced inaccurate passives in 12.57% of trials (versus 2.05% of trials among children without autism), instead of defaulting to active syntax or responding with an alternative verb usage if task demands were high. Ambridge et al. (2020, pp 15–17) discuss two possibilities. The first is that children with autism struggle to understand the discourse-pragmatic conditions under which typical AGENT-PATIENT order is reversed (e.g., when the PATIENT is topical; *Have you heard the news about YouTube? It got bought by Google*; Pullum, 2014: 64). The second and related possibility is that reversal errors are part and parcel of the same narrative deficit that sometimes causes children with autism to mention characters or events in the wrong order. Both of these possibilities are consistent with the replication and cumulative datasets presented here, which converge on a very similar pattern of results. Rather than re-describe these possibilities, then, we here present an alternative account that nevertheless remains compatible with those summarised in Ambridge et al. (2020).

Under construction-based accounts of language acquisition (e.g., Tomasello, 2003; Dabrowksa, 2004; Goldberg, 2019), children build constructions – including the passive – by analogizing across input utterances that exemplify these constructions. This is true even for those accounts that explicitly retain the original exemplars (e.g., Abbot-Smith & Tomasello, 2006; Ambridge, 2020). For example, suppose that a child without autism hears sentences such as *Chloe was hit by Danny, James was kicked by Billy* and *Sarah was dressed by her Dad*. The assumption is that, on the basis of such utterances, the child forms a construction schema of the form *[PATIENT] [BE] [ACTION] by [AGENT]* (even if only very approximately; Ambridge, 2020). This construction will allow her to produce an appropriate passive sentence such as *Bob was pushed by Wendy* (a target utterance in the present study). Suppose, now, that a child with autism hears sentences such as *Chloe was hit by Danny, James was kicked by Billy* and *Sarah was dressed by her Dad,* but instead forms a construction schema of the form *[PERSON] [BE] [ACTION] by [PERSON]*. This more general construction will allow her to produce both appropriate passive sentences such as *Bob was pushed by Wendy* and (as a description of the same event) incorrect reversed passive sentences such as *Wendy was pushed by Bob*.

This account, as it is presented above, would seem to predict – incorrectly – that children with autism will produce correct and reversed passives at rates of around 50/50. In fact, however, the notion of a child forming either a *[PATIENT] [BE] [ACTION] by [AGENT]* or a *[PERSON] [BE] [ACTION] by [PERSON]* construction is a gross oversimplification. In reality, 'constructions' are probabilistic and multi-facetted: The first slot is neither PERSON nor PATIENT but a probabilistic cluster of all the properties of all of the different entities that have appeared in this position in input utterances (see Ambridge, 2020, for a detailed discussion of how re-representing exemplar utterances at an increasingly abstract level in a computational model results in abstractions that *approximate* – but never map on to entirely – linguistic constructions at various levels of abstraction).

An advantage of this account is that it can potentially also explain the finding of Paul, Fisher, and Cohen (1988) that children with autism make more reversal errors of this type than do IQ-matched controls, when assessed using comprehension methods (though see Tager-Flusberg for a null finding using a similar methodology). But is there any reason to believe that children with autism are more likely than children without autism to form (probabilistically) these overly general constructions? We are not aware of any directly-relevant research evidence, but the possibility is generally consistent with the empathizing-systemizing view of Baron-Cohen and colleagues (e.g., Baron-Cohen, 2009), under which people with autism lie at the more systemizing end of the continuum. Classifying verb arguments as AGENT, PATIENT, EXPERIENCER or THEME might require a degree of empathising, of understanding others' perspectives and emotions. Classifying verb arguments as PERSON does not, and is a more systematic approach, in that it posits a higher level of generalization; that is, of systematicity.

Of course, this possibility is highly speculative at present but could potentially be investigated in future research, for example by investigating whether children with autism make similar errors for other constructions that require human participants to be classified into fine-grained psychological categories like RECIPIENT (e.g., dative/ditransitive constructions). Another potentially illuminating direction for future research would be to replicate the priming task described in this study using animations depicting a mixture of human interactions (e.g., Wendy surprising Bob) and systematic physical processes (e.g. a cam rotating and making a lever move). People with autism and Asperger syndrome are reported to show better understanding of physical systems than people without autism, despite apparent deficits in interpreting human intentions among this population (Lawson, Baron-Cohen, & Wheelwright, 2004). It would be interesting to test, therefore, whether among children with autism the rate of reversal errors would be lower for passive sentences describing systems (e.g., *the cam was moved by the lever*) than for sentences describing human interactions (e.g., *Wendy was surprised by Bob*).

In the pooled analysis presented in this study, the odds of a child with autism producing a reversal error were approximately thirteen times higher than the odds of a child without autism doing likewise. Nevertheless, we noted that children with autism produced reversal errors on only 12.57% of their total responses. Despite substantial proportional odds, then, it may be argued that this modest magnitude on an absolute scale makes the passive reversal effect trivial, particularly considering how rarely passive syntax occurs in natural speech. That is, passive sentences may occur so rarely in natural speech that apparently mild deficits in mapping thematic roles among some children with autism may not cause significant problems in language use. It is important to note, however, that the current study looked at a sample of children with relatively low scores on the SCQ measure of autism (some only a few points above the cut-off of 15). It may well be, therefore, that children with higher scores would produce more reversal errors (or even a different pattern of responses entirely). Determining how patterns of performance in the current paradigm link to specific cognitive profiles will enable us to determine whether the results reported here may guide the fine-tuning of programmes of language support for children with autism. It is likely that the task will need to be modified for use with participants showing different symptomologies.

Conclusion

The current study presented a replication of Ambridge et al. (2020). While children with and without autism produced a similar number of passive responses in general, the responses of children with autism were significantly more likely to include errors in thematic role assignment. Despite age-appropriate knowledge of (at some level) constituent order in passive syntax, the ability of certain children with autism to use word order to appropriately mark thematic roles is impaired.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23,* 275–290. https://doi.org/10.1515/TLR.2006.011

Ambridge, B., Bidgood, A., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2016). Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science*, *40*(6), 1435–1459. https://doi.org/10.1111/cogs.12277

Allen, M. L., Haywood, S., Rajendran, G., & Branigan, H. (2011). Evidence for syntactic alignment in children with autism. *Developmental Science*, *14*(3), 540–548. https://doi.org/10.1111/j.1467-7687.2010.01001.x

Alvares, G. A., Bebbington, K., Cleary, D., Evans, K., Glasson, E. J., Maybery, M. T., ... & Whitehouse, A. J. (2020). The misnomer of 'high functioning autism': Intelligence is an imprecise predictor of functional abilities at diagnosis. *Autism*, *24*(1), 221-232. https://doi.org/10.1177/1362361319852831

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, 014272371986973. https://doi.org/10.1177/0142723719869731

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

Ambridge, B., Bidgood, A., & Thomas, A. (2020). Disentangling syntactic, semantic and pragmatic impairments in ASD: Elicited production of passives. *Journal of Child Language,* 1–18. https://doi.org/10.1017/S0305000920000215

Baron-Cohen, S. Autism: The empathizing–systemizing (E-S) theory. *Ann. N. Y. Acad. Sci.* 1156, 68–80 (2009). https://doi.org/10.1111/j.1749-6632.2009.04467.x

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *British Journal of Psychiatry, 194*(6), 500–509. https://doi.org/10.1192/bjp.bp.108.059345

Bird, H., Franklin, S., & Howard, D. (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers, 33,* 73–79. https://doi.org/10.3758/BF03195349

Bürkner, P.-C. (2018). Bayesian Regression Models using "Stan." *Journal of Statistical Software*. CRAN repository. https://doi.org/10.18637/jss.v080.i01>

Dąbrowska, E. (2004). *Language, mind and brain: Some Psychological and neurological constraints on theories of grammar*. Edinburgh University Press. http://www.jstor.org/stable/10.3366/j.ctvxcrgdw

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Eigsti, I.-M., Bennetto, L., & Dadlani, M. B. (2007). Beyond pragmatics: Morphosyntactic development in autism. *Journal of Autism and Developmental Disorders, 37*(6), 1007–1023. https://doi.org/10.1007/s10803-006-0239-2

Goldberg, A. E. (2019). *Explain me this: Creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Lawson, J., Baron-Cohen, S., & Wheelwright, S. (2004). Empathising and systemising in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders, 34*(3), 301–310. https://doi.org/10.1023/B:JADD.0000029552.42724.1b

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

Naigles, L. R., & Tek, S. (2017). 'Form is easy, meaning is hard' revisited: (Re) characterizing the strengths and weaknesses of language in children with autism spectrum disorder. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(4), e1438. https://doi.org/10.1002/wcs.1438

Norbury, C. F. (2014). Autism spectrum disorders and communication. In L. Cummings (Ed.), *The Cambridge Handbook of Communication Disorders* (pp. 141–157). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139108683.011

Paul, R., Fisher, M., & Cohen, D. (1988). Brief report: Sentence comprehension strategies in children with autism and specific language disorders. *Journal of Autism and Developmental Disorders, 18*, 669–679. https://doi.org/10.1007/BF02211884

Pullum, G. K. (2014). Fear and loathing of the English passive. *Language & Communication, 37*, 60-74. http://dx.doi.org/10.1016/j.langcom.2013.08.009

R Core Team. (2016). R. *R Core Team*. Retrieved from https://www.r-project.org/

Rutter, M., Bailey, A., & Lord, C. (2003). *The social communication questionnaire: Manual*. Los Angeles, CA: Western Psychological Services.

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162*, 42–45. https://doi.org/10.1016/j.bandl.2016.08.001

Tager-Flusberg, H. (1981). On the nature of linguistic functioning in early infantile autism. *Journal of Autism and Developmental Disorders, 11*(1), 45–56. https://doi.org/10.1007/BF01531340

Tek, S., Mesite, L., Fein, D., & Naigles, L. (2014). Longitudinal analyses of expressive language development reveal two distinct language profiles among young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 44*(1), 75–89. https://doi.org/10.1007/s10803-013-1853-4

Tomasello, M. (2003). *Constructing a language*. Harvard University Press.
Wechsler, D. (2012).

Wechsler, D. (2012). *WPPSI-IV: Wechsler preschool and primary scale of intelligence* (4th ed.). Bloomington, MN: Pearson, Psychological Corporation.

Data, code and materials availability statement

All data, code, and materials are available at the Open Science Framework repository accompanying this article: https://osf.io/c2pjd/.

Ethics statement

Ethics approval was obtained from the ethics committee of the University of Liverpool.

Authorship and Contributorship Statement

Samuel Jones analyzed the data, wrote the first draft of the manuscript, and revised the manuscript during peer review. Madeline Dooley collected the data. Ben Ambridge conceived of and designed the study, oversaw data analysis, and revised the manuscript prior to and during peer review. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

Samuel Jones and Ben Ambridge are members of the ESRC International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/S007113/1 and ES/L008955/1] is gratefully acknowledged.

Appendix

By-participant SCQ, seven-subset WPPSI-IV, and task scores. WA = without autism; ASC = autism spectrum condition.

Participant	Group	SCQ	Vocab	Blocks	Pictures	Information	Assembly	Zoo	Naming	Age	Correct Passive	Incorrect Passive	Correct Active	Other Verb	Irrelevant
1	WA	NA	20	16	22	13	20	6	21	6.3	0.22	0	0.67	0.06	0
2	WA	NA	20	18	20	18	17	11	17	4.44	0.39	0	0.61	0	0
3	WA	NA	22	18	20	20	22	12	17	5.2	0.67	0.06	0.28	0	0
4	WA	NA	19	18	16	18	17	11	16	4.96	0.06	0.06	0.89	0.06	0
5	WA	NA	24	22	22	20	17	14	17	5.45	0.33	0	0.67	0	0
6	WA	NA	24	20	18	20	20	13	18	5.47	0	0	1	0	0
7	WA	NA	22	18	16	20	22	11	19	4.5	0.22	0	0.78	0	0
8	WA	NA	20	18	15	21	16	13	16	5.47	0.56	0	0.44	0	0
9	WA	NA	22	24	21	20	19	14	17	4.03	0.39	0	0.61	0	0
10	WA	NA	20	18	22	19	22	14	17	4.33	0.56	0	0.44	0	0
11	WA	NA	24	22	17	23	25	12	18	5.47	0.5	0	0.44	0	0.06
12	WA	NA	24	21	17	25	20	13	20	5.34	0.61	0	0.39	0	0
13	WA	NA	20	18	16	18	20	12	16	6.51	0.17	0	0.72	0.06	0.06
14	ASC	23	21	20	16	19	24	11	14	9.43	0	0.06	0.39	0	0.61
15	ASC	24	23	18	20	22	28	12	18	8.45	0.11	0.17	0.28	0.17	0.28
16	ASC	21	23	24	20	18	28	13	18	8.96	0.06	0	0.56	0.11	0.28
17	ASC	23	22	20	18	23	30	12	14	7.79	0.11	0	0.39	0	0.5
18	ASC	21	21	16	18	14	24	13	17	9.07	0.11	0	0.28	0.11	0.5
19	ASC	26	15	20	18	16	24	11	18	8.35	0.33	0.11	0.28	0.11	0.17

Appendix continued

By-participant SCQ, seven-subset WPPSI-IV, and task scores. WA = without autism; ASC = autism spectrum condition.

Participant	Group	SCQ	Vocab	Blocks	Pictures	Information	Assembly	Zoo	Naming	Age	Correct Passive	Incorrect Passive	Correct Active	Other Verb	Irrelevant
20	ASC	29	28	24	26	25	28	14	22	6.63	0.56	0.06	0.33	0	0.06
21	ASC	25	23	18	22	23	19	9	17	7.63	0.11	0.28	0.39	0.06	0.17
22	ASC	22	21	20	23	19	24	14	19	6.65	0.17	0.06	0.5	0	0.28
23	ASC	20	24	26	23	25	28	12	17	8.27	0.06	0	0.61	0.11	0.22
24	ASC	23	21	16	21	18	22	11	13	6.58	0.17	0	0.67	0	0.17
25	ASC	19	25	22	18	23	25	14	19	9.01	0.39	0	0.5	0.06	0.06

License

Can phones, syllables, and words emerge as side-products of cross-situational audiovisual learning? — A computational investigation

Khazar Khorrami
Unit of Computing Sciences, Tampere University, Finland

Okko Räsänen
Unit of Computing Sciences, Tampere University, Finland
Department of Signal Processing and Acoustics, Aalto University, Finland

Abstract: Decades of research has studied how language learning infants learn to discriminate speech sounds, segment words, and associate words with their meanings. While gradual development of such capabilities is unquestionable, the exact nature of these skills and the underlying mental representations yet remains unclear. In parallel, computational studies have shown that basic comprehension of speech can be achieved by statistical learning between speech and concurrent referentially ambiguous visual input. These models can operate without prior linguistic knowledge such as representations of linguistic units, and without learning mechanisms specifically targeted at such units. This has raised the question of to what extent knowledge of linguistic units, such as phone(me)s, syllables, and words, could actually emerge as latent representations supporting the translation between speech and representations in other modalities, and without the units being proximal learning targets for the learner. In this study, we formulate this idea as the so-called *latent language hypothesis* (LLH), connecting linguistic representation learning to general predictive processing within and across sensory modalities. We review the extent that the audiovisual aspect of LLH is supported by the existing computational studies. We then explore LLH further in extensive learning simulations with different neural network models for audiovisual cross-situational learning, and comparing learning from both synthetic and real speech data. We investigate whether the latent representations learned by the networks reflect phonetic, syllabic, or lexical structure of input speech by utilizing an array of complementary evaluation metrics related to linguistic selectivity and temporal characteristics of the representations. As a result, we find that representations associated with phonetic, syllabic, and lexical units of speech indeed emerge from the audiovisual learning process. The finding is also robust against variations in model architecture or characteristics of model training and testing data. The results suggest that cross-modal and cross-situational learning may, in principle, assist in early language development much beyond just enabling association of acoustic word forms to their referential meanings.

Keywords: early language acquisition; computational modeling; visually grounded speech; language representation learning; neural networks.

Corresponding author(s): Okko Räsänen, Unit of Computing Sciences, Tampere University, P.O. Box 553, FI-33101, Tampere, Finland. Email: okko.rasanen@tuni.fi.

ORCID ID(s): https://orcid.org/0000-0002-0537-0946

Citation: Khorrami, K., & Räsänen, O. (2021). Can phones, syllables, and words emerge as side-prod-ucts of cross-situational audiovisual learning? — A computational investigation. *Language Development Research*, 1(1), p 123-191. DOI: 10.34842/w3vw-s845

Introduction

When learning to communicate in their native language, infants face a number of challenges that they need to overcome in order to become proficient users of the language. In order to understand speech, they need to figure out how to extract words from the running acoustic signal and how the words relate to objects and events in the external world (cf. Quine, 1960). In order to develop syntactic skills and become creative and efficient users of the language, they must understand that speech is made of units smaller than individual words, allowing combination of these units to form new meanings. In essence, this means that the child learner has to acquire understanding of spoken language as a hierarchical compositional system. In this system, smaller units such as phonemes or syllables make up larger units such as words and phrases, and where these units are robust against different sources of non-phonological variability in the acoustic speech.

The journey from a newborn infant without prior linguistic knowledge to a proficient language user consists of several learning challenges. While one body of developmental research has investigated how infants can utilize distributional cues related to phonetic categories of their native language (e.g., Werker and Tees, 1984; Maye et al., 2002; see also Kuhl et al., 2007 for an overview), another set of studies has focused on the question of how infants could segment acoustic word forms from running speech where there are no universal cues to word boundaries (e.g., Cutler and Norris, 1988; Mattys et al.,1999; Saffran et al., 1996; Thiessen et al., 2005; Choi et al., 2020). Yet another line of research has investigated word meaning acquisition, assuming that words as perceptual units are already accessible to the learner. In that research, the focus has been on the details of the mechanisms that link auditory words to their visual referents when they co-occur at above-chance probability across multiple infant-caregiver interaction scenarios (e.g., Smith and Yu, 2008; Trueswell et al., 2013; Yurovsky et al., 2013), also known as cross-situational learning.

All these different stages have received a great deal of attention in the existing research, both experimental and computational. However, we still have limited understanding on how the different stages and sub-processes in language learning interact with each other, what drives learning in all these different tasks, and what type of acoustic or linguistic representations infants actually develop at different stages of the developmental timeline. For instance, does adaptation to phonetic categories pave the way for lexical development (cf. NLM-e framework by Kuhl et al., 2017), or is early lexical learning a gateway to refined phonemic information (cf. PRIMIR theory by Werker & Curtin, 2005)? How accurately do words have to be segmented before their referential meanings can be acquired?

In contrast to viewing language learning as a composition of different learning tasks, an alternative picture of the process can also be painted: what if processes such as word segmentation or phonetic category acquisition are not necessary stepping stones for speech comprehension, but that language learning could be bootstrapped by meaning-driven predictive learning, where the learner attempts to connect the (initially unsegmented) auditory stream to the objects and events in the observable surroundings (Johnson et al., 2010; Räsänen and Rasilo, 2015; also referred to as discriminative learning in Baayen et al., 2015; see also Ramscar and Port, 2016). While tackling this idea has been challenging in empirical terms, a number of computational studies have explored this idea along the years (e.g., but not limited to, Yu et al., 2005; Roy and Pentland, 2002; Räsänen and Rasilo, 2015; Chrupała et al., 2017; Alishahi et al., 2017; Räsänen and Khorrami, 2019; ten Bosch et al., 2008; Ballard and Yu, 2004). These models have demonstrated successful learning of speech comprehension skills in terms of connecting words in continuous speech to their visual referents with minimal or fully absent prior linguistic knowledge.

Since rudimentary semantics of spoken language seem to be accessible to (computational) learners without having to first learn units such as phone(me)s, syllables or words, it is of interest whether some type of representations for such units could actually *emerge as a side-product* of the cross-modal and cross-situational learning process. The idea is that, instead of learners separately and sequentially tackling a number of sub-problems on the road towards language proficiency, linguistic knowledge could emerge as a latent representational system that effectively mediates the "translation" between auditory speech and other internal representations related to the external world or the learner itself. While not precluding the fact that certain aspects of language skills are likely to emerge earlier than others, the key value of this idea—here referred to as latent language hypothesis (LLH)—is that it replaces a number of proximal language learning goals (phoneme category learning, word segmentation, meaning acquisition) with a unified learning goal of minimizing the predictive uncertainty in the multisensory environment of the learner. This goal aligns well with the popular view of the mammalian brain as a powerful multimodal prediction machine (Friston, 2010; Clark, 2013; see also Meyer and Damasio, 2009, or Bar, 2011), and also fits to the picture of predictive processing at various levels of language comprehension (e.g., Warren, 1970; Jurafsky, 1996; Jurafsky et al., 2001; Watson et al., 2008; Kakouros et al., 2018; Cole et al., 2010). Even if cross-modal learning would not be the primary mechanism for acquisition of linguistic knowledge, it is important to understand the extent the cross-modal dependencies can facilitate (or otherwise affect) the process.

The goal of this paper is to review and explore the feasibility of LLH as a potential mechanism for bootstrapping the learning of language representations at various levels of granularity without ever explicitly attempting to learn such representations. We specifically focus on the case of audiovisual associative learning between visual scenes and auditory speech. We build on the existing computational studies on the topic, and attempt to provide a systematic investigation of LLH by comparing a number of artificial neural network (ANN) architectures for audiovisual learning. We first define LLH in terms of high-level computational principles and review the existing research on the topic in order to characterize the central findings so far. We then present our computational modeling experiments of visually-grounded language learning, where we investigate a large battery of phenomena using a unified set of evaluation protocols: the potential emergence of phone(me)s, syllables, words, and word semantics inside the audiovisual networks. We study whether individual artificial neurons and layers of neurons become correlated with different linguistic units, and whether this leads to qualitatively discrete or continuous nature of acquired representations in terms of time and representational space. Finally, we summarize and discuss our findings and the extent that the LLH could explain early language learning.

While our experiments largely rely on existing body of work in this area (see section Earlier Related Work), our current contributions include i) a coherent theoretical framing of the present and earlier studies under the concept of LLH, ii) an integrative summary of the existing research, iii) systematic experiments investigating several different aspects of language representation learning in terms of linguistic units of different granularity and in terms of unit selectivity and temporal dynamics, iv) comparison of alternative neural model architectures within the same experimental context, and v) comparing learning and representation extraction from both synthetic and real speech. In addition, we propose a new objective technique to evaluate the semantics learned by the audiovisual networks.

Theoretical Background

One of the key challenges in early language acquisition research is to identify the fundamental computational principles responsible for the learning process. Young learners have to solve an apparently large number of difficult problems ranging from unit segmentation and identification to syntactic, semantic, and pragmatic learning on their way to become proficient language users. Is thereby unclear what type of collection of innate biases, constraints, and learning mechanisms are needed for language learning to succeed. In terms of parsimony, a theory should aim to explain the different aspects of LA with a minimal number of distinct learning mechanisms.

The key idea behind LLH is to replace several separate language learning processes and their proximal learning targets with a single general overarching principle for learning, namely predictive optimization. In short, LLH relies on the idea that the mammalian brain has evolved to become an efficient uncertainty reduction (=prediction) device, where input in one or more sensory modalities is used to construct a set of predictions regarding the overall state of the present and future sensorimotor environment (cf., e.g., Friston, 2010; Clark, 2013). This strategy has several ecological advantages. For instance, complete sensory sampling of the environment would take excessive time and effort, and actions often need to be taken with incomplete information of a constantly changing environment. In addition, predictive processing allows focusing of attentional resources on those aspects of the environment that have high information gain to the agent (see, e.g., Kakouros et al., 2018, for a review and discussion). As a result, the ability to act based on partial cues of the "external world state" (also across time) results in a substantial ecological advantage. Importantly, predictive processing necessitates some type of probabilistic processing of the stochastic sensory environment. This is because evaluation of the information value of different percepts requires a model of their relative likelihoods in different contexts (or degrees of "surprisal"; see also the Goldilocks effect; Kidd et al., 2012). This connects the overarching idea of predictive processing to the concept of *statistical learning* in developmental literature, as infants appear to be adept learners of temporal (Saffran et al., 1996) and cross-modal probabilistic regularities (Smith & Yu, 2008).

In the context of LLH, we postulate that statistical learning is a manifestation of general sensorimotor predictive processing, and where language learning could also be driven by optimization of predictions *within* and *across* sensory modalities[1] during speech perception. In order to efficiently translate heard acoustic patterns to their most likely visual referents or to predict future speech input, intermediate latent representations that best support this goal are needed. More specifically, the question is whether representation of the *linguistic structure* underlying the variable and noisy acoustic speech could emerge as a side product of such a predictive optimization problem (see also van den Oord et al., 2018).

In case of audiovisual associative learning, this idea can be illustrated by a simple high-level mathematical model such as

$$\arg_{\theta} \max p(\bar{v}_t \mid \bar{x}_t, \theta) \mid \forall t \ \in [0, T] \qquad (1)$$

where $\bar{v}_t$ is visual input at time *t*, $\bar{x}_t$= {x_0, x_1, x_2, ..., x_t} is the speech input up to time *t*, θ is a statistical model (or biological neural system) enabling evaluation of the probability, and *T* is the total cumulative experience ("age") of the learner so far. Now, assuming that 1) θ consists of several *plastic* processing stages/modules $\theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ (e.g., layers or cortical areas in in artificial or biological neural networks), 2) Eq. (1) can be solved or approximated using some kind of learning process, and that 3) observed speech and visual input are statistically coupled, θ must result in intermediate representations that together lead to effective predictions of the corresponding visual world, given some input speech. If a solution for θ is discovered, i.e., *the model has learned to relate speech to visual percepts,* we can ask whether the intermediate stages of θ have become to carry emergent representations that correlate with

[1] In the most generic form also including motor aspects with articulation and manual gestures.

how linguistics would characterize the structure of speech. Alternatively, if the model becomes able to understand even basic level semantics between speech and the visual word without reflecting any known characteristics of spoken language, that would be a curious finding in itself.

The basic formulation in Eq. (1) can be extended to model the full joint distribution $p(\bar{x}_t, \bar{v}_t|\theta)$ of audiovisual experiences. Alternatively, assuming stochasticity of the environment, it can be reformulated as minimization of Kullback-Leibler divergence between $p(\bar{v} \mid \psi)$ and $p(\bar{v} \mid \bar{x}, \theta)$, where ψ is a some kind of stochastic generator of visual experiences (due to interaction with the world) and the latter term is the learner's model of visually grounded speech. However, the main implication of each of these models stays the same: discovering a model θ that provides an efficient solution to the cross-modal translation problem between spoken language and other representations of the external world. The same idea can be applied to within-speech predictions across time by replacing $\bar{v}$ with $\bar{x}_{t+k}$ ($k > 0$) in Eq. (1). In this case, if k is set sufficiently high, the learned latent representations must generalize across phonemically irrelevant acoustic variation in order to generate accurate predictions for future evolution of the speech signal given speech up to time t; evolution which is primarily governed by phonotactics and word sequence probabilities in the given language (see van den Oord et al., 2018, for phonetic feature learning with this type of approach; cf. also models of distributional semantics, such as Mikolov et al., 2013, that operate in an analogous manner with written language).

Given the existence of modern deep neural networks, LLH can be investigated using flexible hierarchical models that can tackle complicated learning problems with real-world audiovisual data, and without pre-specifying the representations inside the networks. This is also what we do in the present study. While such computational modeling cannot tell us what exactly is happening in the infant brain, it allows us to investigate the fundamental feasibility of LLH under controlled conditions in terms of learnability proofs.

Note that we wish to avoid taking any stance on the debate whether discrete linguistic units are something that exist in the human minds or computational models. In contrast, we adopt a viewpoint similar to Ramscar and Port (2016) and use linguistic structure as an idealized description of speech data, investigating how the representations learned by computational models correlate with the manner that linguistics would characterize the same input. In addition, we do not claim that *audiovisual* learning is necessarily the only mechanism for early acquisition of primitive linguistic knowledge. We simply want to study the extent that this type process can enable or facilitate language learning, and generally acknowledge that purely auditory learning is also central to language learning.

Earlier Related Work

A number of existing computational studies and machine learning algorithms have studied the use of concurrent speech and visual input to bootstrap language learning from sensory experience. In the early works (e.g., Roy and Pentland, 2002; Ballard and Yu, 2004; Räsänen et al., 2008; ten Bosch et al., 2008; Driesen and Van hamme, 2011; Yu et al., 2005; Mangin et al., 2015; Räsänen and Rasilo, 2015), visual information has been primarily used to support concurrent word segmentation, identification, and meaning acquisition. The basic idea in these models has been to combine cross-situational word learning (Smith & Yu, 2008)—the idea that infants learn word meanings by tracking co-occurrence probabilities of word forms and their visual referents across multiple learning situations—with simultaneous "statistical learning" of patterns from the acoustic speech signal. In parallel, a number of robot studies have investigated the grounding of speech patterns into concurrent percepts or actions (e.g., Salvi et al., 2012; Iwahashi, 2003). However, the acoustic input of some studies has been pre-processed to phoneme-like features (Roy and Pentland, 2002; Ballard and Yu, 2004; Salvi et al., 2012) or word segments (Salvi et al., 2012) using supervised learning. Alternatively, visual input to the models have been rather simplified, such as simulated categorical symbols for visual referents e.g., ten Bosch et al., 2008; Räsänen and Rasilo, 2015; Driesen and Van hamme, 2011).

In terms of LLH, the older models have had relatively rigid and flat representational structure, limiting their capability to produce emergent hierarchical representations. In contrast, the older models contain a series of signal processing and machine learning operations to solve the audiovisual task, including initial frame-level signal representation steps such as phoneme recognition or speech feature clustering, followed by pattern discovery from the resulting representations using transition probability analysis (Räsänen et al., 2008; Räsänen & Rasilo, 2015), non-negative matrix factorization (ten Bosch et al.,2008; Mangin et al., 2015), or probabilistic latent semantic analysis (Driesen & Van hamme, 2011), to name a few. Despite these limitations, these studies already demonstrate that access to units such as phonemes or syllables is not required for early word learning, as long as the concurrent visual information is related to the speech contents systematically enough. In addition, they show that word segmentation is not required before meaning acquisition, but that the two processes can take place simultaneously with referential meanings actually defining word identities in the speech stream. Such models can also account for a range of behavioral data from infant word learning experiments using auditory and audiovisual stimuli (Räsänen & Rasilo, 2015).

More recent developments in deep learning have enabled more advanced and flexible hierarchical models that can tackle richer visual and auditory inputs with unified elementary processing mechanisms. These models have their origins in methods for learning relationships between images and natural language descriptions of them, such as photographs and their written labels or captions (e.g., Frome et al., 2013; Socher et al., 2014; Karpathy & Li, 2015). These text-based models have been expanded to deal with acoustic speech input, such as spoken image captions (Synnaeve et al., 2014; Harwath and Glass, 2015; Harwath et al., 2016; Chrupała et al., 2017). Early works applied separate techniques for segmenting words-like units prior to alignment between audio caption data and images e.g. Synnaeve et al., 2014; Harwath and Glass, 2015). The more recent audiovisual algorithms operate without prior segmentation by mapping spoken utterances and full images to a shared high-dimensional vector space (Harwath et al., 2016; Chrupała et al., 2017). However, compared to text, dealing with acoustic speech data is a more difficult task: time-frequency structure of speech is not invariant similarly to orthography, but varies as a function of many different factors ranging from speaker identity to speaking style, ambient noise, or recording setup/listener situation. Moreover, acoustic forms of the elementary units such as phonemes or syllables are affected by the linguistic context in which they occur, causing substantial variation also within otherwise controlled speaking conditions. These are also challenges that language learning infants face, and which cannot be studied with transcription- or text-based models.

In a typical visually grounded speech (VGS) model (Harwath et al., 2016; Chrupała et al., 2017; see Fig. 1 for an example), the model consists of a deep neural network with two separate branches for processing image and speech data: an image encoder responsible for converting pixel-level input into high-level feature representations of the image contents, and a speech encoder doing the same for acoustic input. Both branches consist of several layers of convolutional or recurrent units, and outputs from the both branches are ultimately mapped to a shared high-dimensional semantic space, aka. *embedding space*, via a ranking function. The idea is to learn neural representations for images and spoken utterances so that the embeddings produced by both branches are similar when the input images and speech share semantic content. Once trained, distances between the embeddings derived from inputs can then be used for audiovisual, audio-to-audio, or visual-to-visual search, such as finding the semantically best matching images for a spoken utterance, or finding utterances with similar semantic content than a query utterance (Harwath et al., 2016; Chrupała et al., 2017; see also Azuh et al., 2019, and Ohishi et al., 2020, for cross-lingual approaches).

Training of these models is carried out by presenting the network with images paired with their spoken descriptions (whose mutual embedding distances the model tries to minimize) and pairs of unrelated images and image descriptions (whose embedding distances the model tries to increase). The visual encoder is often pre-trained on some other dataset using supervised learning (but see also Harwath et al., 2018), whereas the speech encoder and mappings from both encoders to the embedding space are optimized simultaneously during the training. Model training is typically conducted on datasets specifically designed for the image-to-speech alignment tasks,

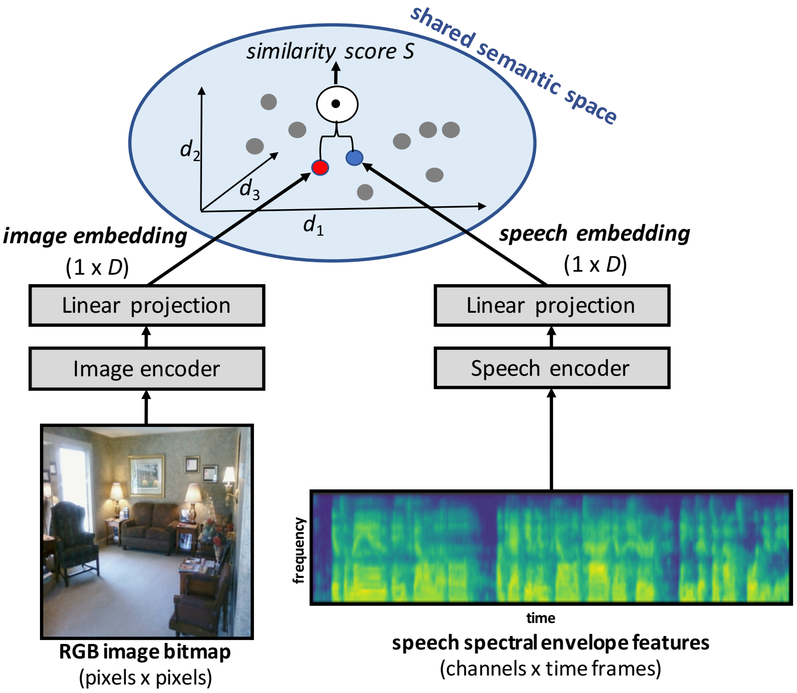


Figure 1. ***The basic architecture of the VGS models explored in the present study. Visual and auditory input data are processed in two parallel branches, both consisting of several neural network layers. Outputs from both branches are mapped into a shared "amodal" embedding space that encodes similarities shared by the two input modalities.***

either by adding synthesized speech to captioned image databases, such as SPEECH-COCO by Havard et al. (2017) or Synthetic Speech COCO (SS-COCO; Chrupała et al., 2017) derived from images and text captions of MS-COCO (Chen et al., 2015), or acquiring spoken descriptions for images using crowd-sourcing, such as Places Audio Caption Corpus (Harwath et al., 2016) derived from Places image database (Zhou et al., 2014) or SpokenCOCO (Hsu et al., 2020) derived from MSCOCO.

Evidence for Language Representations in VGS Models

From the perspective of LLH, the question of interest is whether the audiovisual models learn latent representations akin to linguistic structure of speech, as the models learn to map auditory speech to semantically relevant visual input and vice versa. In this context, a number of studies have investigated phonemic learning in VGS models.

Alishahi et al. (2017) used a recurrent highway network (RHN)—a variant of recurrent neural network (RNN)—VGS model with 5 recurrent layers to investigate how phonological information is represented in intermediate layers of the model (same model as used by Chrupała et al., 2017). Using synthetic speech from SS-COCO, they trained supervised phone classifiers with input-level Mel-frequency cepstral coefficients (MFCCs) and hidden layer activations as features to test how informative the features are with respect to phonetic categories. Alishahi et al. found that, even though the MFCCs already led to approximately 50% phone classification accuracy, the accuracies improved substantially when using activations from the first two recurrent layers of their model (up to approx. 77.5%) and then decreased slightly for the last recurrent layers. To further probe phonetic and phonemic nature of their network representations, Alishahi et al. (2017) also applied a so-called minimal-pair ABX-task (Schatz et al., 2013) to the networks to test whether the hidden representations can distinguish English minimal pairs in speech. Again, the best phonemic discriminability was obtained for the representations of the first two recurrent layers. Alishahi et al. (2017) also applied agglomerative clustering to activations within each layer, and found that the pattern of feature organization in MFCCs and in the first recurrent layer were better correlated with the ground-truth phoneme categories than the activations computed from other layers.

Drexler and Glass (2017) also used the ABX-task to investigate phonemic discriminability of the hidden layer activations of a CNN-based VGS model from Harwath and Glass (2017). Similar to Alishahi et al. (2017), they found that the hidden layer activations were better than the original spectral input features in the ABX-task (among other tasks), that the early layers were phonemically more informative than the deeper ones, and that the network also learned to discard speaker-dependent information from the signal due to the visual grounding. However, they also found that somewhat higher phonemic discriminability was still obtained using purely audio-based unsupervised learning algorithms compared to their VGS model. Another study by Harwath et al. (2020) augmented the CNN-based VGS model from Harwath et al. (2018) with automatic discretization (vector quantization) of the internal representations during the training and inference process. Then they investigated how this affects the phonemic and lexical discriminability of the hidden layer representations. They found that phoneme discrimination ABX scores of the early layer representations were much higher than those typically observed for spectral features in the same task or with a number of baseline speech representation learning algorithms. They also found that discretized representations from early layers primarily carried phonemic information, while representations quantized in deeper layers corresponded better to lexical units. However, discretization did not improve phonemic discriminability beyond the original distributed multivariate representations of the hidden layers.

Recently, Räsänen and Khorrami (2019) trained a weakly supervised convolutional neural network (CNN) VGS model to map acoustic speech to the labels of concurrently visible objects attended by the baby hearing the speech, as extracted from headmounted video data from real infant-caregiver interactions of English-learning infants (Bergelson & Aslin, 2007). They then measured the so-called phoneme selectivity index (PSI) (Mesgarani et al., 2014) of the network nodes and layers. Their results indicated that, in addition to learning a number of words and their referents from such data, hidden layer activations of the model also became increasingly representative of phonetic categories towards deeper layers of the network. The model was also able to handle referential ambiguity in the visual input when the infant was not attending the correct object. However, Räsänen and Khorrami did not use actual visual inputs but categorical labels of the perceived objects, simplifying the visual recognition process substantially.

In terms of phone segmentation, Harwath and Glass (2019) investigated whether activation dynamics of a CNN-based VGS model reflect underlying phonetic structure of speech. They compared temporal activation patterns of VGS-model hidden layers to phone boundaries underlying the input speech data from TIMIT corpus (Garofolo et al., 1993). As a result, they found that peaks in the change-rate of activation magnitudes of the early CNN layers were highly correlated with transitions between phone segments. In contrast to studying whether the models learn to segment, Havard et al. (2020) studied how the performance of VGS models improves if linguistic unit segmentation is provided as side information to the model during the training. They found that explicit introduction of segmentation cues led to substantial performance gains in the audiovisual retrieval task compared to regular VGS training. The effect was the most pronounced when the system was supplemented with a hierarchy of phone, syllable, and word boundaries across different layers of the model.

Several studies have also investigated lexical representations in VGS-based models. Chrupała et al. (2017) used the same RHN-RNN networks as Alishahi et al. (2017) and showed that the RHN model outperformed the earlier CNN model of Harwath et al. (2016) on audio-to-image retrieval task. Then they investigated how linguistic form- and semantics-related aspects of the input are encoded in the hidden layers of the network. Through a number of experiments, Chrupała et al. (2017) showed that form related features become represented within the first layers of their model, whereas deeper layers tended to encode semantics better than the early layers. They also studied how the network responds to homonyms (i.e., words with similar pronunciation but different meaning, such as "*sail*" and "*sale*") and concluded that the representations of deeper network layers became increasingly better at distinguishing homonyms. In other words, the deep representations also contained cues for contextual semantic disambiguation.

Harwath and Glass (2017) investigated whether word segments in speech can be connected to the bounding boxes of corresponding objects in images using a convolutional neural model of VGS, and showed that this was indeed the case. As an extension to their work, Harwath et al. (2018) created a method to map segments of spoken utterances to their associated objects in the pictures (referred to as “match-map” network) in order to investigate how object and word localization emerges as a side-product of training a network using caption-image pairs. In another study, Havard et al. (2019b) studied if lexical units can be segmented from the representations of recurrent layers of a RNN-based VGS model. By using a variety of metrics, they showed that the network learns an implicit segmentation of word-like units and manages to map individual words to their visual referents in the input images.

Kamper et al. (2017) have also studied if visual data can be employed as an auxiliary intermediate tool for detecting words within speech signals. They designed a speech tagging algorithm which is trained using a dataset of aligned speech-image pairs. They first trained a supervised vision tagging system which, given an image, generates probabilities for the presence of different objects within that picture. Next, they integrated their trained vision model with an audio processing network and trained a joint system which maps spoken utterances to the visual object probabilities. As a result, their network learned to output a list of keywords (object category names) given continuous speech input, again without ever receiving direct information on what constitutes a word in an acoustic sense.

Merkx et al. (2019) further improved the audiovisual search performance of the RNN-based VGS model of Chrupała et al. (2017) and used it to study how different layer activations of the model encode words in speech. They used acoustic input features and hidden layer activations as inputs to a supervised word classifier to test if the representations are informative with respect to underlying word identities. They concluded that the presence of individual words in the input can be best predicted using activations of an intermediate (recurrent) layer of their model.

Havard et al. (2019a) studied neural attention mechanism (Bahdanau et al., 2015) in an RNN-based VGS model using English and Japanese speech data. They found that similar to human attention (Gentner, 1982), neural attention mostly focuses on nouns and word endings. This is in line with the knowledge that infant early vocabulary tends to predominantly consist of concrete nouns. In another study, Havard et al. (2019b) examined the influence of different input data characteristics in a word recognition task by feeding the VGS model with synthesized isolated words with varying characteristics. They observed a moderate correlation between word recognition accuracy and frequency of the words in training data, and a weak correlation for image-related factors such as visual object size and saliency. Havard et al. (2019b) also investigated word activations in the same RNN model using the so-called gating paradigm from speech perception studies (Grosjean, 1980). For this purpose, they fed the network with individual spoken words and truncated the words from different positions at the beginning or end of the words. They found that the precision of word recognition dropped steeply if the first phoneme of a word was removed. In contrast, removal of the word-final phonemes had little impact on precision, and the precision decreased steadily when truncating additional phonemes from the end. This was generally in line with data from human lexical decision tasks.

Inspired by the work of Havard et al. (2019b), Scholten et al. (2020) recently studied word recognition in an RNN-VGS model. Instead of using synthesized speech, they conducted their experiments using real speech data from Flickr8k (Harwath & Glass, 2015). Scholten et al. evaluated their model on word recognition by examining how well word embedding vectors can retrieve images with the correct visual object corresponding to the query word, measuring the impact of different factors on word recognition performance. They found that longer word lengths and faster speaking rates were negatively correlated with performance, while word frequency in the training set had a substantial positive impact on the task performance.

Overall, the general finding from the earlier work has been that the representations learned by VGS models exhibit many characteristics related to the underlying linguistic structure of the input speech, and they learn this structure without ever receiving specifications of how speech or language are organized into some kind of elementary units. This suggests that phonetic and lexical representations and segmentation capabilities could emerge as a side-product from meaning-driven learning. However, it is not yet clear in which conditions these phenomena can occur, and how different levels of language representation are related to each other inside the same models. This is since the studied model architectures (RNNs vs. CNNs), model analysis methods (discriminability, clusteredness, node vs. layer selectivity etc.), and data (synthetic vs. real speech) utilized by the previous studies have varied from one study to another. No individual study has attempted to look at the emergence of linguistic units at phonetic, syllabic, and lexical levels in a single model or study, nor compared multiple model architectures within the same experimental context. In addition, the existing studies have rarely reported baseline measures from untrained models, making it unclear how much of the findings are actually driven by the visually-guided parameter optimization compared to the effects of non-linear network dynamics also present with randomly initialized model parameters (see also Chrupała et al., 2020). This leaves unclear questions such as: 1) Can a single neural model reflect emergence of several levels of linguistic structure at the same time, including phone(me)s, syllables, and words, both in time and selectivity? 2) If so, does the network encode such units preferentially in terms of individual selective nodes or distributed representations? 3) How robust these findings are to variations in the neural architecture of VGS models? 4) Do the analysis findings (primarily carried out on synthetic speech) also generalize to real speech with higher acoustical variability?

In our experiments, we seek to address the above questions by systematically investigating the audiovisual aspect of LLH in three alternative VGS network architectures and at phoneme, syllable, and word level, both in terms of selectivity and in terms of temporal characteristics, and using both synthetic and real speech datasets. The second section describes the alternative speech processing networks used in our experiments, followed by methodology to analyze the internal representations of the models with respect to linguistic structure underlying the speech input to the model. In the third section, we describe the data and experimental setup of our study, followed by results, discussion, and conclusions.

Methods

The goal of our experiments was to investigate the extent that linguistic units of different granularity may emerge as a side product of audiovisual cross-situational learning in neural models of visually grounded speech. We also study the extent that the architecture of the model or type of data (real vs. synthetic) affects the nature of the learned representations.

We first explain the adopted VGS model structure in more detail, including three alternative speech encoder architectures explored in our experiments. We then describe our toolkit used to analyze the hidden layer representations of the networks with respect to linguistic characteristics of the input speech. In addition, we propose a new automatic method for evaluating the semantic relevance of the audiovisual associations learned by the models.

Model Architecture and Speech Encoder Variants

VGS systems are generally trained to align between speech and image modalities so that they learn semantic similarities between the two modalities without any explicit supervision in the form of labels. Here our aim is to use VGS models to simulate infants' audiovisual learning, where they hear speech that is related to the observable visual contexts, but does not contain unambiguous and isolated speech-referent pairs. The setup thereby simulated cross-situational word learning under a high degree of referential uncertainty, and without access to prior segmentation of acoustic word forms.

We follow the methodology by Harwath and Glass (2017) and Chrupała et al. (2017), where input to the model consists of images (photographs) and their spoken descriptions. Speech and image data are initially processed in different encoders consisting of several ANN layers, followed by encoder-specific mappings to a shared "amodal" embedding space. In this space, a chosen similarity metric can be used to measure the pairwise similarity of any representations resulting from auditory or visual channels. During training, the model is optimized to assign a higher similarity score for embeddings resulting from images and image descriptions that match with each other (so-called *positive samples*). At the same time, the model tries to assign higher distances for embedding pairs from unmatched images and utterances (*negative samples*). As a result, the model learns to generate embeddings that encode concepts available in both input modalities. The basic architecture of the image-to-speech mapping network is shown in Fig. 1.

In our current visual encoder network, pixel-level RGB image data are first resampled to 224x224 pixels and then transformed into high-level features using VGG16 image classification network (Simonyan and Zisserman, 2015), which is a deep CNN consisting of 16 layers pretrained on ImageNet data (Russakovsky et al., 2015). Output features of the first fully connected layer (14th layer) of VGG16 are then projected linearly to a D-dimensional space to form the final visual embeddings, and where the linear layer weights are optimized during the VGS model training.

Compared Speech Encoder Architectures

We compare three alternative speech encoder networks, all consisting of a stack of convolutional and/or recurrent neural layers applied on speech input. In all models, the input speech is represented by 40-dimensional log-Mel filterbank energies extracted with 25-ms windows with 10-ms window hop-size, which is a representation that simulates the frequency-selectivity of the human ear. The following three speech encoder architectures were investigated in our experiments (Fig. 2):

CNN0 (Fig. 2, left) is a multi-layer convolutional network with an architecture adopted from Harwath and Glass (2017). It includes five convolutional layers with increasing temporal receptive fields, each followed by a max pooling layer. The output of the last convolutional layer is pooled over the entire utterance in order to discard the effects of absolute temporal positioning of the detected patterns.

As an alternative convolutional model, we designed a **CNN1** network (Fig. 2, middle) with 6 convolutional layers and hand-crafted receptive field time-scales in different layers. We specified the convolutional and pooling layers such that the filter receptive field sizes at different layers would approximately correspond to the known typical time-scales of phones, syllables, and words while gradually expanding towards the larger units (see Fig. 2 for details). As in CNN0, the output of the last convolutional layer is maxpooled across all the time steps.

Our third model variant, **RNN** (Fig. 2, right), was adapted from the model introduced originally by Chrupała et al. (2017) and also used by Alishahi et al. (2017). It includes a convolution layer as the first layer, followed by three residualized recurrent layers with Long Short-Term Memory (LSTM) units. Unlike Chrupała et al. (2017), we use three layers instead of the original five layers, as we observed in our initial tests that the three layer model was already capable of achieving comparable performance to the CNN models in the audiovisual mapping task while training much faster than the original model. Also, in order to maintain comparability of the three networks, we do not utilize a separate attention mechanism in the RNN model. The first two recurrent layers of the RNN feed their frame-by-frame activations to the next layer, allowing measurement of their temporal activations. In contrast, the last layer outputs an activation vector for the entire test sentence after processing it fully, discarding the frame-based temporal information.

In all three variants, the utterance-level activations of the final layer are L2 normalized and linearly projected to *D*-dimensional latent space to form the final speech embeddings. These can then be compared to other embeddings within and across the modalities. We use cosine similarity to measure a similarity score *S* between any two embeddings.

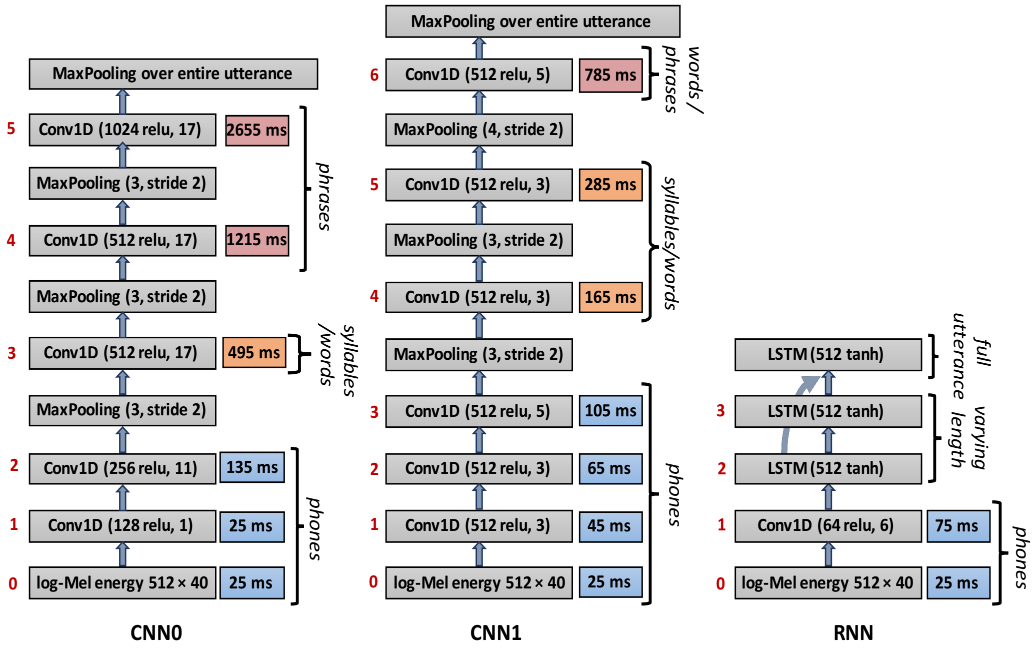

MaxPooling over entire utterance
5 Conv1D (1024 relu, 17) 2655 ms
MaxPooling (3, stride 2)
4 Conv1D (512 relu, 17) 1215 ms
phrases
MaxPooling (3, stride 2)
3 Conv1D (512 relu, 17) 495 ms
syllables /words
MaxPooling (3, stride 2)
2 Conv1D (256 relu, 11) 135 ms
1 Conv1D (128 relu, 1) 25 ms
0 log-Mel energy 512 × 40 25 ms
phones
CNN0
MaxPooling over entire utterance
6 Conv1D (512 relu, 5) 785 ms
words / phrases
MaxPooling (4, stride 2)
5 Conv1D (512 relu, 3) 285 ms
MaxPooling (3, stride 2)
4 Conv1D (512 relu, 3) 165 ms
syllables/words
MaxPooling (3, stride 2)
3 Conv1D (512 relu, 5) 105 ms
2 Conv1D (512 relu, 3) 65 ms
1 Conv1D (512 relu, 3) 45 ms
0 log-Mel energy 512 × 40 25 ms
phones
CNN1
LSTM (512 tanh)
full utterance
3 LSTM (512 tanh)
2 LSTM (512 tanh)
varying length
1 Conv1D (64 relu, 6) 75 ms
0 log-Mel energy 512 × 40 25 ms
phones
RNN

Figure 2. ***Three speech encoders studied in our experiment together with the maximum temporal receptive field lengths of the network nodes. Left: CNN0. Middle: CNN1. Right: RNN. Unit descriptions next to the layers denote the approximate linguistic unit time-scale that the receptive fields of the convolutional layers correspond to. Numbers in red denote layer identifiers used in the analyses of section Results.***

Note that both the CNN and RNN -based models are capable of modeling temporal structure of the data. On one hand, recurrent layers are specifically designed for processing sequential data because they can potentially memorize the history of all past events and therefore recognize patterns across time. On the other hand, convolutional layers are also capable of capturing temporal structure through the hierarchy of increasingly large temporal receptive fields (Gehring et al., 2017), where the largest receptive field size also sets the limit on the temporal distance up to which they can capture statistical dependencies in the data. However, the manner that CNNs and RNNs models capture the temporal structure is very different. Therefore it was of interest whether we can see commonalities or differences in their strategy of encoding linguistic structure of the speech data in order to solve the audiovisual mapping problem.

Model Training

The method we applied for training our networks followed the same strategy as in Harwath et al. (2016) and Chrupała et al. (2017) by using the so-called *triplet loss*: first, a triplet set is made by taking one matching image-speech pair (i.e., an image and an utterance describing it), and adding two negative samples by pairing the original image with a random speech utterance and the original utterance with a random image. The data are then organized into a collection of *B* such triplets. At training time, error backpropagation is used to minimize the following loss function:

$$L(\theta) = \sum_{j=1}^{B} \max\left(0, S_j^c - S_j^p + M\right) + \max\left(0, S_j^i - S_j^p + M\right) \qquad (2)$$

where S_j^p is the similarity score of *j*th ground-truth pair S_j^c the score between original image and the impostor caption, and S_j^i is the score between original caption and the impostor image. In practice, the loss function decreases when ground-truth pair embeddings become more similar to each other. Similarly, the loss decreases when mismatched pairs get further away from each other until they reach distance of *M*, which is referred to as the *margin* of the loss. Intuitively, this means that when the embeddings of a false pair are more than *M* units apart, they are considered as semantically unrelated and the pair no longer affects further parameter updates of the model. As a result, the model learns to tell apart semantically matching and mismatching audiovisual inputs.

Model Evaluation

Our model evaluation consisted of two stages. We first verified that the trained networks have successfully learned to associate auditory and visual patterns to each other, as measured in terms of semantic retrieval tasks. We then proceeded to analyzing whether and how the hidden layer representations of the models correlate with linguistic characteristics of the input speech. Methods and metrics for these analyses are described next.

Audiovisual Search Performance

After training, audio and visual embedding layers can represent semantic similarities between images and spoken captions using the similarity score. Therefore, within a pool of test images and utterances, semantically related examples can be distinguished by sorting instances based on the mutual similarities between their embedding vectors. As a quantitative evaluation of model performance, we studied *recall@k* introduced by Hodosh et al. (2013) and frequently applied in VGS model literature. In the present case, recall@k measures performance of the trained models on image search, given an input utterance as a query ("speech-to-image search"), and on automatic image caption search, given an image as a query ("image-to-speech search", sometimes also referred to as automatic image annotation; see also Harwath et al., 2016 and Chrupała et al., 2017).

For measuring recall@k, spoken captions and images from a test dataset are presented to speech encoder and image encoder branches of the model, respectively, resulting in speech and image embedding vectors. In speech-to-image search task, the similarity of each speech sample with all test images is then calculated by applying a similarity metric (here: cosine similarity) to their embedding vectors, and *k* nearest matches are maintained. Recall@k is then obtained as the percentage of utterances for which the image corresponding to the utterance is within the *k* closest matches. Similarly, for image-to-speech search task, recall@k measures the percentage of query images for which the correct caption is within the *k* closest retrieved utterances.

In our experiments, we report recall@10 as it is also commonly used in earlier studies (Harwath et al., 2016; Chrupała et al., 2017).

Quantitative Evaluation of Audiovisual Search Semantics

While previous studies have primarily used recall@k to measure performance in audio-visual alignment tasks, the problem of recall@k is that it is unable to account for semantically relevant matches beyond the pre-defined image-caption pairs of the database (see Kamper et al., 2019). For instance, the data might contain a large number of food pictures, and hence a spoken query such as "There's leftover food on the table" could result in many relevant search results with food in them, but only the one for which the caption was originally created for would be counted as a correct search result. For this reason, Kamper et al. (2019) used human judgments for evaluating semantic retrieval in his VGS model. However, despite crowdsourcing, this can be time consuming and expensive.

In order to objectively evaluate and compare the quality of the learned semantic representations of the alternative speech encoder architectures, we developed a new method to objectively and automatically evaluate semantic similarity between input speech and the corresponding retrieved audio captions. For this purpose, we utilized Word2Vec (Mikolov et al., 2013) and SBERT (Reimers et al., 2019), distributional word semantics models trained on large-scale text data, that allow measurement of semantic similarity between different words (Word2Vec) or sentences (SBERT) in textual form. Since semantic similarity judgements of distributional semantic models correlate highly with human ratings of similarity and synonymity (Landauer and Dumais, 1997, or Günther et al., 2019, and references therein; but see also Nematzadeh et al., 2017 or Deyne et al., 2021, for recent analysis), we use these two models as proxies for human judgement for semantic relatedness between different spoken captions.

With SBERT[2], the semantic similarity of two captions can be obtained simply by taking the cosine similarity of the sentence-level embeddings extracted from the utterance transcripts. However, the maximum similarity score is strongly affected by presence of repeated words in the two compared sentences. An alternative measurement can be obtained by excluding repeating words between the sentences, but we hypothesized that removing of content words might cause unwanted problems with context-dependent embeddings of SBERT. In order to measure semantic similarity of two spoken captions at the word level, we first extracted content words of the utterance transcripts using the Natural Language Toolkit (NLTK) in Python by including nouns, verbs, and adjectives while ignoring other parts of speech. We then calculated semantic relatedness score (SRS $\in$ [0, 1]) between the two utterances as:

$$SRS(reference, candidate) = \frac{1}{N_r}\sum_{i=1}^{N_r} \max\{S_{\mathrm{w2v}}(r_i, c_j) \mid \forall j\} \qquad (3)$$

where S_{w2v} is Word2Vec similarity score between individual words (cosine similarity of the pre-trained word embedding vectors) and r and c are content words in reference and candidate sentences, respectively. In other words, for each content word in the reference utterance, the most semantically similar word is chosen from the candidate utterance, and the total similarity score is the average across all such pairings. By excluding the repeating words between the sentences before SRS calculation, this measurement is then an indicator of semantic relatedness of the utterances while ensuring that the similarity is not simply driven by identical lexical content. In our experiments, we used both SBERT and SRS semantic similarity measurements to test

[2] We used pre-trained SBERT model "paraphrase-distilroberta-base-v1" trained on paraphrase data (Reimers and Gurevych, 2020)

whether audio-to-audio search results produce semantically meaningful outputs even if the utterances do not correspond to the same original image, thereby enabling more representative evaluation of semantic retrieval beyond recall@k.

Selectivity Analysis of Hidden Layer Activations

The literature on interpreting linguistic structure learned by deep neural networks has shown that multiple alternative metrics are needed to understand hidden representations. This is since there is no unanimous view of what "linguistic representations" should look like in such a distributed multi-layer representational systems, and hence it is difficult to operationalize broad concepts such as as "phonemic or lexical knowledge" in terms of specific and sensitive measures to probe the hidden layer activations (see, e.g., Belinkov & Glass, 2020; Chrupała et al., 2020). Given this starting point, our metrics for analyzing the relationship between model activation patterns and linguistic units in the speech input focus on four complementary measures: selectivity of individual nodes in network layers towards specific linguistic units, clusteredness of entire activation patters of a layer, and linear and non-linear separability of layer activations w.r.t. different linguistic unit types. We deliberately focus on statistical and classifier-based measures of analysis that are suitable for basic level categorical data (phone, syllable, or word types), whereas measures such as representational similarity analysis (RSA; Kriegeskorte et al., 2008) used in some other works (e.g., Chrupała et al., 2020) are better suited for non-categorical reference data[3].

This section uses phones as the example units of analysis, but the same analysis process was also carried out for syllables and words in each layer of each of the compared models, as described in section Model Evaluation. As is customary, we use *types* to refer to unique phones in the corpus and *tokens* for individual occurrences of phones in the data.

The first measure, ***node separability***, describes how well activations corresponding to the different phones in the speech input can be separated by individual nodes of a layer. The metric is based on d-prime measure (aka. sensitivity index) from the signal

[3] The main advantage of RSA is its sensitivity to different grades of similarity between the analyzed entities. However, derivation of reference metrics for linguistic representations could be conducted in various ways, including factors such as phonotactics or articulatory attributes for phones, focusing on semantics, syntactic role, or lexical neighborhood density for words, or using human similarity judgements or brain imaging data for any of the units. Different choices on the relative importance of such factors could also lead to different analysis findings.

detection theory. While standard d-prime describes the separation of two normal distributions in terms of how many standard deviations (SDs) their means are apart, *D*-dimensional generalization of the metric can be written as:

$$\bar{d}'_{i,j} = \frac{\bar{\mu}_i - \bar{\mu}_j}{\sqrt{\frac{1}{2}(\sigma_i^2 + \sigma_j^2)}} \tag{4}$$

where $\bar{\mu}_i$ and $\bar{\mu}_j$ indicate the means and σ_i^2 and σ_j^2 SDs of the *D*-dimensional activations (of a layer with *D* nodes) during specific phones $i, j \in \{1, 2, ..., M\}$, respectively. By taking the root-mean-square of across the *D* nodes and then averaging the result across all possible unique pairs of phones, we obtain the multidimensional node separability measure $d' \in [0, \infty]$ for the given layer:

$$d' = \frac{2}{M^2 - M}\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\sqrt{\frac{1}{D}\sum(\bar{d}'_{i,j})^2} \tag{5}$$

The metric is independent of representation space dimensionality. It is zero if all nodes have identical activation distributions for all phone types, and grows with increasing separation of the distributions for different phone types. Intuitively, if individual nodes of a layer specialize in encoding different phone categories, we should observe a high value of *d’* for the given layer.

Our second measure investigates the degree that the distributed activation pattern of an entire layer encodes phonetic identity. We measure this ***clusteredness*** of the representations by applying *k*-means clustering to the extracted activations of each layer, where the number of clusters *k* is specified to be the same as the number of phone types in the corpus (i.e., *k* = *M*; see also Alishahi et al., 2017, for an agglomerative approach). Clustering is initialized randomly, and then all activation vectors get assigned to one of the clusters by the k-means algorithm. The proportions of samples from each phone type in each cluster are then calculated, and each cluster is assigned to represent a unique phone type. The assignment is based on greedy optimization, where the cluster with the highest proportion of samples from a single phone category (i.e., having the highest phone *purity*) is chosen as a representative of that type, and then that cluster and phone type are excluded from the further assignments. The process is repeated until all clusters have been mapped to their best-matching types (with the aforementioned constraints). The overall phonetic purity of the clustering is then measured as the average of the cluster-specific purities w.r.t. to the assigned phone categories. The result is averaged across 5 independent runs of k-means to account for the variance due to the random initialization. Mean and SD of the overall purity across the runs are then reported in the experiments. Purity ranges from 1/*M* (different phones are uniformly distributed across all clusters) to 1 (phones group into perfectly pure clusters in an unsupervised manner).

Besides analyzing the activations of individual nodes and full layers, we use two additional measures to investigate whether the full layers or their node subsets separate between different phone types: ***linear separability*** and ***non-linear separability***, as measured by machine learning classifiers that are trained to classify phones using the activation patterns as features (also known as *diagnostic classifiers*; see also Belinkov & Glass, 2020; Chrupała et al., 2020). For linear separability, we use support vector machines (SVMs) with a linear kernel. For non-linear separability, we use a k-nearest neighbors (KNN) classifier. Both classifiers are trained with a large number of phone tokens from each phone type, and then tested on held-out tokens from the same types (see section Model Evaluation for details). Separability is measured in terms of unweighted average recall (UAR %), corresponding to the average of phone-specific classification accuracies.

On top of the four reported metrics, we also calculated a number of other metrics. For the node selectivity, we measured the so-called Phoneme Selectivity Index (PSI) by Mesgarani et al. (2014). Since PSI was very highly correlated with the d-prime separability across the different layers and test conditions, we do not report it separately. In addition, we measured the difference and ratio between cross- vs. within-type cosine distances of layer activation vectors as a measure of separability. However, we found the k-means-based metric more representative and straightforward to interpret for the phenomenon of interest. Finally, we also calculated overall classification accuracies (aka. weighted average recall / WAR) for the SVM and KNN classifiers. Since WAR is simply the proportion of tokens correctly classified, it is biased towards classification accuracy of more frequent phones. However, UAR and WAR were also highly correlated, and therefore we report UAR only.

In addition, we initially performed word-level analyses separately for content words only, as the we hypothesized that the audiovisual learning paradigm may support learning of nouns and verbs better than, e.g., function words. However, the results were highly correlated to those using all word types in the analyses. For the sake of clarity, we only report the results for words from all parts of speech.

Temporal Analysis of Hidden Layer Activations

We also compared temporal dynamics of the network activations with ground truth phone, syllable, and word boundaries. Our question was whether the temporal activation patterns would somehow reflect the underlying linguistic unit boundaries, i.e., whether the models reflect emergent speech segmentation capabilities even though they were not trained for such a purpose. In earlier work, Harwath and Glass (2019) reported that activation magnitudes of a VGS model (similar to our present CNN0) were related to phone boundaries on TIMIT corpus (Garofolo et al., 1993) after the model had been trained on Places Audio Caption Dataset (Harwath et al., 2016). Our present aim was to replicate the finding on other corpora, and to investigate segmentation of syllables and words in addition to phones.

In order to do so, we first measured activations of each layer for each input utterance as a function of time, and then characterized the overall temporal dynamics using a 1-D time-series representation for the given input. We then compared the peaks of this representation with known linguistic unit boundaries. We investigated three types of 1-D representations for the network temporal dynamics: activation magnitudes $m_l[t] \in [0, \infty]$ (from Harwath & Glass, 2019), instantaneous normalized entropy $h_l[t] \in [0, 1]$, and linear regression from instantaneous node activations to pseudo-likelihoods of unit boundaries, $r_l[t] \in [-\infty, \infty]$. The first one is simply the L2-norm of activations of all nodes *n* in layer *l* at time *t*. Entropy was defined as

$$h_l[t] = -\frac{\sum_{n=1}^{D} \hat{a}_n[t] \log_2 (\hat{a}_n[t])}{\log_2(N)} \qquad (6)$$

where $\hat{a}_n[t]$ denotes the node- and layer-specific activations after the sum of activations has been normalized to 1 for each t and l, and where D is the total number of nodes in the given layer. In essence, $m_l[t]$ quantifies how well the input matches to the receptive fields of the filters in each layer, whereas $h_l[t]$ quantifies how the activity of the layer is distributed: small values close to zero indicate that only few neurons are active at the given time, whereas $h_l[t]$ close to 1 (high entropy) means that all nodes have very similar activation levels and hence little information is transmitted by the instantaneous activations.

Linear regression was performed by first creating a target temporal signal for each utterance, where the signal had a Gaussian kernel with a maximum amplitude of one centered at each unit boundary (see Landsiedel et al., 2011, for a similar approach for syllable nuclei detection). Duration of the kernels was set so that approximately 95% of the kernel mass was within ±20 ms from the annotated target boundary for each phone and within ±40 ms for syllables and words. This was done to account for the uncertainty in defining the exact unit boundary positions in time (see, e.g., Kvale, 1993). Then an ordinary least-squares linear mapping was estimated from the instantaneous node activations to the target signal. After estimating the mapping, the regression representation $r_l[t]$ was obtained by applying the mapping to all activations in the corpus, representing the estimated "score" that a boundary is located at each temporal position. A separate mapping model was trained for phones, syllables, and words to be used in respective evaluations. Due to computational constraints, a subsample of 2,500 target corpus utterances was always used to train the regression model.

The first two representations, $m_l[t]$ and $h_l[t]$, were normalized to have zero mean and unit variance at the utterance-level before further analysis. Due to the nature of the regression targets, $r_l[t]$ was already targeted between 0 and 1 (except for regression inaccuracies) and did not require further normalization.

Following Harwath and Glass (2019), a difference of a Gaussian filter of $\sigma = 5$ ms was applied to the normalized 1-D curves from L2-norm and entropy to measure their rate of change, followed by filter delay correction (see Fig. 3 for visualization). After preprocessing each of the 1-D representations, peak-picking was applied to detect local maxima in the rate of change in magnitude or entropy or maximum boundary score in the regression output. The outputs of the peak-picking were then considered as boundary hypotheses and compared to annotated linguistic unit boundaries. Sensitivity of the peak picking algorithm was controlled by a detection threshold θ_d—the minimum required difference between the last local minimum and current local maximum in order for the maximum to be considered as a peak.

Phone segmentation was evaluated using standard metrics, where a reference phone boundary was considered as correctly detected if the algorithm had produced a hypothesized boundary within ±20 ms from the reference (Räsänen et al., 2009). Similar procedure was used for syllable and word segmentation but using a ±50-ms criterion for the detection, as used in the earlier literature on syllable segmentation (Räsänen et al., 2018, and references therein). Recall (proportion of boundaries detected), precision (proportion of hypothesized boundaries correct), and F-score (harmonic mean of the previous two) were then calculated as the primary metrics for segmentation.

For conciseness, we only report results for the optimal θ_d determined separately for phones, syllables, and words across the full test set. This is since we are primarily interested in whether the model activation patterns reflect boundaries of linguistic units, not whether the settings of our algorithm generalize to novel test conditions as required for proper speech segmentation algorithms. For the same reason, the linear regression model was trained on the same data as to which it was then applied to.

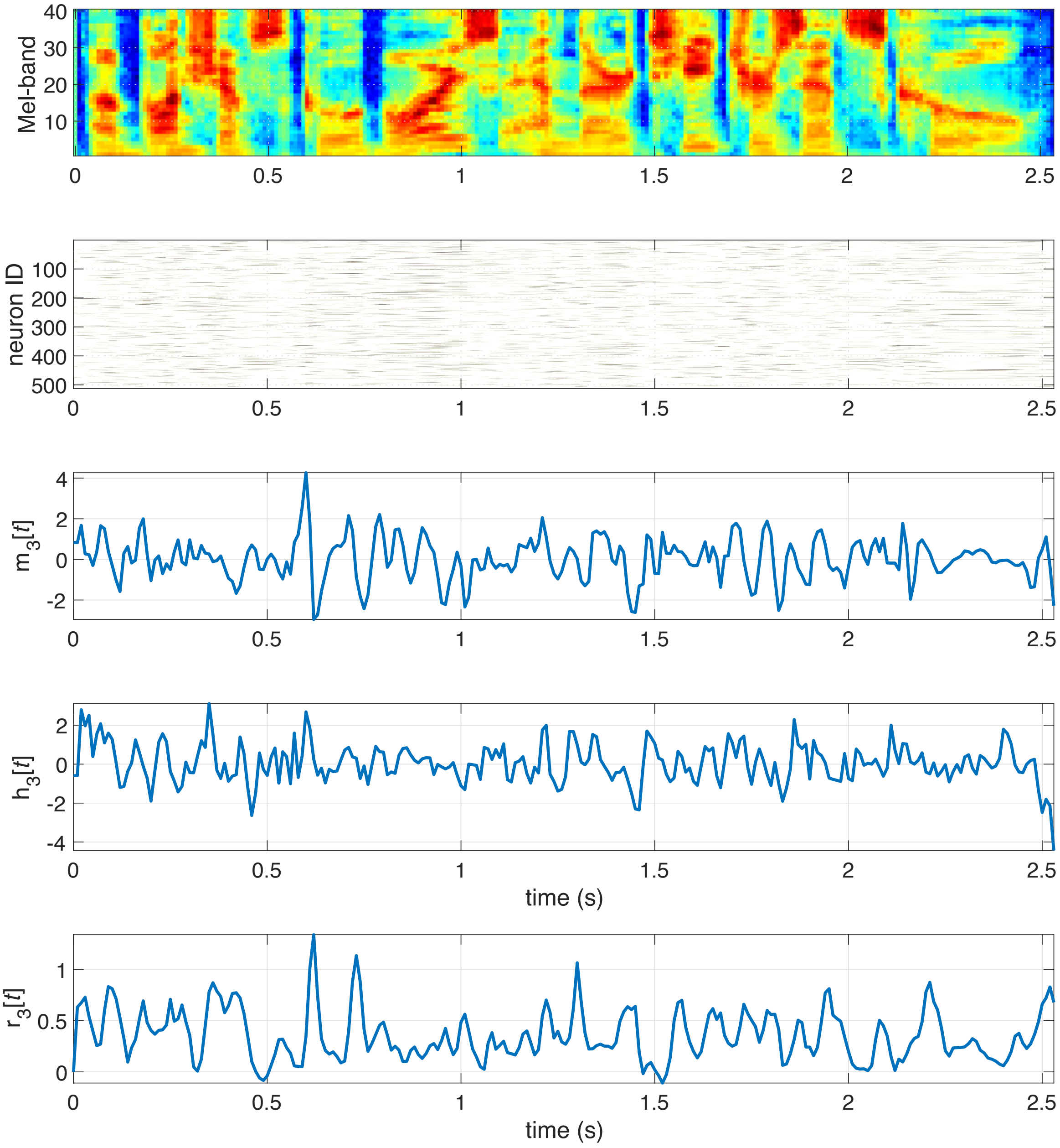


Figure 3. ***An example of the temporal analysis process for a spoken utterance. Top: original speech input as a log-Mel spectrogram. Second panel: corresponding neuron activations from layer 3 of the CNN1 model. Third panel: instantaneous magnitude of the node activations. Fourth panel: instantaneous entropy of the activations. Bottom panel: linear regression from the node activations to phone boundary scores. Segment boundaries are obtained from the curves with peak-picking. The first two activation curves are z-score normalized.***

Experimental Setup

Data

We investigated model training and representation analysis with both synthetic speech and real speech. The use of synthetic speech allows highly controlled experiments with clean signals, limited number of speakers, and accurate ground truths for the linguistic units in the speech data. In contrast, real speech is, by definition, more natural, and comes with higher within- and across-speaker acoustic variability. This is especially due to crowd-sourced nature of the speech audio in the existing audiovisual datasets. Therefore it was also of interest whether analysis findings from synthetic data would generalize to real speech, but also whether the VGS models trained on synthetic data would generalize to real speech and vice versa, as this also affects the general applicability of synthetic data for computational research on language learning in general.

For synthetic training and evaluation data, we used SPEECH-COCO dataset (Havard et al., 2017) based on MSCOCO (Lin et al., 2014). MSCOCO was originally collected to train computer vision systems, and consists of images paired with their verbal descriptions provided by human subjects. The dataset focuses on object recognition in context, and thus provides a variety of images of scenes and objects commonly observed in everyday life. The dataset includes a total of 123,287 images and covers 11 super-categories (e.g. animal, food, furniture etc.) and 91 common object categories (e.g. dog, pizza, chair) of which 82 categories contain more than 5,000 labeled samples (cases). Each image is paired by at least five written captions describing the scene using the object categories. SPEECH-COCO (Havard et al., 2017) was derived from MSCOCO by using a speech synthesizer to create spoken captions for more than 600k of the image descriptions in the original MSCOCO dataset (Chen et al., 2015). The speech was generated using a commercial Voxygen text-to-speech (TTS) system, which is a concatenative TTS system with four UK and four US English voices. SPEECH-COCO has the same datasplit as in MSCOCO 2014; the training set includes 82,783 images with corresponding 414,113 image descriptions and the validation set consists of 40,504 images paired with 202,654 captions. Each audio sample comes with synthesizer metadata on the audio caption, such as timestamps and identities of phones, syllables, and words synthesized, and we treated these as the gold standard phonetic reference of our speech data.

In our experiments, we randomly sampled two sets of 5k images from the original SPEECH-COCO validation set to be used for model validation and test data. The rest of the validation set (~30k images) were included in the training data. As a result, there were a total number of 113,287 images and 566,432 spoken captions for training, and two sets of 5,000 images with 25,000 utterances for validation and testing. In the linguistic representation analyses, one randomly chosen caption was used for each test image.

For real speech -based model training, we used Places400K corpus. It is based on Places205 image database (Zhou et al., 2014) that contains over 2.5 million images illustrating 205 different everyday scene types. Places Audio Caption (English) 400K data (Harwath et al., 2016) contains approximately 400,000 speech captions created for an equal number of images from Places205. Audio captions were collected from hundreds of speakers through Amazon's Mechanical Turk. During the data collection process, the user was asked to provide a free-form speech for each image describing the salient objects in it. There is only one verbal description per image, but compared to SPEECH-COCO, the average duration of the utterances is longer. The authors of Places400K automatically transcribed the spoken captions using automatic speech recognition (ASR) and reported approximately 23% word error rate for the results. Since we only use the text captions for the semantic retrieval analysis, and since there is no obvious reason why the ASR errors would bias the relative comparison of alternative models in the task, we find quality of the captions acceptable for the purpose. We split 10,000 validation and 10,000 testing images-caption pairs from the full dataset, and used the rest (392,385 image-speech pairs) for model training. In contrast to SPEECH-COCO that only consists of 8 different synthetic voices, Places400K represents notable variety in speakers and speaking styles due to its crowdsourced nature. Hence, their use in our experiments allows us to probe the impact of acoustical variety on learned model representations.

Since Places400K does not have existing phonetic annotations, we used a third corpus to investigate model representations with real speech. For this purpose, the so-called "*Large Brent*" subset (see Rytting et al., 2010) of Brent-Siskind corpus (Brend & Siskind, 2001) was used. The corpus consists of recordings of infant-caregiver interactions from four preverbal babies. The transcripts of the adult speech were transformed into phone- and word-level annotations using ASR-based forced-alignment by Rytting et al. (2010). In Räsänen et al. (2018), the transcripts were further syllabified based on the phone strings, and we use the 6,253 utterances with phone-, syllable- and word-level annotations as described in that paper. In contrast to SPEECH-COCO and Places400K, audio quality of Brent is significantly worse due to its at-home recordings. It also represents very different speaking style from the two other corpora. Therefore it was of interest to compare analysis results from SPEECH-COCO synthetic speech to those of Brent. Note that since Brent consists of audio only, it was not possible to evaluate audiovisual retrieval on that corpus.

The three datasets and their roles in model training, validation, audiovisual search evaluation, and model representation analysis are summarized in Tables 1 and 2.

Table 1. ***Datasets used for audiovisual model training (train), early stopping and model selection (dev), and analysis of audiovisual semantic retrieval (test). N refers to the number of utterances and corresponding spoken captions.***

	N train	*N* dev	*N* test	speakers	speaking style
Places400K	382,385 (images & utt.)	10,000	10,000	2,683	crowsourced real
SPEECH-COCO	113,287 (images) x 5 (utt.)	5,000	5,000	8 (4 x US, 4 x UK)	synthetic

Table 2. ***Datasets used for analyzing model representations with respect to linguistic annotations. SPEECH-COCO test set is the same as the one used to test audiovisual semantic retrieval (listed in Table 1).***

					phones		syllables		words	
	N test	duration	speakers	style	*types*	*tokens*	*types*	*tokens*	*types*	*tokens*
SPEECH-COCO	5,000	214 min	8 (4 x US, 4 x UK)	synthetic	47	190,629	232	51,855	168	37,558
Brent	6,253	93 min	4	real IDS	44	71,569	113	13,849	86	13,218

Model Training

Since all inputs to the models had to be of equal length, we first zero-padded/truncated input log-Mel spectra to the length of 1024 frames (10.24 s) and 512 frames (5.12 s) for the Places and COCO datasets, respectively. Following the literature, the embedding space dimensionality *D* was set to 1024 for Places (Harwatha & Glass, 2017) and 512 for COCO (Chrupała et al., 2017). The resulting model parameter counts were approximately 21.2M (CNN0), 9.8M (CNN1), and 10.1M (RNN) parameters on Places, and 13.9M (CNN0), 7.4M (CNN1), and 8.8M (RNN) on COCO.

For model training, we used mini-batch size of 120 triplets and shuffled mini-batch sample assignments after each epoch. A new set of negative samples was also drawn for each epoch. Adam optimizer with a fixed learning rate of to 1e-4 was used. Early stopping based on development set recall@10 score with patience of 5 was used to control the training, and the best model according to the validation recall was then used for testing purposes. In all models, rectifier linear units were used as activation functions for all convolutional layers and hyperbolic tangents were applied for recurrent layers. Based on pilot experiments with several triplet loss margins, a margin *M* = 0.2 was ultimately chosen based on its superior audiovisual retrieval performance.

Evaluation Protocol

For speech-to-image and image-to-speech retrieval tasks, we measured recall@10 for randomly sampled subsets of 1,000 image-caption pairs from our test data sets. For

Places, we used all 10,000 test set image-caption pairs and sampled 1,000 pairs, measured the recall, and repeated the processes until distribution of recall scores for all testing samples converged to a normal distribution. We report recall scores then based on the mean and standard deviation of the obtained distribution. For COCO, we used all the possible 25,000 image-caption pairs (5,000 unique images, each paired with 5 captions), first divided them to five subsets of 5,000 image-caption pairs, and then sub-sampled a random set of 1,000 image-caption pairs for each subset until the mean recall@10 scores across all subsets converged.

For objective evaluation of audio-to-audio search results, semantic relatedness score (SRS) was calculated on the orthographic test set captions of both corpora. On Places, all the 10,000 test utterances were used, whereas one randomly chosen caption was used for each of the 5,000 images on COCO. COCO captions were textual to begin with, and we used the Places captions generated by Harwath et al. (2016) using ASR. For each of the test set query utterances, the corresponding textual caption was compared to the captions of the top 5 retrieved utterances using the SRS score in Eq. (3). As a reference, the process was repeated for 5-top dissimilar and 5-random captions. The analyses were conducted for all test set captions whose content words passed spell checking based on the Word2Vec model. This left us with 9,242 and 4,877 utterances for Places and COCO test sets for semantic similarity measurement, respectively.

For linguistic analyses, audio encoder activations were first recorded for all utterances in the test corpora. Similarly to Alishahi et al. (2017), activations were averaged across the duration of each annotated phone, syllable, or word token, so that each resulting activation sample corresponded to one linguistic token at the given level of analysis. All unit types with less than 50 tokens in a test set were then discarded from the analyses, and the resulting type and token counts are summarized in Table 2. Classifier-based separability analyses were conducted for all the tokens of a test corpus, where 80% of the tokens were used for training and 20% for testing of the classifiers (ensuring that the tokens in training and testing were from different utterances). For the KNN, $k = 15$ nearest neighbors were used based on initial optimization on a subset of data. Node selectivity and clustering analyses were conducted for a random sample of 50 tokens from each type, sampling uniformly and randomly from the full pool of test set tokens. This was done to ensure that the reported metrics reflect equally all the phonetic/syllabic/lexical types instead of being strongly biased towards the most frequent ones. The same random samples were used for all models and layers. Note that the classifier-based UAR metric is inherently unaffected by test class frequencies, and no such sampling procedure was needed for the classifier analyses. For the temporal segmentation analyses, original activations for all utterances in the test corpora were used. In addition, utterance onsets and offsets were automatically scored as correctly detected (and not counting any additional algorithm boundaries at those locations as insertions), as their detection can be considered as trivial due to the definition of an utterance as a stretch of speech separated by pauses or a change in speaker turn.

As a reference point, we also report measures obtained from the same models before their training (i.e., using the initial random parameters). This allows us to disentangle any effects of audiovisual learning from the potential benefits of simply performing a series of random non-linear transformations on the input speech data (see Chrupała et al., 2020, for a discussion).

Results

Results of the experiments are divided into two parts: first, we ensure that all three model variants have learned the audiovisual mapping problem, and investigate their relative performance in capturing the semantics between the two modalities. In the second part, we focus on the internal representations used by the models in the multimodal learning task.

Validation of Model Performance

We first ensured that training of all models converged to a meaningful solution of the audiovisual learning task by examining their validation losses and recall@10 scores (Fig. 4). This was also the case, and all three models obtained quite comparable training and validation losses and recall scores on both corpora despite their architectural differences. The only exception to the rule was CNN1, which exhibited superior recall and slightly lower loss on Places validation set compared to the CNN0 and RNN models. Monotonic convergence of all the measures suggests that there was no overfitting in any of the models.

The corresponding test set recall@10 measures on both speech-to-image and image-to-speech search tasks are shown in Table 3. The results for all models are very close to each other, with exact ranking depending on the dataset and task type. In general, the performance between the present models and those reported by Harwath et al. (2016) and Harwath and Glass (2017) are also within similar range. However, exact comparison is not possible, as the details of the test set were not identical due to different sampling strategies. For an unknown reason, our implementation of CNN0 replicated from Harwath and Glass (2017) falls behind the original study on both speech-to-image and image-to-speech search on Places. On the other hand, our CNN1 is similar to results from Harwath and Glass (2017) in performance. RHN-RNN results by Chrupała et al. (2017) are also shown in Table 3 as a reference, although they used different synthesized captions for the COCO data and a larger image search space. In general, the within-corpus results show that the three compared models all succeed in the audiovisual learning task, and they do so with a comparable performance.

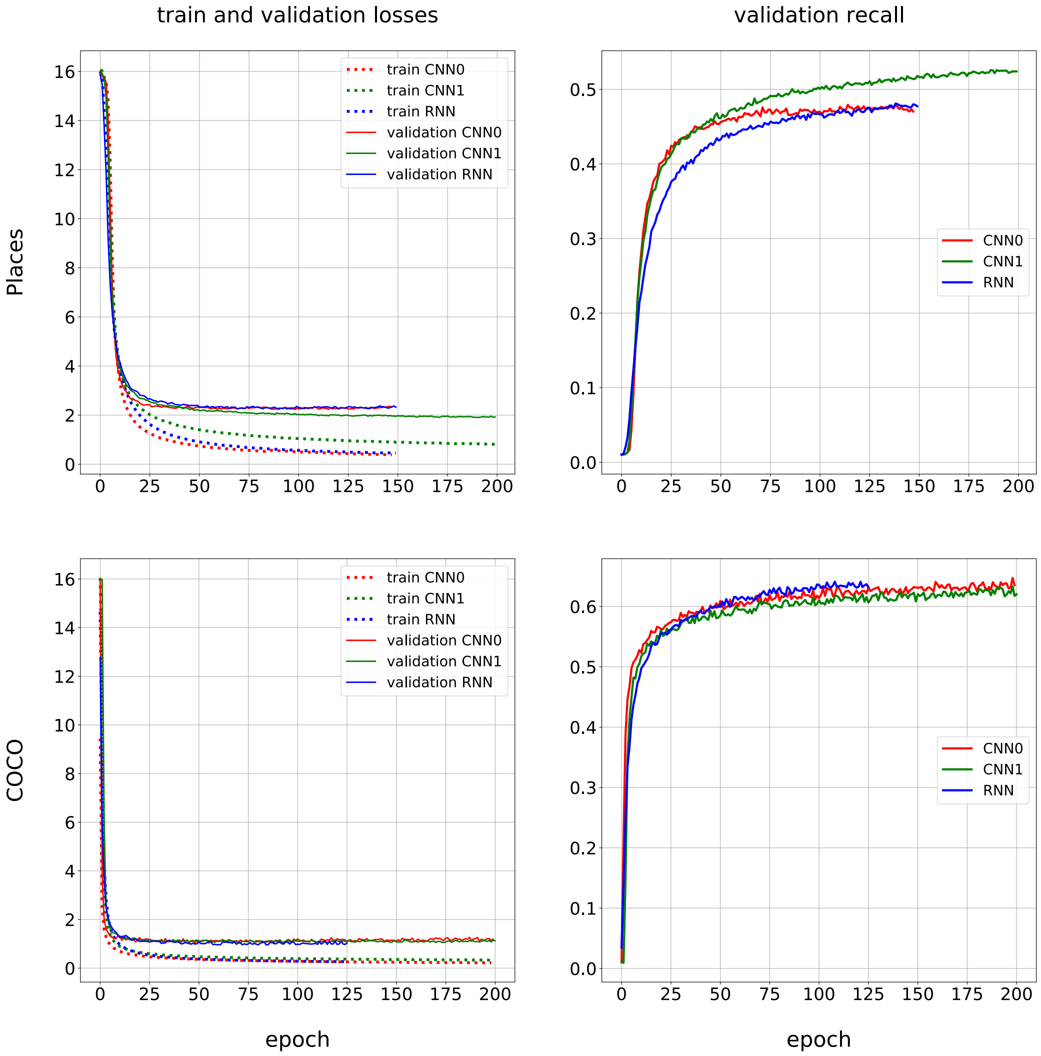
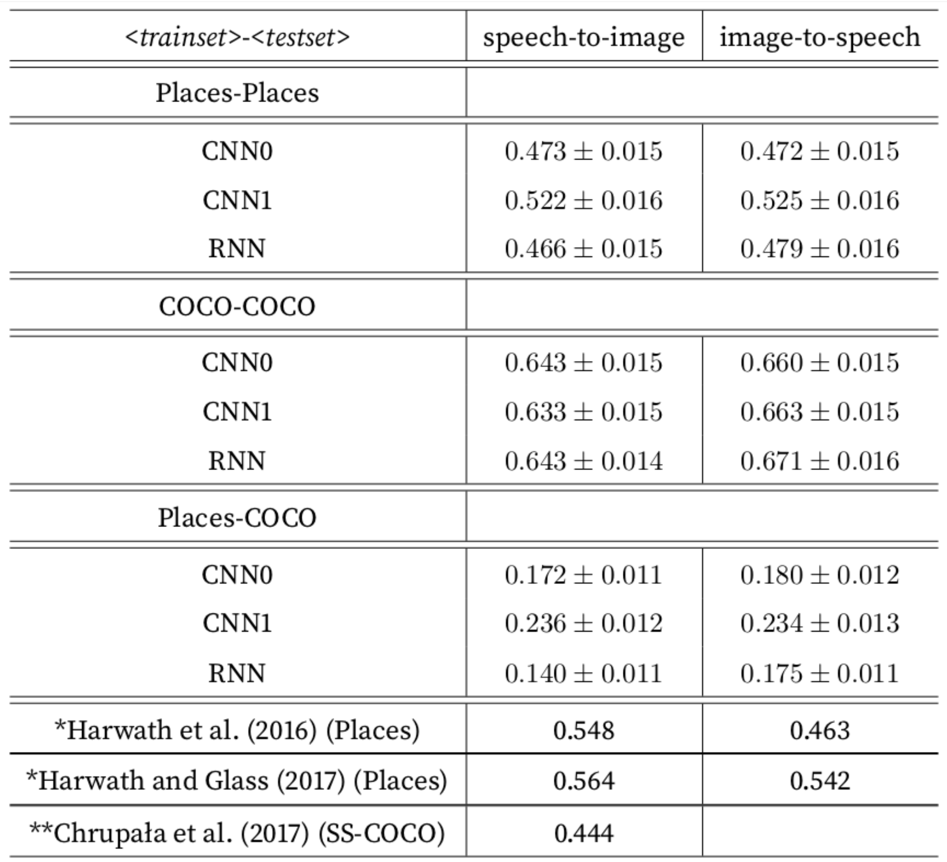


Figure 4. ***Training and validation losses and validation recall@10 scores for the three compared models as a function of training epoch. Top: Places corpus. Bottom: COCO corpus. Left: triplet loss scores. Right: recall@10 scores.***

Table 3: ***Recall@10 values obtained for "speech-to-image" (left) and "image-to-speech" (right). Means and SDs across different samplings of 1,000-sample search spaces are shown. Results from earlier two models tested on Places400K are shown as reference (*) (updated numbers from Harwath et al., 2018instead of the original papers). Note that the experimental setups are not identical, and therefore detailed comparison of numbers is not possible. Results from Chrupała et al. (2017) on another synthesized corpus, SS-COCO, are also shown (**), although they used a larger search space of 5,000 images in their experiments.***



<*trainset*>-<*testset*>	speech-to-image	image-to-speech
Places-Places		
CNN0	0.473 ± 0.015	0.472 ± 0.015
CNN1	0.522 ± 0.016	0.525 ± 0.016
RNN	0.466 ± 0.015	0.479 ± 0.016
COCO-COCO		
CNN0	0.643 ± 0.015	0.660 ± 0.015
CNN1	0.633 ± 0.015	0.663 ± 0.015
RNN	0.643 ± 0.014	0.671 ± 0.016
Places-COCO		
CNN0	0.172 ± 0.011	0.180 ± 0.012
CNN1	0.236 ± 0.012	0.234 ± 0.013
RNN	0.140 ± 0.011	0.175 ± 0.011
*Harwath et al. (2016) (Places)	0.548	0.463
*Harwath and Glass (2017) (Places)	0.564	0.542
**Chrupała et al. (2017) (SS-COCO)	0.444	

In terms of models trained on Places and tested on COCO (bottom part of Table 3, the performance of the models degrades substantially from the within-corpus experiments, even though the performance is still far above chance-level (0.01). Whether this is due to differences in acoustic signals (synthetic vs. real speech) or semantics of the images (common objects vs. scenes of "places") is unclear, although linguistic analyses discussed in section Results suggest that the acoustic mismatch may not be the issue. Among the models, CNN1 generalizes across corpora somewhat better than the CNN0 and RNN models.

Qualitative Analysis of Semantic Retrieval

To further investigate how speech embeddings capture semantic similarities between image and speech modalities, we manually verified a number of retrieval results using the embeddings derived from the speech or image data. Table 4 shows an example of speech-to-speech search obtained using the speech embeddings. As can be observed, the first five most similar captions are semantically connected to the query caption. In the extracted examples, the query utterances and resulted utterances either include same objects or activities or share the same super-category (food, animal, etc.). Note that this matching is not possible by trivial matching of acoustic patterns alone due to lack of temporal alignment between the utterances. This indicates that the model has learned to link semantically similar utterances to each other without any supervised learning.

Table 4. ***Example output for speech-to-speech search using the CNN1 model on Places corpus. Query utterance and utterance transcripts corresponding to the five closest utterance embeddings are shown (spelling as they appear in the ASR-generated transcriptions).***

query	**two cars traveling on a narrow suspension bridge under a blue cloudy sky.**
1	the sky is blue with lots of clouds.
2	the parking lot is empty the sky is cloudy.
3	a blue fairy goes under a blue bridge in a body of water on a clear sunny day with white clouds in the background and green tree surrounding it.
4	train crossing large bridge over bright blue water by the sun's going down with orange blue cloudy skies.
5	a blue cloudy sky bass casting a dark shadow on a roll of cars that are on the street and heavy traffic and there are trees on the side of the road.

Table 5. ***Example output for speech-to-image search using the CNN1 model on Places corpus. Caption of the query utterance (as it appears in annotation files) and images corresponding to the five closest image embeddings are shown.***

a small girl with blue and white striped shirt play.

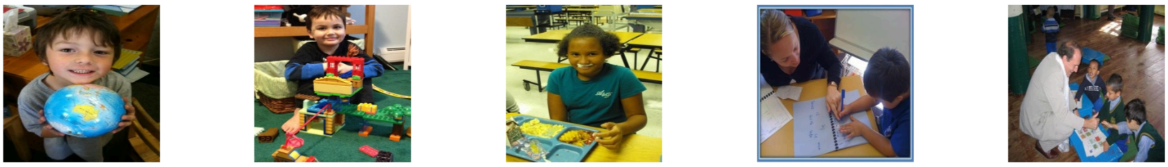


Table 6. ***Example output for image-to-speech search using CNN1 model on Places corpus. Left: query image. Right: transcripts of the utterances corresponding to the five closest utterance embeddings (spellings as they appear in annotation files).***

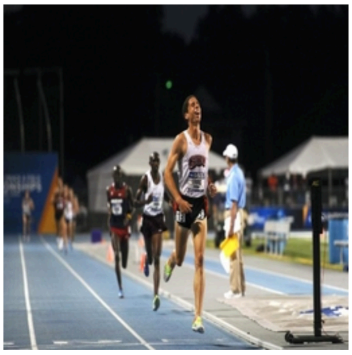
query image	5 highest scoring audio captions
	• google people are racing around checked the first mermaid is wearing a blue uniform. • pictures of man running in a race on the right track if people are around and starts with hurdles on the right. • a woman running on a track there are people looking on in the center of the track are cameron and co. • and there are several people standing along the back of the truck watching them. • two women are running around a track there one is wearing blue and red and the other is black and white.

Table 5 illustrates an example when spoken captions are used to find five best matching images. In the majority of observed samples of speech-to-image search, the resulting pictures contain objects corresponding to one or more of the content words spoken in the query caption. Finally, Table 6 shows five top similar captions resulting from a search based on a query image. In this example, as in the most cases of image-to-speech search results, the extracted captions are semantically related to the query utterance. While these examples are shown for the CNN1 model only, we manually verified that all the three models were able to extract semantic relations between audio captions and images in a qualitatively similar manner.

Evaluation of Semantic Relatedness

By using our Word2Vec and SBERT-based measures of semantic relatedness, we calculated semantic similarity scores between captions corresponding to the five closest, five furthest, and five random embeddings with respect to every possible query utterance and corresponding caption drawn from the test set. For SRS, the measurement was done separately for all content words, and for content words after removing orthographically matching words from the compared captions. Table 7 shows the mean and SD SRS and SBERT similarity scores from the analysis. Fig. 5 also illustrates the obtained distributions of semantic similarities, including SRS with and without the repeating content words, when using the CNN1 model (results for the CNN0 and RNN are essentially similar and hence not shown separately). To further measure the degree that each model managed to capture the semantics of the data, we also quantified the difference in SRS and SBERT similarity score distributions between the top 5 similar captions and the 5 most distant/random captions using the Wilcoxon ranksum statistic. The corresponding test statistics are reported in Appendix C.

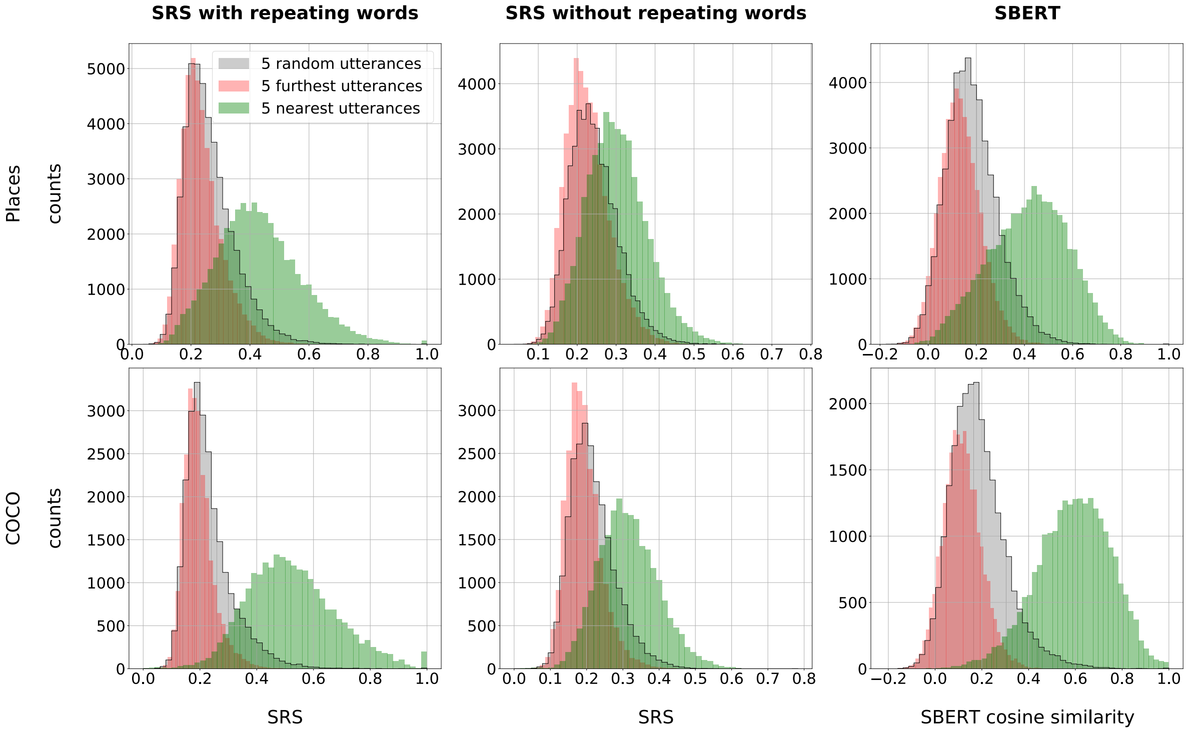


Figure 5. ***Semantic relatedness scores (SRS) and SBERT similarity scores for speech-to-speech embedding search results for the CNN1 model. The graphs show the distributions of similarities between query utterances and the five nearest, the five most distant, and five random captions collected for all test utterances.***

In all the models, semantic similarity of the nearest utterances is significantly higher than in a random sample of utterance pairs ($p < 0.001$ for all comparisons; see Appendix C for details). This is also the case after ignoring repeated words in the semantic similarity calculation. On the other hand, semantic similarity between the query caption and captions of the most distant embeddings are largely overlapping with distances to random embeddings. This likely reflects the use of margin in the VGS model loss function in Eq. (2), which basically constrains the model to focus on the structure of the multimodal embedding space only in the neighborhood of each data point. In contrast, different "degrees of semantic unrelatedness" are not captured by the model, as long as the embeddings of unrelated input pairs are already sufficiently distinct. Although this feature is already implicitly built in to the loss function, the

SRS and SBERT metrics quantitatively demonstrate that the effect also seems to take place in practice.

As for the difference between SRS and SBERT, the SBERT model produces higher average similarity score for the closest utterances and lower average score for the furthest and random utterances compared to SRS model, reflecting its higher capacity in capturing semantics of full sentences with sentence-level embeddings. However, overall there is clear qualitative resemblance between SBERT and the SRS scores with repeating words (see Table 7).

Table 7. ***Medians (Mdn) and standard deviations (SD) of SRS and SBERT scores in case nearest, furthest, and random embeddings w.r.t. query utterances. Left: SRS using all content words in the utterances. Middle: SRS for content words excluding repeating words between query and search result utterances. Right: SBERT scores for full captions.***

	SRS with repeating words						**SRS without repeating words**						**SBERT**					
	5 nearest		5 furthest		random		5 nearest		5 furthest		random		5 nearest		5 furthest		random	
	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD	Mdn	SD
COCO																		
CNN0	0.49	0.16	0.2	0.06	0.21	0.09	0.31	0.08	0.2	0.05	0.21	0.06	0.59	0.16	0.14	0.09	0.17	0.12
CNN1	0.51	0.16	0.19	0.05	0.21	0.09	0.31	0.08	0.19	0.05	0.21	0.06	0.6	0.16	0.11	0.08	0.17	0.12
RNN	0.49	0.16	0.21	0.06	0.21	0.09	0.31	0.08	0.2	0.05	0.21	0.06	0.59	0.16	0.15	0.1	0.17	0.12
Places																		
CNN0	0.37	0.13	0.23	0.07	0.25	0.09	0.29	0.07	0.22	0.06	0.24	0.06	0.4	0.17	0.14	0.09	0.17	0.11
CNN1	0.38	0.13	0.22	0.07	0.25	0.09	0.29	0.08	0.22	0.06	0.24	0.06	0.43	0.16	0.13	0.08	0.17	0.11
RNN	0.36	0.13	0.23	0.07	0.25	0.09	0.29	0.07	0.23	0.06	0.24	0.06	0.4	0.16	0.15	0.09	0.17	0.11

Discussion on Semantic Retrieval Experiments

Overall, the retrieval performance in terms of recall@10, the SRS scores, and the qualitative analyses together confirm that all three models had acquired basic understanding of the semantic relationships between continuous speech and the related visual images. In addition, the three models did so in a comparable manner in terms of our analysis metrics despite the architectural differences between the models. Moreover, the analyses with SRS while excluding repeating words indicates that the semantic similarities among spoken utterances were not merely driven by shared words between the utterances. Instead, the models had learned something about semantic relationships of different words through their occurrences in similar visual contexts. This provides a solid starting point for investigating how the models actually learned to represent the spoken language input as a part of their solution to the audiovisual learning task.

Results from Linguistic Selectivity Analyses

Our primary research question was whether the VGS models exhibit signs of emergent linguistic organization. For this, we studied patterns of selectivity across different linguistic units in the hidden layers of our trained speech encoder models.

Fig. 6 shows the analysis results for models trained and tested on the synthetic COCO data, whereas Fig. 7 shows the same analyses for models trained and tested on real speech (Places and Brent corpora). In addition, Fig. 8 shows the results for the Places-trained model tested on COCO, as the result allows us to disentangle the effects of training data from test data characteristics. Layer numbers of each model correspond to the numbers denoted in Fig. 2. L0 stands for the input log-Mel features.

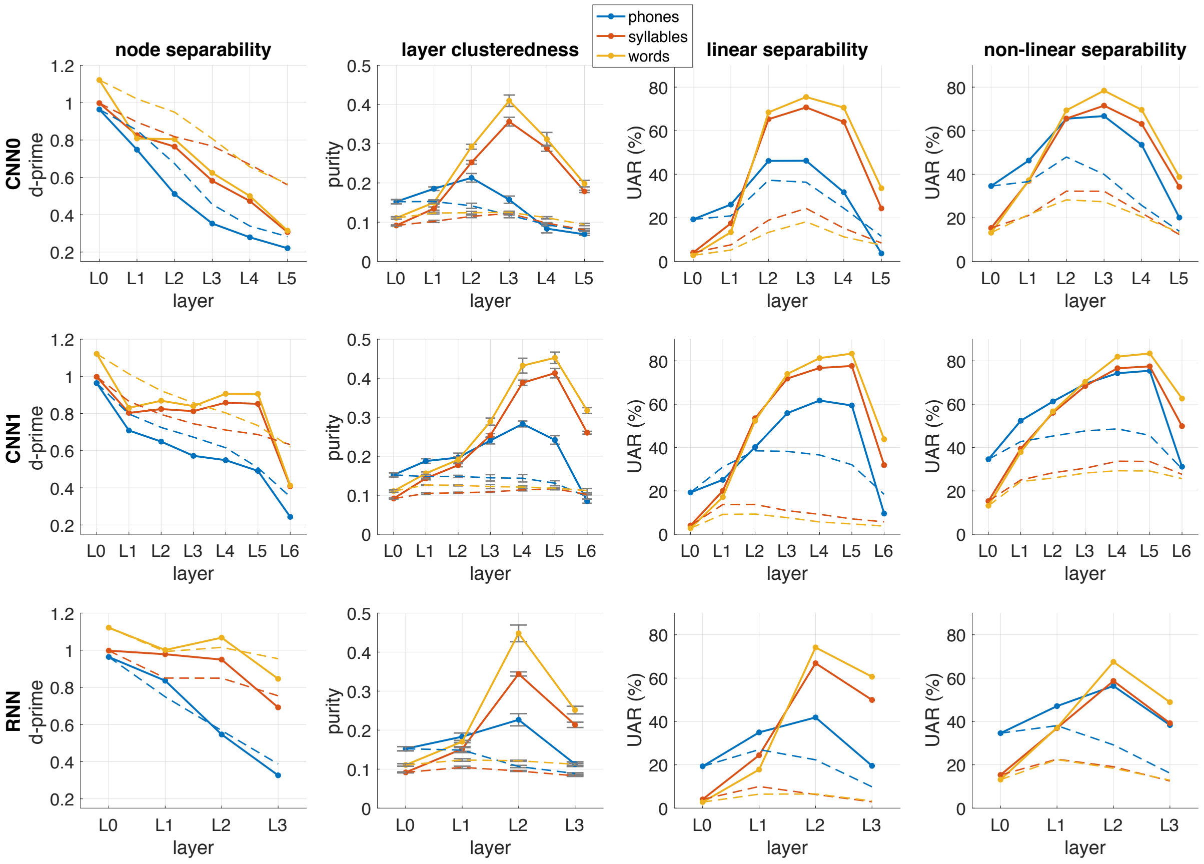


Figure 6. ***Analysis results for models trained and tested on COCO corpus. Each panel row corresponds to one of the models, CNN0, CNN1 or RNN, whereas columns correspond to the four studied selectivity metrics. Blue lines stand for phones, red for syllables, and yellow for words. Solid lines correspond to trained models and dashed lines for the corresponding baseline models before the training. Error bars for clusteredness represent SDs across different runs of the k-means analysis.***

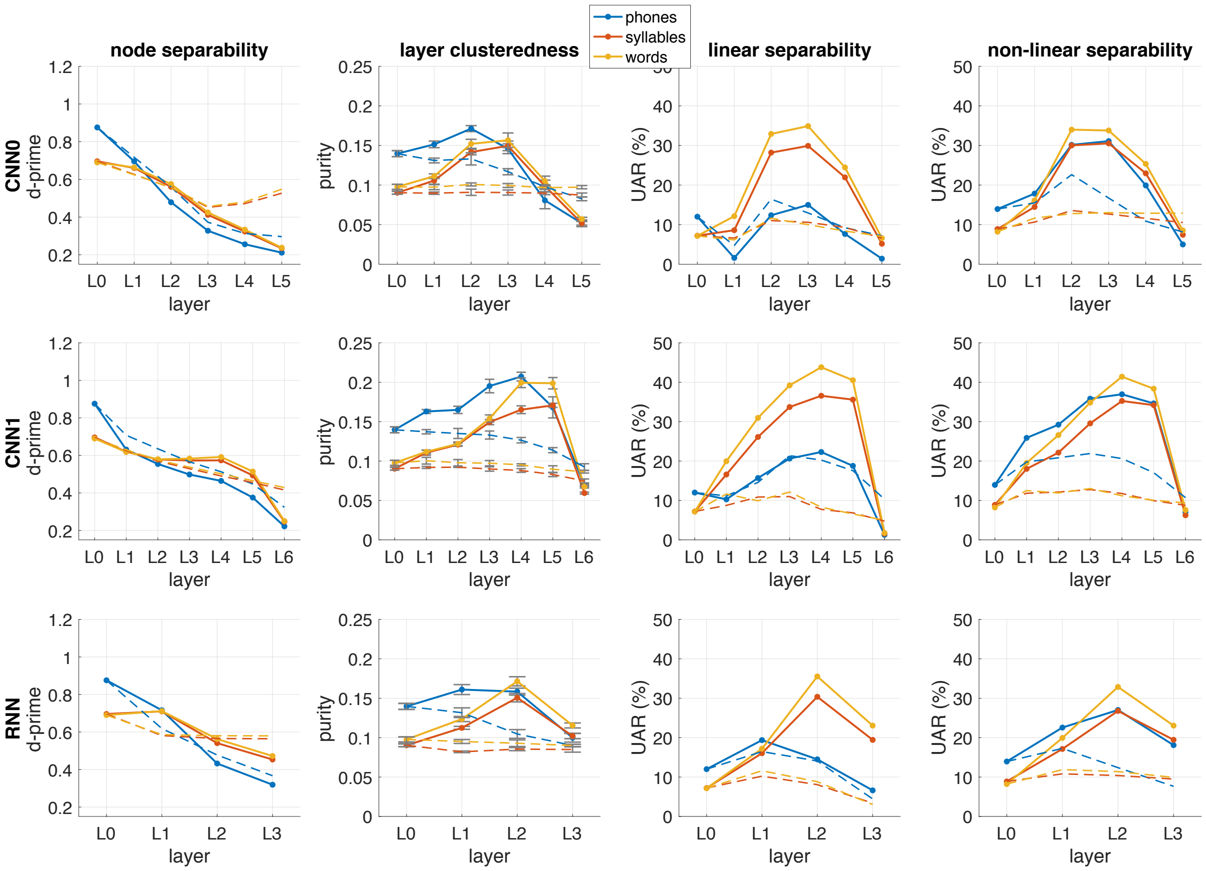


Figure 7. ***Analysis results for models trained on Places corpus and tested on Brent corpus.***

The first observation from the results is that the overall pattern of the analyses is very similar across the different combinations of training and testing corpora. Although the separabilities and classification accuracies are generally lower for Brent than for the synthetic speech from COCO, the relative measures from different layers of each model and in comparison to the untrained baselines are generally similar. Due to this, we will focus on discussing the general findings that hold across the different corpora, and separately mention whenever a finding only applies to a subset of training and testing conditions.

There are several patterns in the results that seem to be robust across the models and datasets. First of all, activation patterns of individual network nodes are poor at separating phone, syllable, or word types from each other, and this is true for all the three model variants (left columns in the figures). In CNN1 and RNN, there is a slight tendency of the node separability to improve in the early or middle layers due to model training. However, even in these cases, node separability is at its maximum or close to the maximum in the input layer. This means that the individual frequency channels

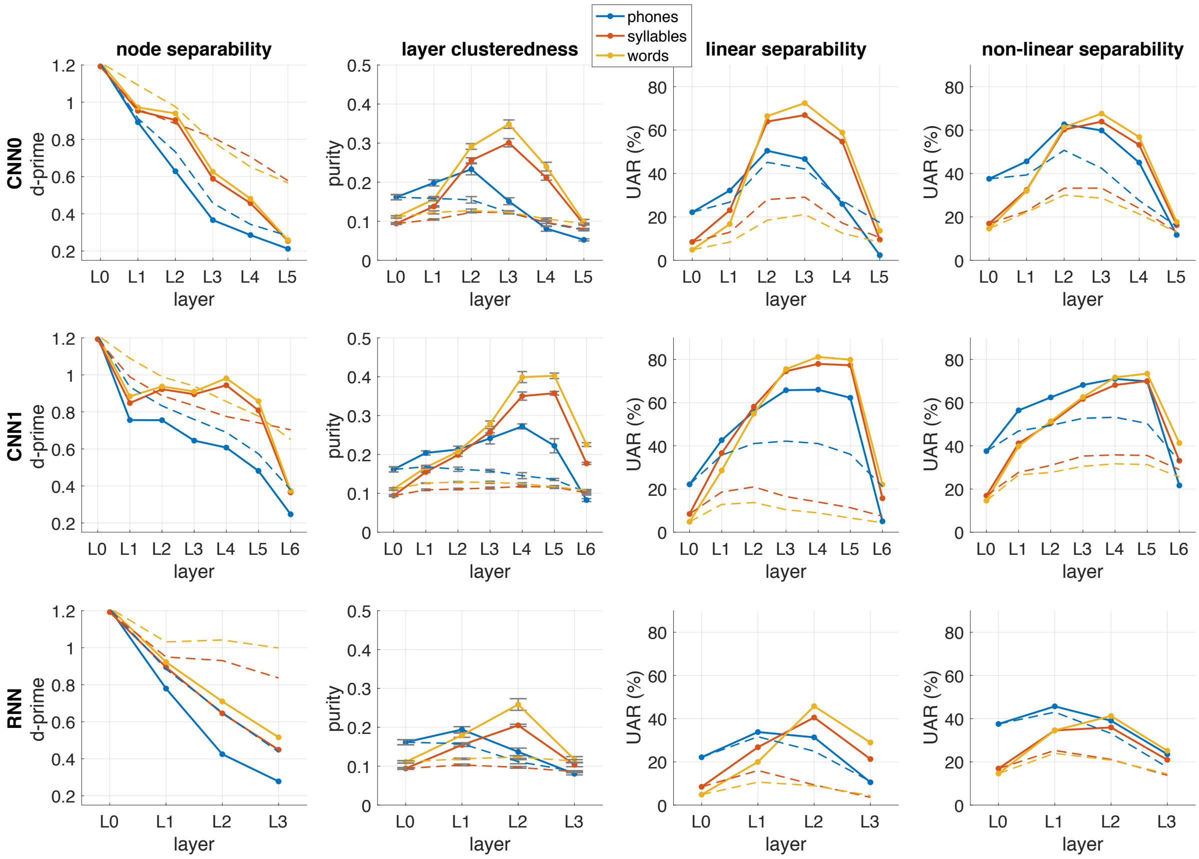


Figure 8. ***Analysis results for models trained on Places corpus and tested on COCO corpus.***

of the Mel-frequency features are more informative of the underlying phonetic/syllabic/lexical identities than the individual nodes in hidden layers of the given models.

The picture changes when all the nodes of a layer are analyzed together (three last columns in each figure). Clusteredness and classification-based measures show clear effects of phonemic, syllabic, and lexical learning taking place in all the models. For instance, unsupervised clustering of CNN1 L4 activations achieves purity of nearly 0.3 in terms of phonetic categories on COCO data (Figs. 6 and 8), whereas clustering of the original log-Mel features leads to purity of 0.16 only. The effect is even larger at the syllabic and lexical levels. CNN1 L5 reaches lexical purity of approx. 0.45 with the vocabulary of 168 unique word types. Purities of RNN and CNN0 have similar patterns to those of CNN1, even though the exact layer in which the purity peaks differs from encoder architecture to others. Purities with the even more numerous syllables follow the same general trend, but with slightly lower purities. Notably, the improvements in syllabic and lexical clusteredness are not simply driven by the increasing receptive field sizes in deeper layers, as the there are no similar improvements in clustering purity for the representations extracted from the untrained but otherwise identical models. Purities measured on the Brent data are lower than those measured from the synthetic data, but there are still clear effects of training with a substantial improvement from the input features. We also conducted post-hoc analyses by measuring purities when the k-means clustering was initialized using the means of phone/syllable/word type activations, and the resulting purity scores were generally 0.1–0.3 higher than those observed for random initialization, but without changing the general pattern across different models, layers and datasets.

The linear and non-linear separability measures also indicate that there are large amounts of phonetic, syllabic, and lexical information encoded in the hidden layer activations of all three models. Representations derived from the middle layers (CNN0), middle to penultimate layers (CNN1), or penultimate layer (RNN) allow relatively accurate classification of syllables and words. On COCO training and testing, the non-linear classification reaches up to approx. 80% accuracy at all three levels of units with the CNN1 model activations, and the result is around 70% for an equivalent model trained on real speech. The corresponding accuracies for log-Mel features are below 40% for phones and below 20% for syllables and words. The qualitative pattern is similar to Brent test data, although the performance numbers are again lower due the more complex and noisy data.

The classification experiments also reveal that phonetic information seems to be embedded within the layers that also encode syllabic or lexical units (cf. the idea of overlapping representational planes in PRIMIR; Werker & Curtin, 2005). The maximum phone classification accuracy is often achieved for the same layers with the best performance on syllables and words, and not for the layers where the receptive field size best fits the typical phone durations (e.g., L2 and L3 in CNN1; see Fig. 2). In addition, phone classification is always clearly more accurate with the non-linear than linear classifier, whereas only minor differences between linear and non-linear classification are observed for syllables or words. This shows that phonetic information becomes the most refined in the same layers that encode syllabic and lexical information (i.e., is concurrently represented with higher levels of linguistic organization), suggesting that phonetic units become encoded in a context-sensitive manner. However, accurate decoding of the phone identities requires non-linear decoding of the activation patterns. Classification analyses also reveal that simply performing a number of random high-dimensional non-linear projections on the data seems to improve classification performance, as observed for the untrained models. However, the improvements are far from the benefits of audiovisual learning.

As for model comparisons, there are certain details that differ between the three architectures, even though all three architectures had very similar performance in the semantic retrieval tasks that they were trained for (see the previous sub-sections). In terms of CNN0 and CNN1, one difference is that phonetic clustering purity is higher for CNN1 than CNN0. In both models, the purity peaks at layers with comparable receptive field lengths (L2 in CNN0: 135 ms, L4 in CNN1: 165 ms) that are somewhat beyond typical phone durations. However, this may be explained by the higher number of nodes and thereby higher representational capability in CNN1 at the given layer (same 512 in all layers for CNN1), whereas CNN0 architecture uses increasing number of nodes as the receptive field increases (as specified in Harwath & Glass, 2017). In addition, non-linear separability of phonetic units is somewhat higher with CNN1 representations than those in CNN0. Patterns for words and syllables are more similar for the two models. In terms of comparison between the CNNs and the RNN, the results are remarkably similar despite their architectural differences. The first recurrent layer (L2) of the RNN is similar to the middle layers of the CNN models, and a similar drop in linguistic selectivity is observed for the last layer of RNN as in those of both CNNs.

Results from Temporal Segmentation Analyses

For the temporal segmentation analyses, we first compared L2-norm, entropy, and linear regression-based representations in the task and found that the regression approach led to somewhat higher segmentation performance than the other two. Due to this, we focus on the regression results here and the full set of L2-norm and entropy-based measures can be found from Appendix B.

Fig. 9 shows the temporal segmentation analysis results for models trained and tested on COCO. Figs. 10 and 11 show the corresponding results for models trained on Places and tested on Brent, and for models trained on Places and tested on COCO, respectively. Baseline performance levels with untrained models are also shown for reference.

Looking at the COCO-COCO results in Fig. 9, there are two key findings: 1) Segmentation performance of all types of units is far above zero, and CNN hidden layers have higher segmentation accuracy for phones and syllables than when using the log-Mel features. 2) Untrained model performance is also relatively high throughout the conditions. This shows that much of the temporal dynamics exhibited by the models (as captured with the present methodology) are already captured by the interaction of input features and non-linear processing steps. In other words, there is only a small effect of training on how the activation patterns reflect linguistic unit boundaries in time. On Places-Brent, the general pattern of results is again very similar to COCO-COCO, but in this case the metrics of the trained models are even closer to the same models with random initial parameters. This is not due to training with real speech, as the performance in the Places-COCO condition again reflects the pattern observed in COCO-COCO. In addition, overall scores for syllables and words are somewhat higher on Brent than on COCO test data. However, this is primarily explained by the substantially shorter average utterance length on Brent, which means that the relative proportion of trivial utterance onset and offset boundaries is much higher on Brent than on COCO.

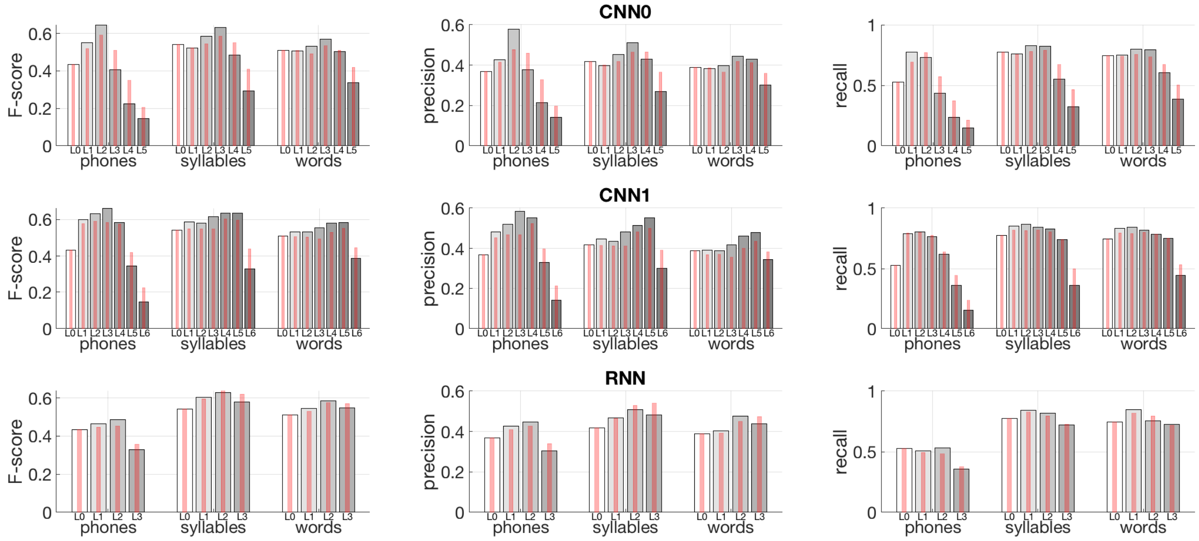


Figure 9. ***Segmentation results from models trained and tested on COCO corpus when using linear regression from activation magnitudes to unit boundary scores as the signal representation. Bars in each subplot represent layer-specific performance metrics for segmenting phones, syllables, and words, respectively. Different layers from input (L0) to last network layer are shown with different shade bars from left to right. Results are shown for F-score (left panels), precision (middle panels), and recall (right panels), and for the three tested models: CNN0 (top), CNN1 (middle), and RNN (bottom). Thin red lines denote baseline performance with untrained models.***

In terms of different layers, there is a small trend of earlier CNN layers to better reflect phonetic unit boundaries while syllable and word boundaries are better accessible from deeper layers. In addition, on COCO test data, the CNN models lead to higher phone and syllable segmentation performance compared to words, whereas on Brent syllables and words are actually more accurately segmented than phones. The best phone segmentation F-score of 0.66 is obtained by CNN1, followed by 0.65 for CNN0, both on COCO training and testing. The corresponding numbers for syllables are 0.64 and 0.63, respectively. The RNN has lower phone segmentation performance than the CNNs, whereas its syllable and word segmentation scores are generally comparable to those of other models (e.g., F-score of 0.64 for syllables on COCO-COCO).

Comparing the current regression-based segmentation results to the L2-norm and entropy-based representations reported in Appendix B, the results are largely qualitatively similar across the approaches. The clearest difference is a modest performance

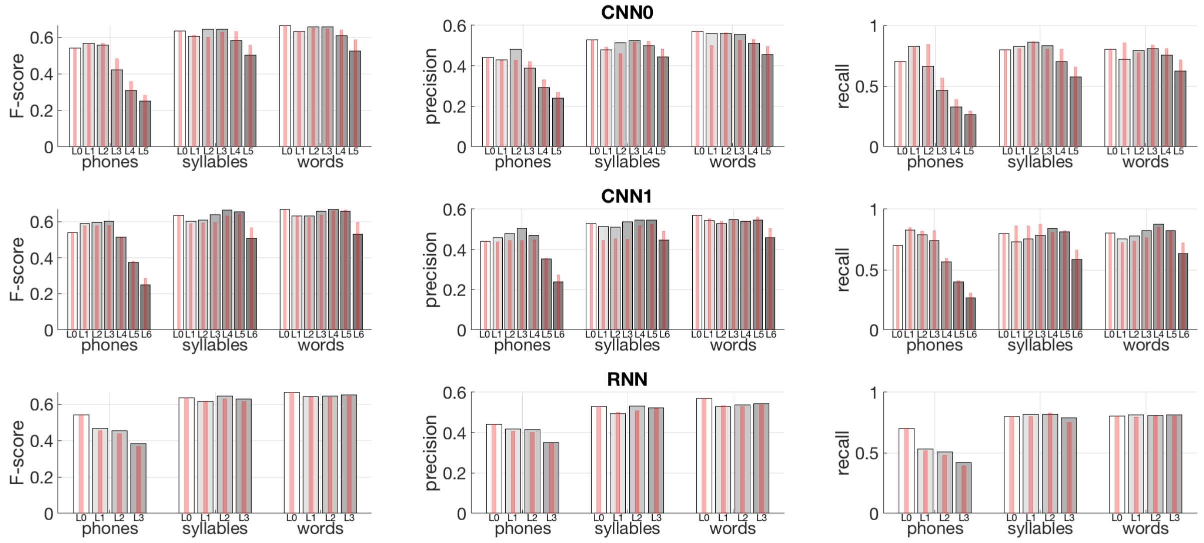


Figure 10. ***Segmentation results for models trained on Places and tested on Brent data.***

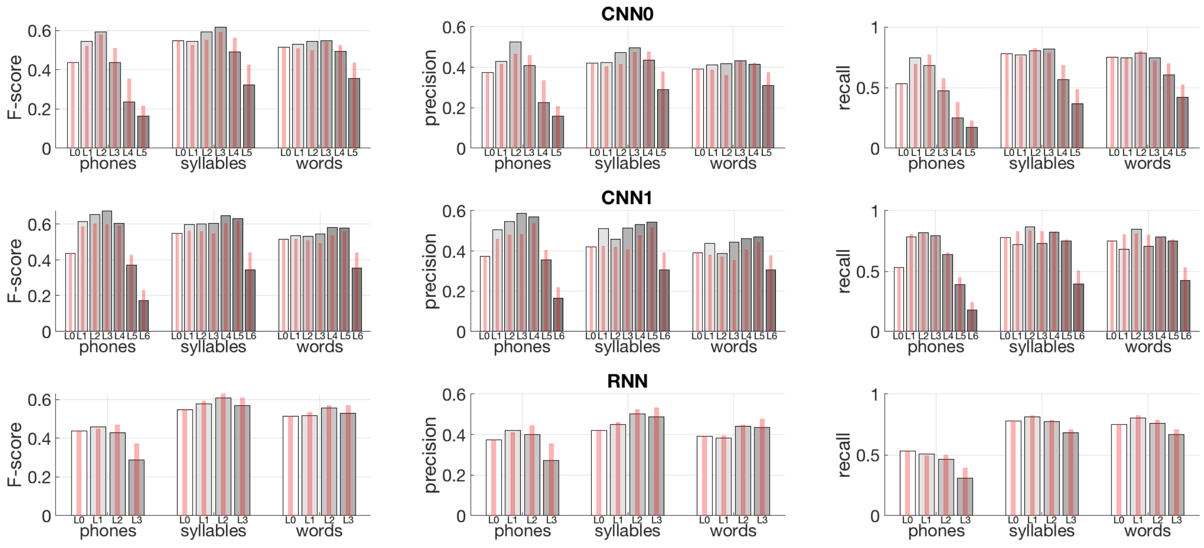


Figure 11. ***Segmentation results for models trained on Places and tested on COCO data.***

gain in syllable and word segmentation for the regression approach, as compared to the other two methods.

Compared to earlier studies, the observed segmentation scores are far from the F-scores of 0.72–0.80 reported for modern unsupervised phone segmentation algorithms (see, e.g., Hoang & Wang, 2015, for an overview), typically tested on TIMIT corpus (Garofolo et al., 1993). In addition, relatively low precision of the segmentation scores means that there is a substantial amount of oversegmentation involved in the process. In other words, the network dynamics have some correspondence to linguistic unit boundaries in time, but the segmentation behavior is far from being perfect. The observed performance is also somewhat worse than that reported by Harwath and Glass (2019) despite using the same model architecture and training protocol (CNN0). The reason for this difference is unclear, but may have to do with the different selection of the test data, as Harwath and Glass used read speech from TIMIT to test their model trained on Places. In general, it appears that temporal dynamics of the models are informative of phonetic, and to a smaller degree, syllabic and word boundaries. However, the effects of training are much less pronounced than with the selectivity analyses. Again, the main pattern of results does not seem to depend on whether the data are real or synthetic speech in nature.

Discussion

This work set out to investigate whether cross-modal and cross-situational learning can give rise to emergent latent linguistic structure, as predicted by LLH. After formulating the LLH and reviewing the findings from the existing literature, we investigated the idea systematically using machine learning models relying on statistical dependencies between images and their spoken descriptions. This can be viewed as simulation of cross-situational learning (e.g. Smith & Yu, 2008) with a high degree of referential ambiguity, as the models had to automatically discover which aspects of the utterances (in time and frequency) were related to what kind of visual referents (in terms of visual features and spatial positions). We compared three distinct speech encoder networks in the task to understand how speech encoder architecture impacts the audiovisual mapping performance and the manner that the networks encode linguistic information. In addition, we compared learning from both synthetic and real speech. As we analyzed the models, we observed that all networks exhibited similar capabilities in learning the semantic structure between spoken language and image data. In addition, all networks showed clear signs of linguistic organization in terms of all three unit types of analysis, namely phones, syllables, and words.

In terms of our more detailed linguistic analyses, the present findings largely align with the earlier literature on investigating linguistic units in VGS models (e.g., Chrupała et al., 2017; Alishahi et al., 2017; Havard et al., 2019a, 2019b; Merkx et al., 2019). However, the present study is the first one to show that broadly similar learning takes place in different model architectures (convolutional and recurrent) and on both synthetic and real speech. In addition, for the first time, we analyzed the emergence of phonetic, syllabic and lexical representations with a shared set of metrics, and probing for both linear and non-linear separability of the representations in terms of underlying linguistic units in the input. The analysis revealed that all three levels of representation exist in the trained models in parallel. Also, while phone-level information is generally accessible already from the earlier layers, phonetic information also co-exists in the layers that also encode syllabic or lexical information, but the information requires non-linear decoding. In addition, the results clearly demonstrate that the linguistic information in these type of models becomes encoded as distributed representations, whereas informativeness of individual network nodes with respect to linguistic unit types is very limited. Temporal dynamics of the models also seem to carry information related to linguistic units in the input speech, especially phones and syllables, although it appears that only a limited proportion of this is actually driven by audiovisual learning in the models. However, this is not surprising, as many of the existing unsupervised phone and syllable segmentation algorithms can already perform relatively well by only analyzing the original signal-level acoustic changes (e.g., Hoang & Wang, 2005, and Räsänen et al., 2018, and references therein).

Comparing our selectivity results to the earlier work, Chrupała et al. (2017) and Merkx et al. (2019) found that some model layers specialized for lexical processing whereas some others encoded such information to a lesser degree. According to Chrupała et al. (2017), the accuracy for predicting presence of individual words in speech increases towards the deeper layers of the network but decreases slightly for the last layers of the model. Our results for layer selectivity in the RNN model generally replicate those findings, and similar behavior can be observed for the CNN models as well. Regarding phonemic representations, Alishahi et al. (2017) and Drexler and Glass (2017) found that phone-like information was most evident in the early-to-intermediate layers of their networks, and where deeper layers showed slightly decreasing phonemic specificity. In our results with RNN model (e.g., Figs. 6 and 7), phones are also the most prominent in the first recurrent layer or in the convolutional layer preceding the recurrent layers of our RNN model, when compared to the penultimate recurrent layer (L3 in our RNN model; cf. Fig 2). Moreover, as our results illustrate, a similar general pattern of initially increasing and then suddenly decreasing linguistic unit selectivity is observed for all the model architectures and for all linguistic units. While much of the earlier work has been carried out using RNN-based VGS models, our results also indicate that the findings also apply to CNN-based architectures with different temporal characteristics, demonstrating the robustness of the phenomenon.

In terms of data characteristics, the present study also shows that the selectivity analyses conducted on synthetic speech are qualitatively similar to those on real speech.

This also strengthens the findings from the earlier visual grounding studies that have used synthetic speech (e.g., Alishahi et al., 2017). While not all synthetic speech is equal in terms of naturalness and acoustic variability, the finding also suggests that modern high-quality speech synthesis could be used for speech data creation in other computational modeling studies when lacking suitable real-speech corpora. This is especially relevant for studies where input to the learner needs to be dynamically adjusted depending on the learner behavior, such as simulating learning in infant-caregiver interaction using computational agents (e.g., Asada, 2016) and references therein). However, the potential limitations of synthetic speech should still be carefully and separately considered for each study.

In summary, together with the earlier computational findings reviewed, the present study demonstrates that speech comprehension, as defined in terms of capacity to associate spoken language with its referential meaning, does not necessitate a priori parsing of the speech input into distinct units such as phones or words. Instead, a flexible statistical learning machine focusing on modeling the dependencies between different perceptual channels is sufficient for capturing rudimentary semantics of speech, at least in the present type of simplified audiovisual learning scenarios. In addition, when implemented as a neural network with several hidden layers, these hidden layers start to reflect selectivity towards different types of linguistic units that the input speech consist of. This is in line with earlier findings using neural network models using supervised training (Nagamine et al., 2015; see also Magnuson et al., 2020) or simplified visual input (Räsänen and Khorrami, 2019). Here we show that similar emergence of units can be observed in learning conditions analogous to cross-situational learning. This lends initial support for the LLH: the idea that infant language representation learning may be driven (or at least supported) by learning processes that do not directly aim at learning such representations, but where the linguistic representations become acquired as a byproduct of multimodal sensing and interaction with the environment.

Note that the idea of initially general (non-linguistic) perceptual processing and gradual phonological development enabled by concurrently developing lexicon is in line with PRIMIR theory (Werker & Curtin, 2005). However, PRIMIR also assumes that acoustic word-forms are segmented before being associated with their meanings. The present work together with the earlier reviewed studies (e.g., Alishahi et al., 2017; Chrupała et al., 2017; Havard et al., 2019a; Harwath and Glass, 2019; Merkx et al., 2019; Scholten et al., 2020) demonstrates that explicit segmentation into acoustic word-forms before linking them to their meanings is simply not needed. In contrast, both sub-lexical and lexical representations can gradually emerge from the interaction of rich multimodal experiences available to the learner, when the learner is simply optimizing the multimodal predictive value of the auditory input. This is also in line with computational models demonstrating that there are synergies in learning multiple levels of linguistic structure simultaneously (e.g., Feldman et al., 2013) or in word segmentation and meaning acquisition (Johnson et al., 2010).

We would like to emphasize that the present type of VGS models do not aim at modeling neurophysiology of speech perception, and exploring such a connection is greatly beyond the scope of the present study. In contrast, our present aim has been to investigate LLH at the level of computational principles. However, despite the conceptual gap between artificial and real neurons, ANNs have been successfully applied to modeling of cortical organization in case of visual (Yamins & DiCarlo, 2016) and auditory (Kell et al., 2018) processing. Therefore linking model data from VGS models to neurophysiological data could be attempted in the future, given access to suitable human data. On the other hand, it would be an interesting avenue to explore VGS model architectures further by taking into account known characteristics of the auditory and associative areas in the brain.

Limitations of the Present Study

From the point of view of human language learning, one of the main limitations of the present study is the data used in our experiments. While language learning children observe the world from their own visual perspective and predominantly hear spontaneous speech in interaction with their caregivers (e.g., Yurovsky et al., 2013), our model training datasets consisted of photographs and their verbal descriptions. This means that both the visual and auditory experiences differ from those of a child, and also the relationship between the two modalities is much more systematic than what would be expected from situated caregiver speech. While caregiver speech is not random with respect to otherwise observable environment, caregivers do not tend to narrate everything that the child is observing. On the other hand, factors such as joint attention, prosodic cues of child-directed speech (CDS), skewed statistics of the visual experiences (e.g., Clerkin et al., 2017), and gradual increase in the complexity of the speech input may help infants to resolve audiovisual referential ambiguity, whereas the present VGS models do not receive any "highlighting" of relevant targets in the visual or auditory domains. As for audio quality, the synthesized speech in COCO dataset necessarily has less acoustic variability than authentic caregiver speech, making the speech parsing problem easier. However, the experiments with Places corpus inevitably show that our primary findings also apply to learning from real speech, and similar patterns of linguistic unit emergence (albeit with smaller effects) can be seen when the models are tested on naturalistic IDS speech.

In terms of mere input quantity, our training set totaled to approximately 400 h

(COCO) or 895 h (Places) of descriptive speech paired with visual scenes. In comparison, an average infant hears approx. 3 hours of CDS per day (Bunce et al., in preparation). The total amount of speech heard by the first birthday would then correspond to approx. 1000 h of CDS, i.e., by time when the child starts to comprehend some tens of words (CDI data from Wordbank; Fenson et al., 2007; Frank et al., 2017). In this context, at least the approximate scale of magnitude between toddler language input and our model input does not seem totally implausible, assuming that some tens of percentages of CDS input would relate to situations with opportunities for visual grounding. In addition, the present models can also solve the audiovisual mapping problem with much less data with reasonable performance (we have also tested learning using only one caption per COCO image, corresponding to a total of 80 h of speech; not reported separately). However, for the reasons listed above (and many others, such as developmental factors), detailed comparison between infant input and our study is not feasible. It is also not meaningful, as the present hypothesis is not that language learning would *only* take place through audiovisual learning. In contrast, the main goal has been to investigate the degree that referentially-driven multimodal learning can, *in principle,* explain aspects of early language organization.

As another central limitation, our present data consisted of English speech only. While this necessarily limits the extent that conclusions can be drawn cross-linguistically, the use of English also limits the capability to disentangle syllable- and word-level representations from each other. This is since a substantial proportion of the English word tokens consists of monosyllabic words (Greenberg, 1999), and this was also the case for our present data.

Finally, our viewpoint to the structure learned by the networks is necessarily limited, even though we combined a number of measures in our experiments. For instance, we did not systematically compare the networks in behavioral experimental paradigms such as gating experiments (see Havard et al., 2019b), nor investigated the confusions among different linguistic unit types. In contrast, we focused on understanding the dynamics of the internal representations from the point of view of hierarchy of linguistic units of different granularity.

In the future work, it would be important to test the audiovisual models with real infant language and visual input. Some baby steps to this direction already exist (Räsänen & Khorrami, 2019), but systematic investigation at the scale of real infant language experiences would be ideal to understand the role of visual[4] experience in early organization of language. Ideally, datasets from several different languages

[4] In the general case, this concerns parsing of speech in the context of sensory information from auditory, visual, somatosensory, and olfactory channels. In addition, motor activity and internal representations of emotions and interoception of basic bodily functions should be included as potential factors in speech grounding.

would be also utilized and compared. In addition, comparison and combination of the VGS models with purely auditory predictive models (e.g., van den Oord et al., 2018) should be conducted to understand the relative roles of different perceptual modalities in early language learning.

Conclusions

One of the puzzles in the study of child language acquisition research is the question how infants learn to parse the noisy and highly variable acoustic speech input into a meaningful and structured interpretation of the spoken message, and then to gradually use the language in a compositional and generative manner. Several potential mechanisms have been proposed to solve problems such as phonemic categorization, word segmentation, and word meaning acquisition, but the overall picture of how these bits and pieces fit together remains unclear. It is especially unclear what would be the ecological or functional pressure for the baby brain to solve a series of proximal language processing sub-problems before being able to utilize the benefits of speech perception to comprehend the world around them.

The latent language hypothesis investigated in this paper is aimed to shed some light on this puzzle by proposing that linguistic knowledge could emerge from predictive optimization across sensorimotor modalities and across time. By doing so, separate solutions to the several intermediate speech parsing problems would not be necessary. The reviewed studies and the present findings demonstrate that the audiovisual aspect of LLH is, in principle, a potential mechanism for assisting in language representation learning. However, the link to behavioral data is currently limited, and therefore nothing conclusive can be said on human learning based on the models alone. Yet, the VGS models show that the role of multisensory input in the context of language learning should not be underestimated, and that access to rich multimodal experiences concurrent to speech input have the potential to assist the learners already in the first stages of language learning. In future work, it would therefore be important to better understand the role of multimodality but also purely auditory predictive processing in early language acquisition.

References

Alishahi, A., Barking, M., & Chrupała, G. (2017). Encoding of phonology in a recurrent neural model of grounded speech. *Proc. 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pp. 368–378.

Asada, M. (2016). Modeling early vocal development through infant-caregiver interaction: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 8, 128–138.

Azuh, E., Harwath, D., & Glass, J. (2019). Towards bilingual lexicon discovery from visually grounded speech audio. *Proc. 20th Annual Conference of the International Speech Communication Association (Interspeech 2019),* pp. 276–280.

Baayen, H., Shaoul, C., Willits, J., & Ramscar, M. (2015). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience,* 31, 106–128.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proc. 3rd International Conference on Learning Representations (ICLR-2015).*

Ballard, D., & Yu, C. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions,* 1, 57–80.

Bar, M. (Ed.). (2011). *Predictions in the brain: Using our past to generate a future*. Oxford: Oxford University Press.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics,* 7, 49–72.

Bergelson, E., & Aslin, R. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences,* 114, 12916–12921.

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

Bunce, J., Casillas, M., Doelle, E.-A., & Soderstrom, M. (in preparation). A cross-cultural estimation of child directed speech across development.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint: http://arxiv.org/abs/1504.00325

Choi, D., Batterink, L. J., Black, A. K., Paller, K. A., & Werker, J. F. (2020). Preverbal infants discover statistical word patterns at similar rates as adults: Evidence from neural entrainment. *Psychological Science,* 31, 1161–1173.

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *Proc. 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pp. 613–622.

Chrupała, G., Higy, B., & Alishahi, A. (2020). Analyzing analytical methods: The case of phonology in neural models of spoken language. *Proc. 58th Annual Meeting of the Association for Computational Linguistics,* pp. 4146–4156.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences,* 36, 181–204.

Clerkin, E., Hart, E., Rehg, J., Yu, C., & Smith, L. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 372, 20160055.

Cole, J., Mi, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology,* 1, 425–452.

Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance,* 14, 113–121.

Deyne, S. D., Navarro, D., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science,* 45, e12922.

Drexler, J., & Glass, J. (2017). Analysis of audio-visual features for unsupervised speech recognition. *Proc. International Workshop on Grounding Language Understanding (GLU 2017),* pp. 57–61.

Driesen, J., & Van hamme, H. (2011). Modeling vocabulary acquisition, adaptation and generalization in infants using adaptive Bayesian PLSA. *Neurocomputing,* 74, 1874–1882.

Feldman, N., Griffiths, T., Goldwater, S., & Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review,* 120, 751–778.

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.

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

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience,* 11, 127–138.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Proc. 27th Annual Conference on Neural Information Processing Systems (NIPS 2013),* pp. 2121–2129.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus.* CD-ROM.

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. *Proc. 34th International Conference on Machine Learning (ICML 2017*), pp. 1243–1252.

Gentner, D. (1982). *Why nouns are learned before verbs: Linguistic relativity versus natural partitioning.* Center for the Study of Reading, Technical Report no. 257.

Greenberg, S. (1999). Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation. *Speech Communication,* 29, 159–176.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics,* 28, 267–283.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science,* 14, 1006–1033.

Harwath, D., & Glass, J. R. (2017). Learning word-like units from joint audio-visual analysis. *Proc. 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017),* pp. 506–517.

Harwath, D., & Glass, J. R. (2019). Towards visually grounded sub-word speech unit discovery. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019),* pp. 3017–3021.

Harwath, D., Hsu, W.-N., & Glass, J. (2020). Learning hierarchical discrete linguistic units from visually-grounded speech. *Proc. International Conference on Learning Representations (ICLR-2020).*

Harwath, D., Recasens, A., Surıs, D., Chuang, G., Torralba, A., & Glass, J. R. (2018). Jointly discovering visual objects and spoken words from raw sensory input. *Proc. 15th European Conference on Computer Vision (ECCV 2018),* pp. 659–677.

Harwath, D. F., & Glass, J. R. (2015). Deep multimodal semantic embeddings for speech and images. *Proc. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2015),* pp. 237–244.

Harwath, D. F., Torralba, A., & Glass, J. R. (2016). Unsupervised learning of spoken language with visual context. *Proc. Annual Conference on Neural Information Processing Systems (NIPS 2016),* pp. 1858–1866.

Havard, W., Besacier, L., & Chevrot, J.-P. (2020). Catplayinginthesnow: Impact of prior segmentation on a model of visually grounded speech. *Proc. 24th Conference on Computational Natural Language Learning* (CoNLL 2020), pp. 291–301.

Havard, W., Besacier, L., & Rosec, O. (2017). SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set. arXiv pre-print: http://arxiv.org/abs/1707.08435

Havard, W. N., Chevrot, J.-P., & Besacier, L. (2019a). Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on English and Japanese. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019),* pp. 8618–8622.

Havard, W. N., Chevrot, J.-P., & Besacier, L. (2019b). Word recognition, competition, and activation in a model of visually grounded speech. *Proc. 23rd Conference on Computational Natural Language Learning (CoNLL 2019),* pp. 339–348.

Hoang, D.-T., & Wang, H.-C. (2015). Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America,* 137, 797–805.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research,* 47, 853–899.

Hsu, W.-N., Harwath, D., Song, C., & Glass, J. (2020). Text-free image-to-speech synthesis using learned segmental units. arXiv pre-print: https://arxiv.org/abs/2012.15454

Iwahashi, N. (2003). Language acquisition through a human–robot interface by combining speech, visual, and behavioral information. *Information Sciences,* 156, 109–121.

Johnson, M., Demuth, K., Frank, M., & Jones, B. (2010). Synergies in learning words and their referents. *Proc. Advances in Neural Information Processing Systems 23 (NIPS 2010).*

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20, 137–194.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language*, 45, 229–254.

Kakouros, S., Salminen, N., & Räsänen, O. (2018). Making predictable unpredictable with style — behavioral and electrophysiological evidence for the critical role of prosodic expectations in the perception of prominence in speech. *Neuropsychologia*, 109, 181–199.

Kamper, H., Settle, S., Shakhnarovich, G., & Livescu, K. (2017). Visually grounded learning of keyword prediction from untranscribed speech. *Proc. 18th Annual Conference of the International Speech Communication Association (Interspeech 2017)*, pp. 3677–3681.

Kamper, H., Shakhnarovich, G., & Livescu, K. (2019). Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 27, 89–98.

Karpathy, A., & Li, F.-F. (2015). Deep visual-semantic alignments for generating image descriptions. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pp. 3128–3137.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98, 603–644.49

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLOS One*. https://doi.org/10.1371/journal.pone.0036399

Kiegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, article no. 4. https://doi.org/10.3389/neuro.06.004.2008

Kuhl, P. K., Conboy, B. W., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363, 979–1000.

Kvale, K. (1993). *Segmentation and labeling of speech*. PhD dissertation, The Norwegian Institute of Technology.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review,* 104, 211–240.

Landsiedel, C., Edlund, J., Eyben, F., Neiberg, D., & Schuller, B. (2011). Syllabification of conversational speech using bidirectional long-short-term memory neural networks. *Proc. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011),* pp. 5256–5259.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Proc. European Conference on Computer Vision (ECVV 2014),* pp. 740–755.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive Science,* 44, e12823.

Mangin, O., Filliat, D., ten Bosch, L., & Oudeyer, P.-Y. (2015). MCA-NMF: Multimodal concept acquisition with non-negative matrix factorization. *PLOS One,* https://doi.org/DOI:10.1371/journal.pone.0140732

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, K. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology,* 38, 465–494.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition,* 82, B101–111.

Merkx, D., Frank, S. L., & Ernestus, M. (2019). Language learning using speech to image retrieval. *Proc. 20th Annual Conference of the International Speech Communication Association (Interspeech 2019),* pp. 1841–1845.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science,* 343, 1006–1010.

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences,* 32, 376–382.50

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proc. 1st International Conference on Learning Representations (ICLR 2013).*

Nagamine, T., Seltzer, M. L., & Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. *Proc. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015),* pp. 1912–1916.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models for word representation, or, the unreasonable effectiveness of counting words near other words. *Proc. 39th Annual Meeting of the Cognitive Science Society,* pp. 859–864.

Ohishi, Y., Kimura, A., Kawanishi, T., Kashino, K., Harwath, D., & Glass, J. (2020). Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020),* pp. 4352–4356.

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

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences,* 53, 58–74.

Räsänen, O., Altosaar, T., & Laine, U. (2008). Computational language acquisition by statistical bottom-up processing. *Proc. 9th Annual Conference of the International Speech Communication Association (Interspeech 2008),* pp. 1980–1983.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review,* 122, 792–829.

Räsänen, O., Altosaar, T., & Laine, U. K. (2009). An improved speech segmentation quality measure: The R-value. *Proc. 10th Annual Conference of the International Speech Communication Association (Interspeech 2009),* pp. 1851–1854.

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

Räsänen, O., & Khorrami, K. (2019). A computational model of early language acquisition from audiovisual experiences of young infants. *Proc. 20th Annual Conference of the International Speech Communication Association (Interspeech 2019),* 3594–3598.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, 4512–4525.

Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., Daxenberger, J., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proc. 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pp. 3982–3992.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., & Li, F.-F. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211–252.

Rytting, A., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, 37, 512–543.

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

Salvi, G., Montesano, L., Bernadino, A., & Santos-Victor, J. (2012). Language bootstrapping: Learning word meanings from perception-action association. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 42, 660–671.

Schatz, T., Peddinti, V., Bach, F. R., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proc. 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pp. 1781–1785.

Scholten, S., Merkx, D., & Scharenborg, O. (2020). Learning to recognise words using visually grounded speech. arXiv pre-print: https://arxiv.org/abs/2006.00512

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proc. 3rd International Conference on Learning Representations (ICLR 2015)*.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558–1568.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics,* 2, 207–218.

Synnaeve, G., Versteegh, M., & Dupoux, E. (2014). Learning words from images and speech. *Proc. 28th Conference on Neural Information Processing Systems (NIPS): Workshop on Learning Semantics.*

ten Bosch, L., Van hamme, H., Boves, L., & Moore, R. K. (2008). A computational model of language acquisition: The emergence of words. *Fundamenta Informaticae,* 90, 229–249.

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

Trueswell, J., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology,* 66, 126–156.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv pre-print: http://arxiv.org/abs/1807.03748

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science,* 167, 392–393.

Watson, D., Arnold, J., & Tanenhaus, M. (2008). Tic tac toe: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, 106, 1548–1557.

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

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49–63.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience,* 19, 356–365.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science,* 29, 961–1005.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, 37, 891–921.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. *Proc. Annual Conference on Neural Information Processing Systems (NIPS 2014)*, pp. 487–495.

Data, code and materials availability statement

Program code and intermediate data files:

Python and MATLAB scripts for model definition, training, retrieval evaluation, and linguistic analyses, and result plotting are available at https://github.com/SPEECHCOG/VGS_XSL. Trained TensorFlow models, model hidden layer activations ("model outputs") used in the linguistic analyses, and the corresponding data annotations are available on Zenodo at https://doi.org/10.5281/zenodo.4564283.

Datasets used in the study:

Datasets used in the study are publicly available for download from their original sources after registration to the respective sites:

Brent-Siskind corpus: https://childes.talkbank.org/access/Eng-NA/Brent.html
Places audio captions: https://groups.csail.mit.edu/sls/downloads/placesaudio/downloads.cgi
Places205 images: http://places.csail.mit.edu/downloadData.html
SPEECH-COCO audio captions: https://zenodo.org/record/4282267
MSCOCO images https://cocodataset.org/#download

The derived version of "Large-Brent" with utterance-level waveforms with their phone, syllable, and word-level transcripts (based on forced-alignment from Rytting et al., 2010, and syllabification in Räsänen et al., 2018) is available from the second author upon request. The data cannot be shared publicly, as it would require redistribution of modified (split + pruned) Brent-Siskind corpus audio files. Model outputs and the corresponding annotations for the Large-Brent data are available in the Zenodo-hosted datafile mentioned above.

Authorship and Contributorship Statement

K.K. planned the study together with the second author. She was also responsible for model development and implemented all program code related to model training, retrieval evaluation, and initial linguistic selectivity analysis scripts. K.K. also prepared the data, ran all experiments except for the final linguistic analyses, and produced the corresponding result figures and tables. She wrote the first draft of the article together with the second author and participated to editing of the initial and revised manuscript versions.

O.R. was responsible for the initial idea of the study, for theoretical framing of the work, and generally supervising the work. He planned the work jointly with the first author and participated to methodological development. He also implemented the final linguistic analysis scripts (selectivity and temporal) and conducted and reported the corresponding experiments. O.R. wrote some sections of the first manuscript draft jointly with the first author and was largely responsible for editing the initial and revised versions of the manuscript.

Acknowledgements

This research was funded by Academy of Finland projects no. 314602 and 320053. The authors would like to thank Herman Kamper for his help with VGS model training and for the highly useful discussions on the topic. The authors would also like to thank Grzegorz Chrupała for providing details on his earlier work.

Code used for model training and analysis, trained models, and hidden layer activation data are available for download at https://github.com/SPEECHCOG/VGS_XSL.

Appendix A

Selectivity analyses (Fig. A.1) and temporal segmentation results (Fig. A.2) for models trained on COCO and tested on Brent corpus.

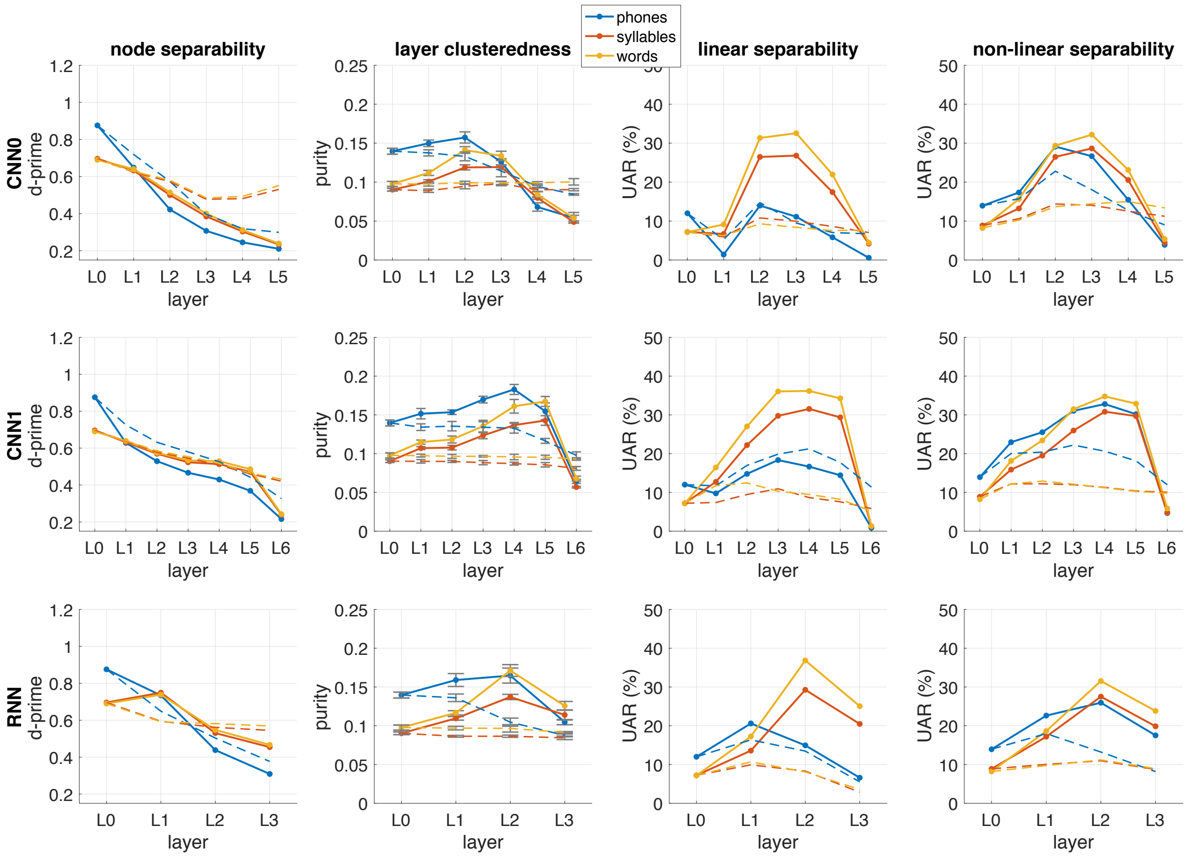


Figure A.1. ***Analysis results for models trained on COCO corpus and tested on Brent corpus. Each panel row corresponds to one of the models, CNN0, CNN1 or RNN, whereas columns correspond to the four studied selectivity metrics. Blue lines stand for phones, red for syllables, and yellow for words. Solid lines correspond to trained models and dashed lines for the corresponding baseline models before the training. Error bars for clusteredness represent SDs across different runs of the k-means analysis.***

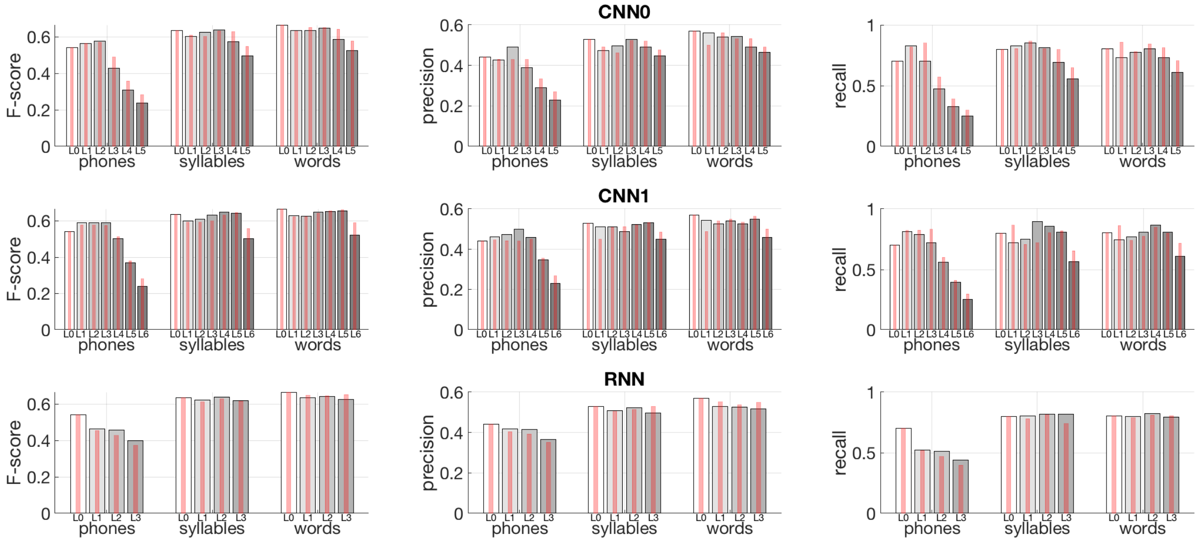


Figure A.2. ***Segmentation results from models trained COCO and tested on Brent corpus when using linear regression from activation magnitudes as the signal representation. Bars in each subplot represent layer-specific performance metrics for segmenting phones, syllables, and words, respectively. Different layers from input (L0) to last network layer are shown with different shade bars from left to right. Results are shown for F-score (left panels), precision (middle panels), and recall (right panels), and for the three tested models: CNN0 (top), CNN1 (middle), and RNN (bottom). Thin red lines denote baseline performance with untrained models.***

Appendix B

Results for segmentation analyses using L2-norm or entropy of layer activations as the signal representation (instead of using the linear regression scores in the main results).

L2-norm measures: Fig. B.1: training and testing on COCO. Fig. B.2: training on Places and testing on Brent. Fig. B.3: training on Places and testing on COCO. Fig. B.4: training on COCO and testing on Brent.

Entropy-based measures: Fig. B.5: training and testing on COCO. Fig. B.6: training on Places and testing on Brent. Fig. B.7: training on Places and testing on COCO. Fig. B.8: training on COCO and testing on Brent.

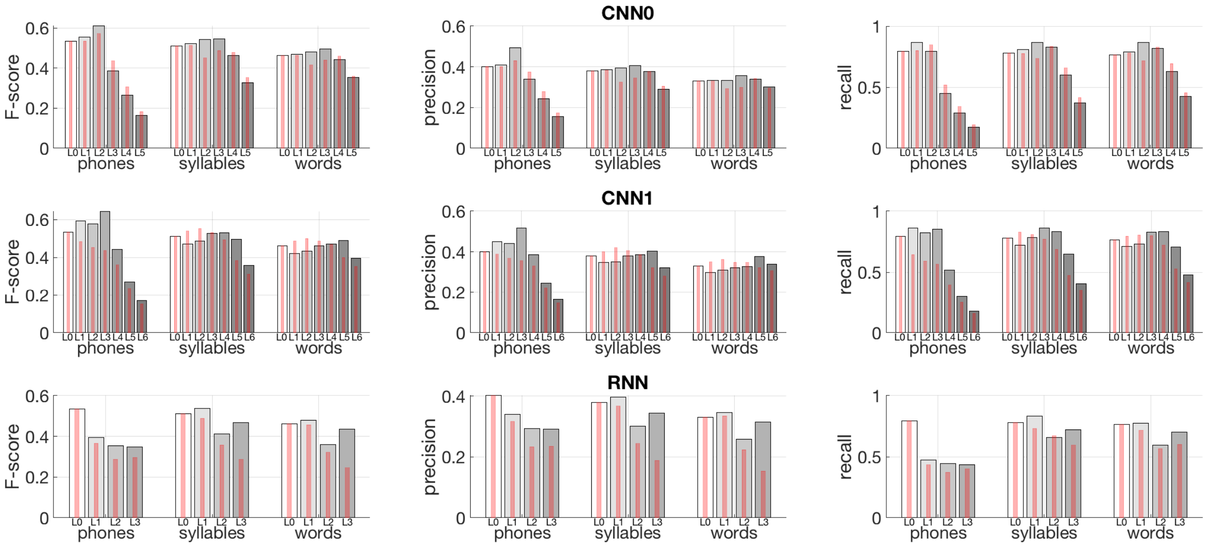


Figure B.1. ***Segmentation results from models trained and tested on COCO corpus when using L2-norm of activation magnitudes as the signal representation. Bars in each subplot represent layer-specific performance metrics for segmenting phones, syllables, and words, respectively. Different layers from input (L0) to last network layer are shown with different shade bars from left to right. Results are shown for F-score (left panels), precision (middle panels), and recall (right panels), and for the three tested models: CNN0 (top), CNN1 (middle), and RNN (bottom). Thin red lines denote baseline performance with untrained models.***

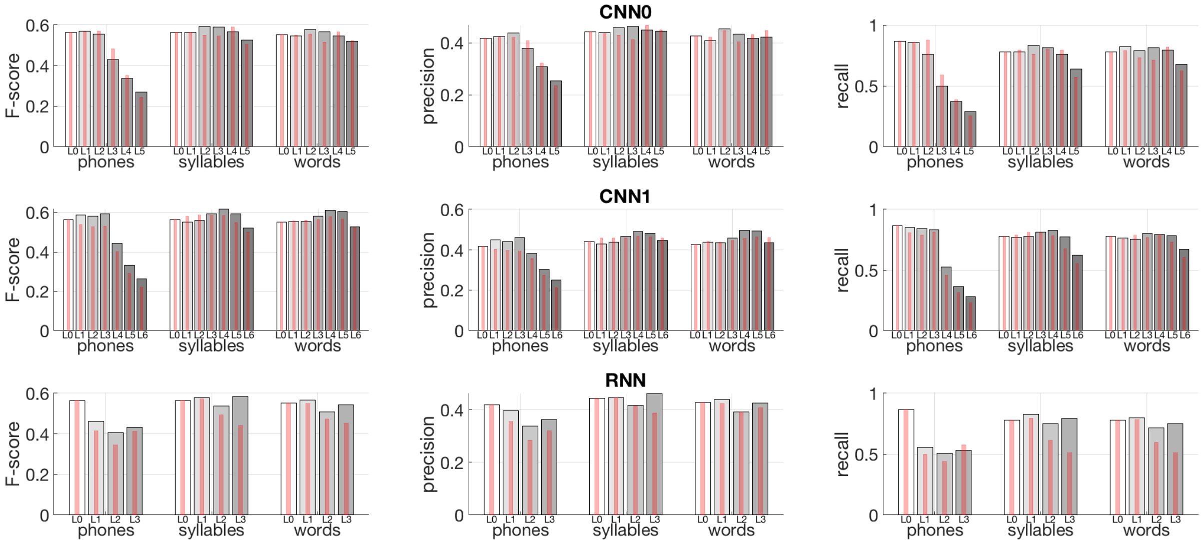


Figure B.2. ***Segmentation results for models trained on Places and tested on Brent data when using L2-norm of activation magnitudes as the signal representation.***

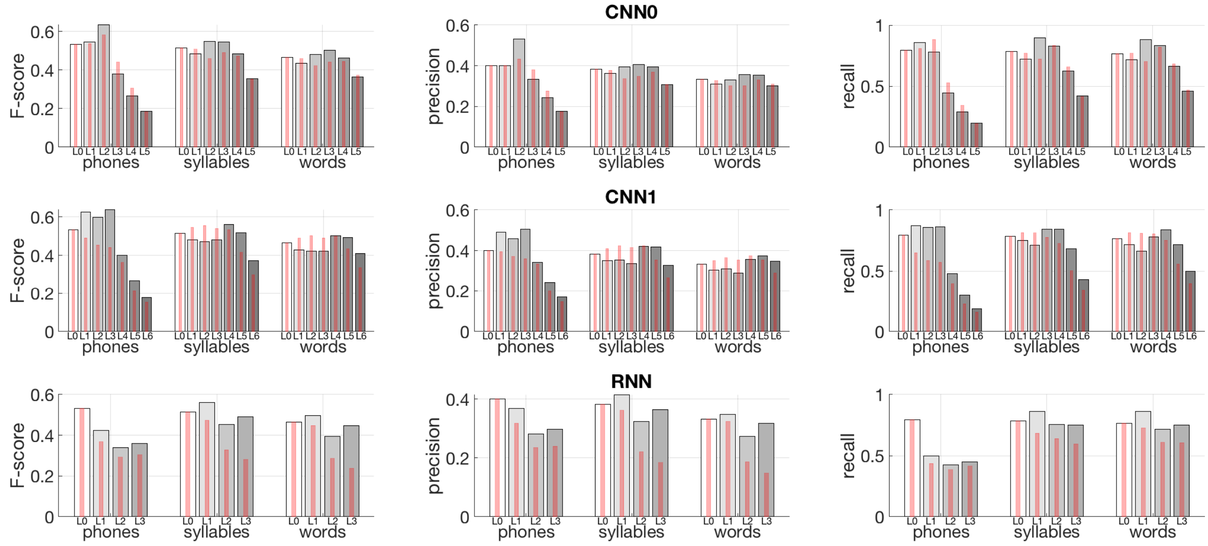


Figure B.3. ***Segmentation results for models trained on Places and tested on COCO data when using L2-norm of activation magnitudes as the signal representation.***

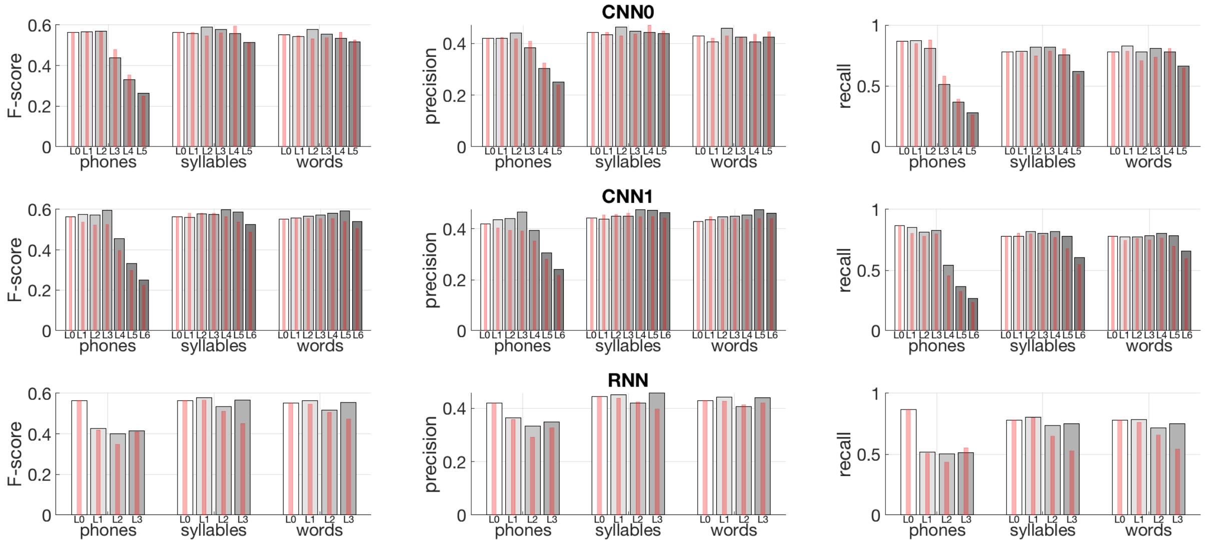


Figure B.4. ***Segmentation results for models trained COCO and tested on Brent data when using L2-norm of activation magnitudes as the signal representation.***

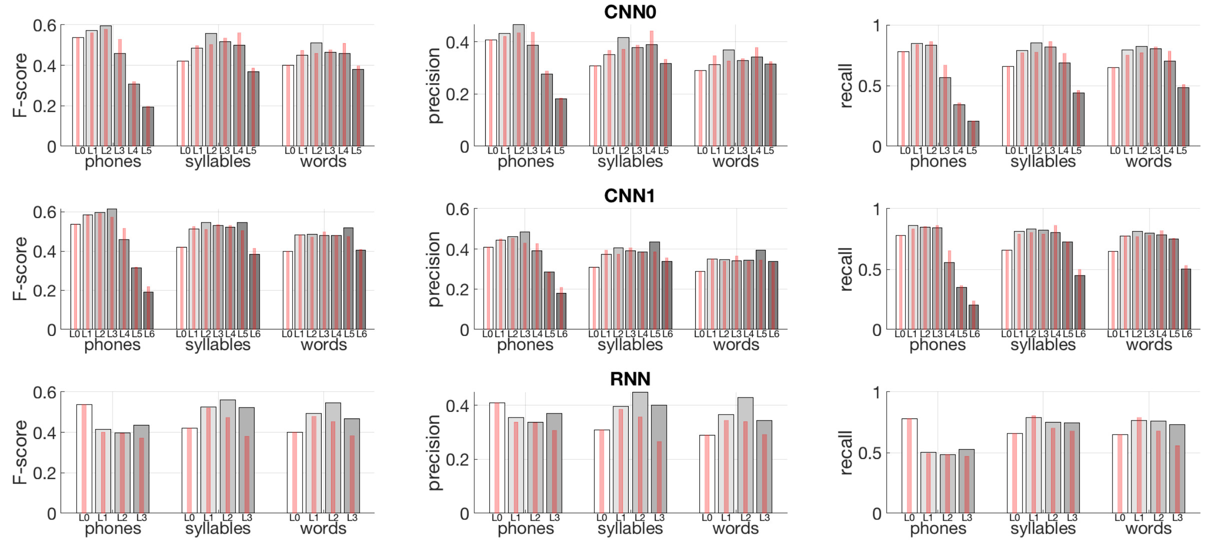


Figure B.5. ***Segmentation results from models trained and tested on COCO when using entropy of activation magnitudes as the signal representation. Bars in each subplot represent layer-specific performance metrics for segmenting phones, syllables, and words, respectively. Different layers from input (L0) to last network layer are shown with different shade bars from left to right. Results are shown for F-score (left panels), precision (middle panels), and recall (right panels), and for the three tested models: CNN0 (top), CNN1 (middle), and RNN (bottom). Thin red lines denote chance-level performance with randomized boundary locations.***

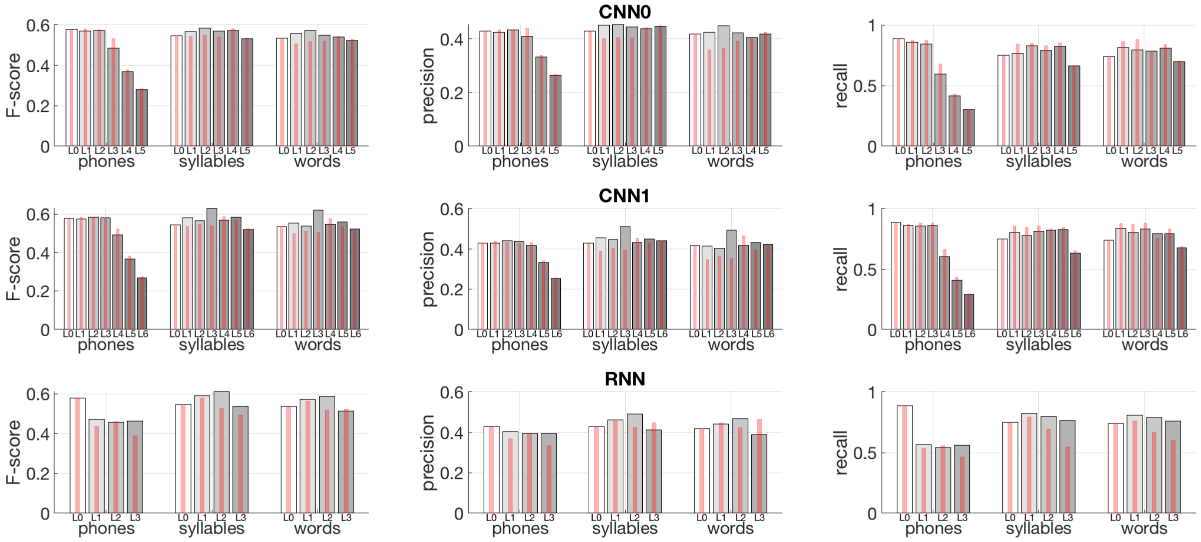


Figure B.6. ***Segmentation results from models trained on Places and tested on Brent when using entropy of activation magnitudes as the signal representation.***

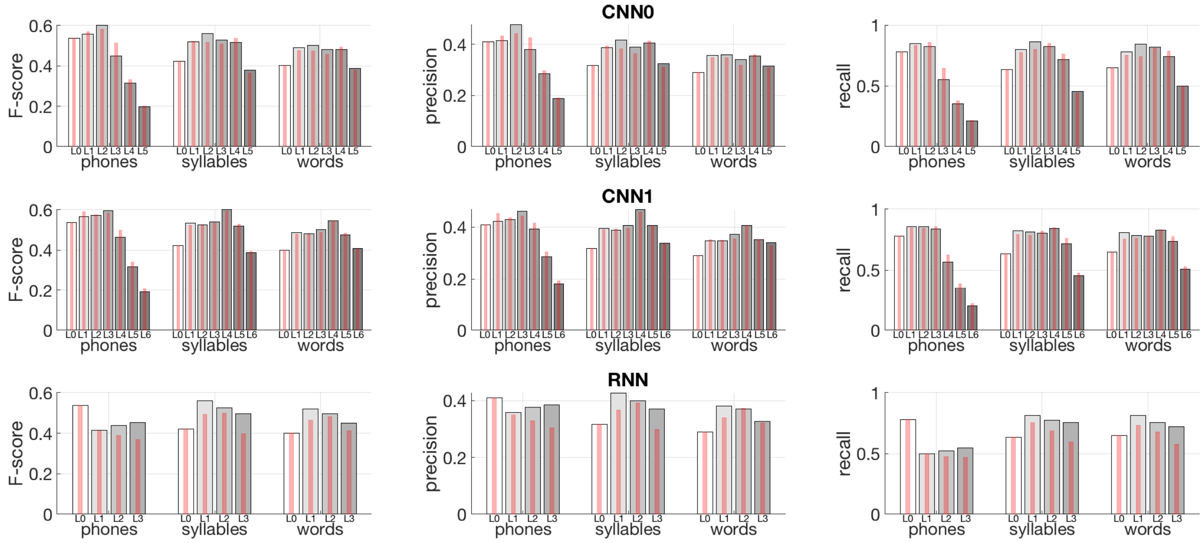


Figure B.7. ***Segmentation results from models trained on Places and tested on COCO when using entropy of activation magnitudes as the signal representation.***

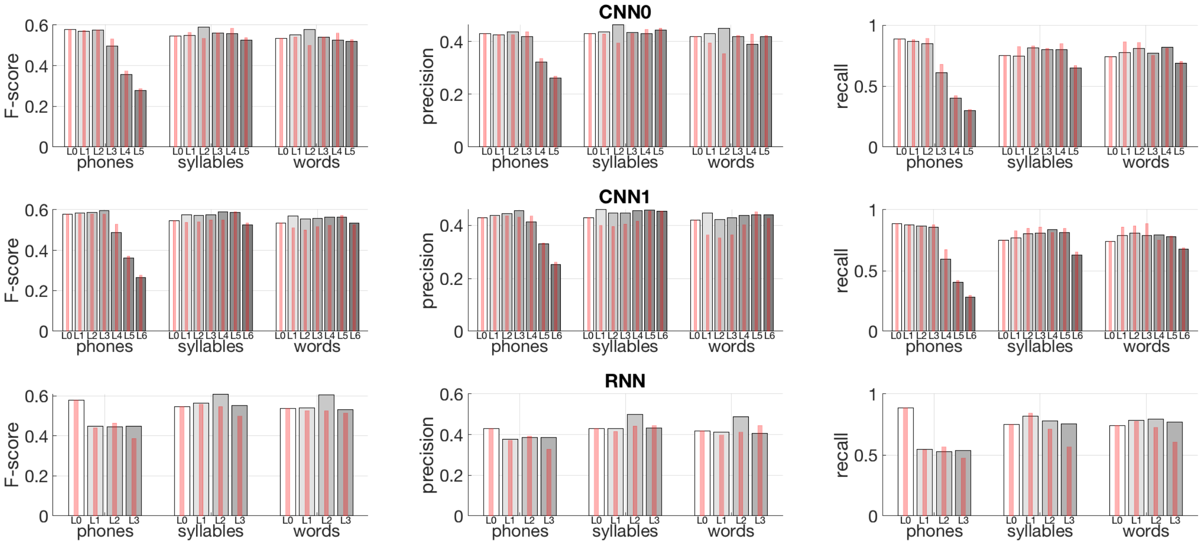


Figure B.8. ***Segmentation results from models trained on COCO and tested on Brent when using entropy of activation magnitudes as the signal representation.***

Appendix C

Wilcoxon rank-sum statistics for SRS and SBERT similarity score distributions from speech-to-speech search are shown in Table C.1.

Table C.1. ***Wilcoxon rank-sum statistic for speech-to-speech search for tested models, comparing the distributions consisting of the SRS scores for 5 nearest vs. 5 furthest utterances ("near vs. far") or 5 nearest vs. 5 random utterances ("near vs. random") for each query utterance (p < 0.001 for all). Left: SRS results using all content words in the utterances. Middle: SRS results excluding repeating words between query and search result utterances. Right: SBERT results for full captions.***

	SRS w. repeating words		**SRS wo. repeating words**		**SBERT**	
COCO (df = 24385)	near vs. far	near vs. rand	near vs. far	near vs. rand	near vs. far	near vs. rand
CNN0	180.41	170.52	143.14	127.91	186.21	178.65
CNN1	185.06	174.12	155.62	128.64	188.73	180.63
RNN	177.65	169.28	139.16	129.79	184.11	178.26
Places (df = 46210)	near vs. far	near vs. rand	near vs. far	near vs. rand	near vs. far	near vs. rand
CNN0	192.90	169.10	146.91	119.20	209.34	184.31
CNN1	208.77	182.91	151.18	119.26	226.44	198.53
RNN	186.41	167.63	145.35	125.72	201.31	184.14

License

Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study

Giulia Bovolenta
Emma Marsden
University of York, UK

Abstract: Prediction error is known to enhance priming effects for familiar syntactic structures; it also strengthens the formation of new declarative memories. Here, we investigate whether violating expectations may aid the acquisition of new abstract syntactic structures, too, by enhancing memory for individual instances which can then form the basis for abstraction. In a cross-situational artificial language learning paradigm, participants were exposed to novel syntactic structures in ways that either violated their expectations (Surprisal group) or that conformed to them (Control group). First, we established a potential expectation to hear feedback that simply repeated the same structure as that just experienced. We then manipulated feedback so that the Surprisal group unexpectedly heard passive structures in feedback following active sentences, while the Control group only heard passive structures following passive sentences. Delayed post-tests examined participants' structural knowledge both by means of structure test trials (focusing on the active / passive distinction, with both familiar and novel verbs), and by a grammaticality judgment task. The Surprisal group was significantly more accurate than the Control group on the structure test trials with novel verbs and on the grammaticality judgment task, suggesting participants had developed stronger abstract structural knowledge and were better at generalising it to novel instances. Tentative evidence suggested the Surprisal group was not significantly more likely to become aware of the functional distinction between the two structures.

Keywords: Second language acquisition; prediction error; cross-situational learning; artificial language learning; structural generalisation.

Corresponding author(s): Giulia Bovolenta, Department of Education, University of York, Heslington, York, YO10 5DD, UK. Email: giulia.bovolenta@york.ac.uk.

ORCID ID(s): Giulia Bovolenta: https://orcid.org/0000-0003-4139-6446;
Emma Marsden: https://orcid.org/0000-0003-4086-5765.

Citation: Bovolenta, G., & Marsden, E. (2021). Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study. *Language Development Research*, 1(1), p 193-243. DOI: 10.34842/c7t4-pz50

Introduction

Is it possible to 'surprise' a learner into acquiring a new structure in a foreign language? A growing body of literature suggests that unpredictable input favours language learning. On one hand, structural adaptation – an increased likelihood to use or expect the syntactic structures we are exposed to, persisting in the long term – is likely one of the mechanisms by which we tune into the patterns of our language (Peter & Rowland, 2019). There is evidence that prediction error drives adaptation to syntactic structure, both from computational modelling (Chang, Dell, & Bock, 2006) and empirical studies with both first language (L1) and second language (L2) speakers (Fazekas, Jessop, Pine, & Rowland, 2020; Montero-Melis & Jaeger, 2020). At the same time, evidence shows that violating expectations facilitates the formation of new individual declarative memories, too, including vocabulary learning (Greve, Cooper, Kaula, Anderson, & Henson, 2017; Stahl & Feigenson, 2017). We are now beginning to form a picture of the ways in which surprisal can aid learning with regards to different aspects of language. If a learner already has the relevant abstract syntactic representation, encountering the structure in a surprising context appears to strengthen that representation. Surprisal can also facilitate the acquisition of new declarative memories for lexical items, such as nouns or verbs, leading to stronger memory formation than non-surprising contexts. But what about the acquisition of *new, syntactic* representations among adult learners who have already established their L1 system? In this study, we address an unexplored gap in the literature, asking whether surprisal could also aid the development of new abstract structural representations, including acquisition of their specific form-meaning mappings, rather than just strengthening existing ones. Following a usage-based approach to language acquisition, we assume that structural knowledge emerges through abstraction from individual learned exemplars (N. C. Ellis, Römer, & O'Donnell, 2016). If expectation violation can aid memory for individual instances, then we hypothesise that it may also aid the acquisition of structural knowledge through abstraction from these individual instances.

We investigated this question in a controlled learning experiment using an artificial language (Yorwegian). Learners were first introduced to a default syntactic structure, the active construction, which they learned while they were also learning the vocabulary of the language. Then, once this structure had been learned and consolidated, participants were exposed on the second day to a (potentially) more complex alternative, the passive construction. This ordering (active then passive) and bias in the input (more active than passive) simulates, to some extent, the likely real-life learning experience of many learners, who would tend to encounter the passive construction less often in their learning due to its lower frequency, relative to the active construction. In this context, we manipulated the utterance containing the passive construction (in what we called a 'feedback' turn), so as to make it either unexpected (Surprisal group) or expected (Control group) relative to the pattern that had been established during training. Participants responded to sentences they heard (by selecting the matching picture) and received feedback on their responses, which consisted of a replay of (the meaning of) the initial sentence they had given their response to. In the first blocks, both groups received feedback using the structure that was always congruent with the structure in the initial sentence, i.e., participants heard the exact same sentence. However, in later trials, the Surprisal group occasionally experienced feedback containing a passive structure immediately following an active structure (though still describing the same picture and with the same meaning in terms of agents and patients), while the Control group always experienced feedback containing the structure that matched the one used in the preceding sentence. We hypothesised that participants in the Surprisal group would develop stronger representations for the passive sentences encountered in feedback, leading to improved learning of the passive syntactic structure itself[1]. In a secondary question, we also hypothesised that surprisal may aid the development of explicit knowledge, either by increasing attention and cognitive effort (Leow, 2015) or by generating stronger representations that would be more likely to emerge in conscious awareness (Cleeremans, 2011).

Background Literature

Structural Priming as a Learning Mechanism

When language users encounter a particular syntactic construction, they are often more likely to expect it again, or to use it in production, than they were before encountering it, a phenomenon known as *structural priming* (Arai, van Gompel, & Scheepers, 2007; Bock, 1986; Ferreira & Bock, 2006; Ledoux, Traxler, & Swaab, 2007). When the priming effect persists over time, it is known as *adaptation* (Kaan & Chun, 2018b). Adaptation to syntactic structure alternations (such as that between prepositional object and double object dative constructions in English) has been observed in L1 production (Jaeger & Snider, 2013; Kaschak, 2007; Kaschak & Borreggine, 2008; Kaschak, Kutta, & Jones, 2011; Kaschak, Loney, & Borreggine, 2006), and in L1 comprehension (Farmer, Fine, Yan, Cheimariou, & Jaeger, 2014; Fine & Jaeger, 2016; Fine, Jaeger, Farmer, & Qian, 2013; Kaan & Chun, 2018a). Adaptation effects have also frequently been observed in L2 speakers (Jackson & Ruf, 2017; Kaan & Chun, 2018a; McDonough & Trofimovich, 2015; Montero-Melis & Jaeger, 2020; Shin & Christianson, 2012; see Jackson, 2018 for a review). The magnitude of these effects tends to be greater for less frequent structures (known as *inverse probability effects*). This has been observed empirically in both the L1 and L2: Structures that have lower frequency in the input elicit greater priming effects (Hartsuiker, Kolk, & Huiskamp,

[1] In this sense, our manipulation is quite different from previous research on surprisal in language processing, as it manipulates expectations about the context in which the 'surprising' language was experienced, rather than the input per se. A reviewer pointed out that another potential way of framing our manipulation could perhaps be as a type of 'recast', which is an interactional and/or feedback (error correction) phenomenon, both in natural discourse (e.g., as a confirmatory turn or as a clarification/comprehension checking mechanism) and in language instruction (e.g., confirmatory to promote continued communication, or corrective to provide feedback on errors) (Goo & Mackey, 2013; Lyster & Saito, 2010; see, however, Foster (Foster, 1998; Foster & Ohta, 2005), who downplays the frequency of recasting in instructional situations). In our study, an incongruent (potentially conceptualised as 'corrective' or 'comprehension checking') recast could be more salient and/or lead to greater awareness relative to a congruent (potentially conceptualised as 'confirmatory' or 'interaction promoting') one, a possibility we raise in the discussion. However, one caveat to keep in mind is that recasts in L2 acquisition studies are normally in response to an utterance produced by the learner, whereas in our study the initial statement is *heard* by the learner, rather than produced. Therefore, if we think of our study in terms of recast, our design could perhaps simulate the cases where a learner hears the interaction and is working out the meaning, rather than actively participate in the interaction.

1999; Hartsuiker & Westenberg, 2000; Jaeger & Snider, 2013; Kaan & Chun, 2018a; Kaschak, 2007; Kaschak et al., 2006; Montero-Melis & Jaeger, 2020; Weber, Christiansen, Indefrey, & Hagoort, 2019).

In L2 acquisition, there is evidence of syntactic priming mechanisms operating from the earliest stages of learning. In Weber et al. (2019), participants were exposed to a novel artificial language in four sessions over the course of nine days. The language consisted of a lexis containing novel nouns and verbs arranged in four possible word orders: three transitive (VOS, OSV, SVO) and one intransitive (SV). In the first session, participants were pre-trained on the nouns. From the second session, participants read novel sentences aloud, which were accompanied by pictures depicting their meaning. Repetition of word order and verb was manipulated from one trial to the next to study priming effects, which were measured by read-aloud times. In the third and fourth session, priming was also assessed through a picture matching task after target trials, where participants had to pick the picture matching the sentence they had just read, out of two possible alternatives (the incorrect picture depicted the same event, but with Agent and Patient roles reversed). Priming effects in read-aloud times were observed from the earliest stages; however, there was no difference in magnitude of priming for infrequent vs. frequent structures (structure frequency was manipulated in the second session, where the frequent word order was twice as likely to occur as the other three). Priming effects were also observed in comprehension, with higher accuracy for repeated structures, but only if they were the frequent word order.

It has been suggested that structural priming is a case of implicit error-based learning (Bock, Dell, Chang, & Onishi, 2007; Bock & Griffin, 2000; Chang, Janciauskas, & Fitz, 2012). Computational modelling of priming data shows that this can be reproduced by a Recurrent Neural Network (RNN) model trained on next-word prediction. As the model encounters more sentences, it gradually improves by adjusting its predictions based on the discrepancy between predicted and actual input, or *prediction error* (Chang et al., 2006). This account is compatible with observed properties of priming and adaptation, such as inverse frequency effects: Low-frequency words would generate greater prediction error, causing a larger adjustment in the weights and therefore a larger learning effect (Chang et al., 2006). More recently, Fazekas et al. (2020) tested this account with an empirical study with both adults and children, and found that exposing participants to surprising dative sentences (using verbs rarely associated with the dative structure) made participants more likely to use the dative structure in a post-test.

Effect of Expectation Violation on New Memory Formation

Structural priming and adaptation phenomena affect representations that have *already* been acquired; what changes as a consequence of exposure, and is further increased by prediction error, is the strength of existing structural representations. However, evidence from a different strand of research, originating mainly in cognitive psychology, shows that prediction error can also enhance the formation of *new* individual memories; events or associations which violate our expectations are remembered better than those that conform with them (*one-shot declarative learning*). Novel associations are better remembered if they violate an established pattern

(Brod, Hasselhorn, & Bunge, 2018; Greve et al., 2017; Greve, Cooper, Tibon, & Henson, 2019), including translation word pairs (De Loof et al., 2018). Surprising feedback, too, is better remembered. Fazio & Marsh's (2009) participants answered general knowledge questions (rating their confidence in their answers) and then were shown the correct answer, which was displayed in either red or green letters. When feedback was unexpected (either following a high-confidence incorrect answer, or a low-confidence correct one) memory for the font colour in which it was displayed was better than for expected feedback. This suggests that surprising feedback can lead to a greater effort to encode it (known as the *surprise hypothesis),* resulting in better 'source memory' (defined as memory for the conditions in which the feedback is encoded, including everything that gets encoded besides the content of the feedback itself).

There is also direct evidence that the effect of violation expectation on novel memory formation can aid language acquisition: Stahl & Feigenson (2017) showed that violation of expectations promotes vocabulary learning in young children. In the study, 3- to 6-year-old children were exposed to novel events which were either entirely possible or which violated core properties of the objects involved (e.g., a cup vanishing and reappearing in a different location). They were then taught the verb corresponding to the action (Experiment 1) or the noun denoting one of the objects (Experiment 2), and were tested immediately on its meaning. Children were significantly more accurate in their responses for verbs and nouns that they had learned in surprising events than for those they had learned in expected events (on which they performed at chance level). The effect was limited to nouns and actions involved in the surprising event: If children were taught the name for an object that was present during the event but did not participate in it, there was no learning effect (Experiment 4). This suggests that violated expectation did not aid learning simply by increasing attention or arousal, but that it led children to revise their predictions about specific objects and events (Stahl & Feigenson, 2017).

The Present Study

From the literature surveyed, it is clear that unexpected input can lead to a strengthening of existing abstract structural representations, in the form of increased priming and adaptation. We also know that violated expectation enhance the formation of new declarative memories, including learning novel vocabulary items. What we do not know, and what is the of focus of this study, is whether surprisal may also favour the acquisition of new *abstract structural* representations. In usage-based accounts of language acquisition, the development of abstract, structural knowledge is assumed to proceed from learned exemplars in the first place (Bybee & Hopper, 2001; N. C. Ellis, 2002; N. C. Ellis et al., 2016). If expectation violation can aid memory for individual instances, then we hypothesise that it may also aid the acquisition of structural knowledge through abstraction from these individual instances.

For our study, we adapted a cross-situational learning paradigm (Smith & Yu, 2008; Yu & Smith, 2007) which has been successfully used in previous studies to investigate the acquisition of syntax in naturalistic settings (Monaghan, Ruiz, & Rebuschat, 2020; Rebuschat, Monaghan, & Schoetensack, 2021; Walker, Monaghan, Schoetensack, &

Rebuschat, 2020). In a cross-situational learning paradigm, participants are exposed to a novel language without any explicit instruction, but instead derive the meaning of novel words by attempting to interpret them across multiple situations. Participants are exposed to novel words or sentences and are required to select the correct interpretation from a range of options. While they are initially at chance in their answers, participants eventually converge on the correct meaning by keeping track of possible interpretations across different trials. Walker et al. (2020) used an artificial language composed of 16 novel words (8 nouns, 4 verbs, 2 adjectives) which could be arranged in either a subject-object-verb (SOV) or object-subject-verb (OSV) word order. Participants were trained and tested on the language over the course of two days, without any explicit instruction. In each learning trial, they heard a sentence in the novel language while two animations appeared on screen; their task was to select the one matching the sentence.

Accuracy in learning trials was above chance from the second block, and results from intermitting test blocks showed that participants succeeded in acquiring both the grammar and vocabulary. This makes it a highly suitable paradigm to investigate the acquisition of syntactic structure in a naturalistic way. To establish expectation and then induce surprisal, we added feedback to critical trials. This feedback always contained a passive structure, which was, at first, always consistent with the trial just heard. We then manipulated the feedback between groups to be either consistent or inconsistent with expectations that participants had established during their first blocks of feedback trials. That is, we assumed that participants would expect feedback turns to replay the sentence in the exact form they had just heard. To generate this expectation, we ensured that feedback was initially congruent for both groups, and only at a later stage did we introduce, for the Surprisal group only, incongruent trials: active sentences that were followed by a passive form, whilst the *same* picture was displayed as during the active sentence. Given that surprising feedback is thought to be better encoded, including its visual features (Fazio & Marsh, 2009), we expected the passive sentences in surprising feedback trials to lead to better learning, not only of the picture itself, but of the specific sentence – picture pairing, too, relative to the learning in the group that experienced the expected feedback trials.

It is also possible that surprising feedback may promote the development of explicit knowledge of the passive structure. While findings like those of Stahl & Feigenson (2017) suggest that the effect of expectation violation on learning is not driven simply by a *general* raising of attention, it seems likely that surprisal has an effect on attention, albeit only to the relevant features (see for instance Greve et al. (2017) on possible mechanisms underlying one-shot declarative learning). In the context of associative learning, it has been suggested that surprisal may increase the salience of a stimulus, which in turn drives learning (Cintrón-Valentín & Ellis, 2016; N. C. Ellis, 2016, 2017). Increased attention may also lead to greater awareness, that is, explicit knowledge of the form-meaning connections being learned. On the one hand, this may happen directly as a consequence of deeper engagement with the stimuli; for example, in L2 research, greater cognitive effort has been reported to correlate positively with the emergence of rule awareness (Cerezo, Caras, & Leow, 2016; Leow, 2015). On the other hand, surprisal may also have a more indirect effect on the emergence of explicit knowledge. According to the *radical plasticity* thesis

(Cleeremans, 2008, 2011), there is a continuum between implicit and explicit knowledge. On initial exposure, implicit knowledge develops, characterised by weak and low-quality representations in memory. As the quality and strength of representation increase with repeated exposure, the knowledge becomes increasingly available to consciousness, that is, becomes explicit. Therefore, if surprisal leads to stronger representations in memory, we may also expect it to lead to more explicit knowledge. More specifically, in our case, stronger representations of individual passive sentences may lead to greater awareness of the form-meaning connections involved (though we do not aim to tease apart these two accounts [greater cognitive effort versus radical plasticity] of how this may happen).

Research Questions and Predictions

Our primary research question (RQ1) was whether being exposed to surprising items in the passive would lead to overall better knowledge of the passive structure. This was assessed by performance accuracy on picture-matching comprehension tests, with both trained and novel lexicon (to assess generalisation to new instances), and a grammaticality judgment task. If expectation violation can aid structural learning, we would expect the Surprisal group (SG) to show better knowledge of the passive structure than the Control group (CG).

Specifically, with regards to comprehension (in picture-matching comprehension tests), we predicted the Surprisal group would perform better than the Control group in structure comprehension test blocks which were placed both at the end of Day 2, after the surprisal manipulation was introduced, and on Day 3. Day 3 included structure test using both previously trained and novel verbs; we expected the Surprisal group to perform better than Control on both tests. Additionally, we introduced individual comprehension test trials on Day 2 immediately after surprising items, to test for any immediate effects of surprisal on structure comprehension. If surprisal led to increased priming effects, too, we would expect the Surprisal group to perform better than the Control group in structure comprehension immediately after surprising passive items. In all comprehension tests (blocks and individual trials) our prediction of an advantage for the Surprisal group concerned the passive structure only, given that this was the structure affected by the surprisal manipulation. We did not expect to observe any effects on the active structure. In the Grammaticality Judgment Task, too, we expected the Surprisal group to perform better than the Control group in their ability to correctly discriminate between grammatical and ungrammatical passive sentences. We did not expect to see any significant differences between groups in their ability to discriminate between grammatical and ungrammatical sentences in the active form. Our secondary research question (RQ2) concerned the possible effects of surprisal on the development of explicit knowledge. Explicit knowledge of the novel structures was assessed by retrospective verbal report, with a debriefing questionnaire administered at the end of study. If expectation violation can promote the development of explicit knowledge, we would expect the SG to show higher rates of awareness than the CG.

This was the first study of a planned project involving data collection from different populations, both online and in the laboratory. Therefore, we also collected a set of cognitive measures (procedural learning abilities and verbal declarative memory) which mediated performance in a previous cross-situational learning study (Walker et al., 2020) in order to control for potential effects of individual differences. However, we did not have any specific predictions regarding possible interactions between the individual differences we measured and the surprisal manipulation.

Methods

Participants

To our knowledge, there are no previous studies that attempted to investigate the effect of expectation violation on structural learning. This means that we had no single point of reference we could use to estimate the potential size of the effect we were interested in, in order to determine a suitable sample size. Therefore, we based our group size calculations on a set of previous studies each investigating one aspect of our manipulations. Walker et al. (2020), from whom we adapted the cross-situational learning paradigm, tested two groups of 32 subjects, which was estimated by the authors to give .99 power for the simultaneous acquisition of two syntactic structures, based on the effect size from their previous study using the same paradigm. Greve et al. (2017), who investigated the effect of prediction error when learning for novel picture-word associations, used a range of group sizes from 20 to 36 subjects, the latter of which was calculated to have .75 power for one-shot declarative learning.

The effect we were interested in was the interaction between these two aspects, namely the effect of surprisal *on* the acquisition of syntactic structure. However, we had no means of estimating the effect size of a potential interaction. Therefore, we designed the study to have at least enough power to detect the two effects separately, on the assumption that this would be a necessary (although not necessarily sufficient) condition to detect the interaction, if one existed. Based on these considerations, we estimated that a group size of at least 35 participants would be the minimum sample size that we should use in the study.

76 native speakers of English (59 females, M_{AGE} = 31, *SD* = 7.62) were recruited via the online research platform Prolific (https://www.prolific.co/) and completed the study over the course of three consecutive days, receiving compensation of £12. The study was given ethics approval by the Education Ethics Committee at the University of York. Participants all reported living in the United Kingdom at the time of taking part in the study. Only one participant reported knowledge of any Scandinavian language (Norwegian) (upon which our artificial language was based); this was at the beginner level and they stated that they had never received formal instruction in the language. Participants were randomly assigned to either the Surprisal (n = 39) or Control (n = 37) group on the first day of the study. The slight numerical imbalance between groups is a consequence of attrition (i.e., participants were evenly assigned to the two conditions on Day 1, but not all completed all three days).

Materials

All stimuli and experimental scripts can be downloaded from the OSF repository for this study (https://doi.org/10.17605/OSF.IO/NKSU8) and from the IRIS database (https://www.iris-database.org/). Participants were trained in an artificial language called Yorwegian, consisting of four nouns (*glim, blom, prag, meeb* – man, woman, boy, girl), eight verbs (*flug-, loom-, gram-, pod-, zal-, shen-, norg-, klig-* – call, chase, greet, interview, pay, photograph, scare, and threaten), one determiner (*lu* - the) and one preposition (*ka* - by). Some of the lexical items used were adapted from Wonnacott, Newport, & Tanenhaus (2008). The specific word-meaning pairs within the noun and verb categories were randomly assigned for every participant. All sentences were SVO, but there were two possible syntactic structures, differentiated by verbal inflection and use of the preposition *ka*. These were the Active structure (e.g. *Lu meeb flug**at** lu prag,* 'The girl calls the boy') and the Passive (e.g. *Lu prag flug**es** **ka** lu meeb,* 'The boy is called by the girl'). This type of passive construction is naturally found in Scandinavian languages. It was chosen so as to have a way of forming passive structures that would not be entirely familiar to L1 English speakers (as there is no equivalent of the BE auxiliary in Yorwegian), while still being ecologically valid.

We used a set of 208 black and white photographs depicting transitive actions, which we adapted from materials created by Segaert and colleagues (Menenti, Gierhan, Segaert, & Hagoort, 2011; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012). The main set of training and testing pictures used on all three days (192 images) depicted the eight verbs: *call, chase, greet, interview, pay, photograph, scare,* and *threaten*. There were four characters which could fill the roles of Agent and Patient: *man, woman, girl* and *boy.* All possible combinations of different characters were included for each training verb, which yielded 12 possible Agent-Patient combinations (the Agent and Patient were always played by different characters). In the training set, the 12 Agent-Patient combinations were repeated for each of the eight verbs, yielding a total number of 96 possible scenes. Each scene was enacted twice, each with different actors, giving a total of 192 unique training pictures. Each picture could appear with one of two possible syntactic structures (Active and Passive constructions), for a total of 384 unique picture-sentence combinations.

The first 96 training pictures (Actor set 1) were used for training blocks on Day 1 and then again on Day 2. On Day 1, all training pictures appeared in the Active construction; on Day 2, half were presented in the Active, the other half in the Passive construction (which pictures appeared in each structure was counterbalanced across participants). Pictures from the second half (Actor set 2) were used for testing blocks distributed across the three days: vocabulary testing blocks on Day 1, Day 2, and Day 3, as well as structure test blocks on Day 2 and Day 3. No unique picture-sentence combination was presented more than once over the course of the experiment. An additional 'generalisation set' was also used (16 images). The pictures in this set depicted four additional transitive verbs (*dress, hug, pull,* and *push*) and were used in a generalisation structure test block on Day 3, to test participants' ability to process the syntactic structures they had been previously exposed to when used with novel verbs. This set used a reduced number of Agent-Patient combinations (four in total: *man-woman, woman-man, boy-girl, girl-boy*).

Procedure

Participants took part in the study online over the course of three consecutive days (Figure 1). The average total duration of the study was ~75 min, with each of the three sessions taking approximately 25 min. On each day, participants had to complete the session between 10am and 6pm. Subjects were randomly assigned to one of two groups, Surprisal or Control. On Day 1, the two groups followed the exact same protocol. On Day 2, all participants followed the same procedure in blocks 1-4. In blocks 2 and 3, feedback was introduced and was the same for both groups. In blocks 5-8, we introduced the between-group surprisal manipulation (described in the next section, on 'Learning trials with feedback'). On Day 3, both groups again followed the same protocol throughout. Participants performed the main task (the cross-situational learning paradigm) over the course of three days. On Day 3, this was followed by a grammaticality judgment task, a serial reaction task, the LLAMA B3 test, and a debriefing questionnaire. All tasks were created using JavaScript library PsychoJS, based on PsychoPy (Peirce et al., 2019), with the exception of the LLAMA B3 test, which was built in jsPsych (De Leeuw, 2015). All experimental scripts were hosted and run online through platform Pavlovia (https://pavlovia.org/). Surveys at the end of the experiment were administered using Qualtrics (www.qualtrics.com).

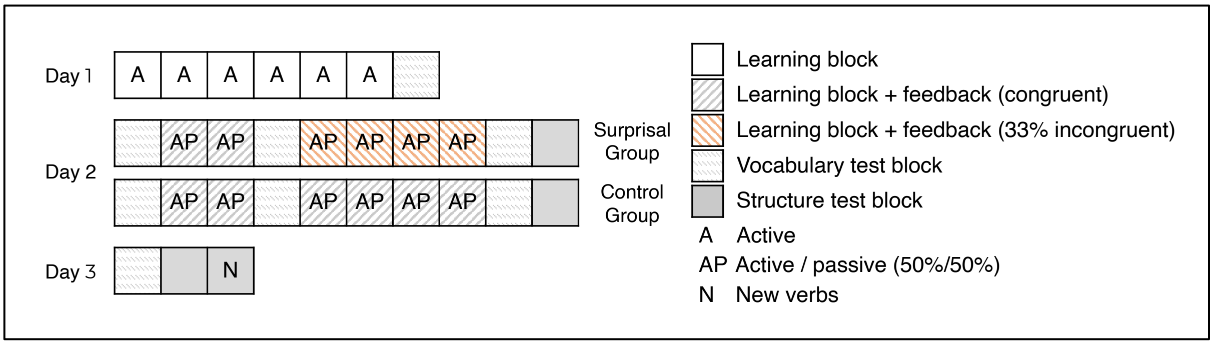


Figure 1. ***Summary of cross-learning task schedule***

Cross-situational Learning Task

Participants received no explicit instruction on either the grammar rules or vocabulary of Yorwegian. They were taught using an adapted version of the cross-situational task used by Walker et al. (2020), which was also used for testing. Participants heard individual sentences in Yorwegian, while two pictures (a target picture and a distractor picture) appeared on screen side by side. Their task was to select the picture that corresponded to the sentence they just heard (the target) by pressing the left or right arrow on their keyboard. There were four different types of trials: normal learning trials, vocabulary test trials, structure test trials, and learning trials with feedback (which included the critical between-group manipulation). In normal learning and testing trials, participants received no feedback on their answers.

Normal learning trials. Distractor Agent, Patient, and verb were picked by the experimental software at random, with the only constraint being that the distractor verb could not be the same as the target verb (to avoid the possibility of participants seeing two pictures depicting the same scene, only enacted by different actors).

Learning trials with feedback. On Day 2, all learning blocks (Blocks 2-3 and 5-8), contained a proportion of learning trials with feedback. 12 out of 16 learning trials in each of these blocks were followed by feedback on the answer just given: after making their choice (in a learning trial), participants were shown the correct picture which they should have picked, regardless of whether they had picked it or not (in a feedback screen). They saw the correct picture displayed on its own, in the centre of the screen, and they also heard the sentence which they had responded to once again. More precisely, they heard a sentence with the same Agent, Patient and verb as the one they had responded to, but, depending on the block and group, the syntactic structure used to describe the scene could either be the same (congruent feedback) or different (incongruent). In Blocks 2 – 3, all feedback was congruent and evenly spread across structures: both groups received feedback on 6 passive and 6 active learning trials per block, and the sentence they heard during feedback matched the one they had responded to, in both content (meaning) and structure. This was done to ensure that both groups would develop an expectation for feedback to replay sentences using the same structure.

Table 1. ***Types of trial included in critical learning blocks (Blocks 5 - 8). Differences between groups are highlighted in bold.***

Group	Feedback	Main structure (heard first)	Feedback structure	Type	n
Control	No	**Active**	-	No feedback	4
Control	Yes	Active	Active	Congruent	4
Control	Yes	Passive	Passive	Congruent	8
Surprisal	No	**Passive**	-	No feedback	4
Surprisal	Yes	Active	Active	Congruent	4
Surprisal	Yes	Passive	Passive	Congruent	4
Surprisal	Yes	**Active**	**Passive**	**Incongruent**	4

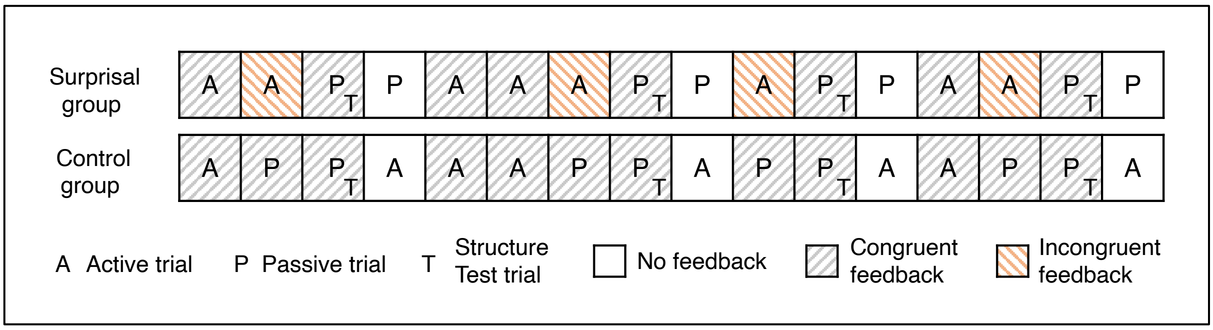


Figure 2. Example of a critical learning block (Blocks 5-8)

In Blocks 5 – 8, we introduced the between-group ‘surprisal’ manipulation. Feedback was still given on 12 out of 16 trials, and both groups still received congruent feedback on 8 of these 12 trials (4 active and 4 passive). The remaining 4 learning trials with feedback were manipulated so that the feedback they were followed by was congruent for the Control group, but incongruent for the Surprisal group (Table 1 & Figure 2). The only difference between congruent and incongruent feedback was the structure used: in both cases, the correct picture was shown, and the sentence which was heard had the same meaning as that heard during training. However, in incongruent trials the sentence was recast in the opposite syntactic structure (which was the 'incongruent' aspect), while in congruent feedback the exact same sentence was re-played, both with regards to meaning and syntactic structure used. In the Control group, these 4 critical trials required participants to respond to a passive sentence, while in the Surprisal group participants would respond to an active one. This was done to ensure that the feedback itself – the sentence learners were exposed after giving their answer, as they saw the correct picture again – would be in the passive for both groups. This manipulation meant that 8 of 12 trials with feedback used an active structure for the Control group, while for the Surprisal only 4 out of 12 were in the active form. To compensate for this imbalance and ensure the same amount of exposure to the structures in both groups, the remaining 4 trials in each block (which did not have feedback) were manipulated to be passive for the Surprisal group, and active for the Control group (Figure 2). Over the course of the whole experiment, participants saw 16 critical learning trials with feedback (with incongruent feedback for the Surprisal group, but congruent for Control), four in each of Blocks 5 to 8. Each of these critical trials was followed by a structure test trial, which is described below.

Structure test trials. All parameters in the pictures were kept constant apart from the Agent and Patient roles, which were reversed from target to distractor pictures (e.g., if the target picture was *The girl interviews the man*, the distractor would be *The man interviews the girl*). Distractor pictures were always picked randomly from either Actor set 1 or 2, regardless of which Actor set the target picture was drawn from (this was done to increase engagement and avoid creating a sense that there was any difference between blocks, which would have been the case if individual blocks only ever showed pictures from one particular set). The following parameters were always randomly chosen: the position of target and distractor picture on screen (left / right), and the position of Agent and Patient characters inside the pictures (left / right). Structure test trials were included in structure test blocks and also immediately following critical feedback trials.

Noun test trials. All parameters in the pictures were kept constant apart from the Patient noun (e.g., if the target picture was *The girl interviews the man*, the distractor could be *The girl interviews the boy* or *The girl interviews the woman*). Noun test trials were included in vocabulary test blocks only.

Verb test trials. All parameters in the pictures were kept constant apart from the verb. Verb test trials were included in vocabulary test blocks only.

Auditory Grammaticality Judgment Task

Following the cross-situational learning task on Day 3, participants did an auditory grammaticality judgment task (a widely used technique – see Plonsky, Marsden, Crowther, Gass, & Spinner (2020) with novel Yorwegian sentences. They were instructed to listen to each sentence and indicate whether it was a correct sentence in the language they had been learning. After each sentence was played, the words CORRECT and INCORRECT appeared side by side on screen, and participants had to press either the left or right arrow on their keyboard to give a response. Responses were untimed and ample time was given to respond; the next sentence was shown only after participants gave a response. They heard a total of 32 sentences, 16 grammatical and 16 ungrammatical. Half of the ungrammatical sentences contained the active verbal inflection followed by the agent marker, while the other had a passive verb, but no agent marker (see Table 2 for example stimuli).

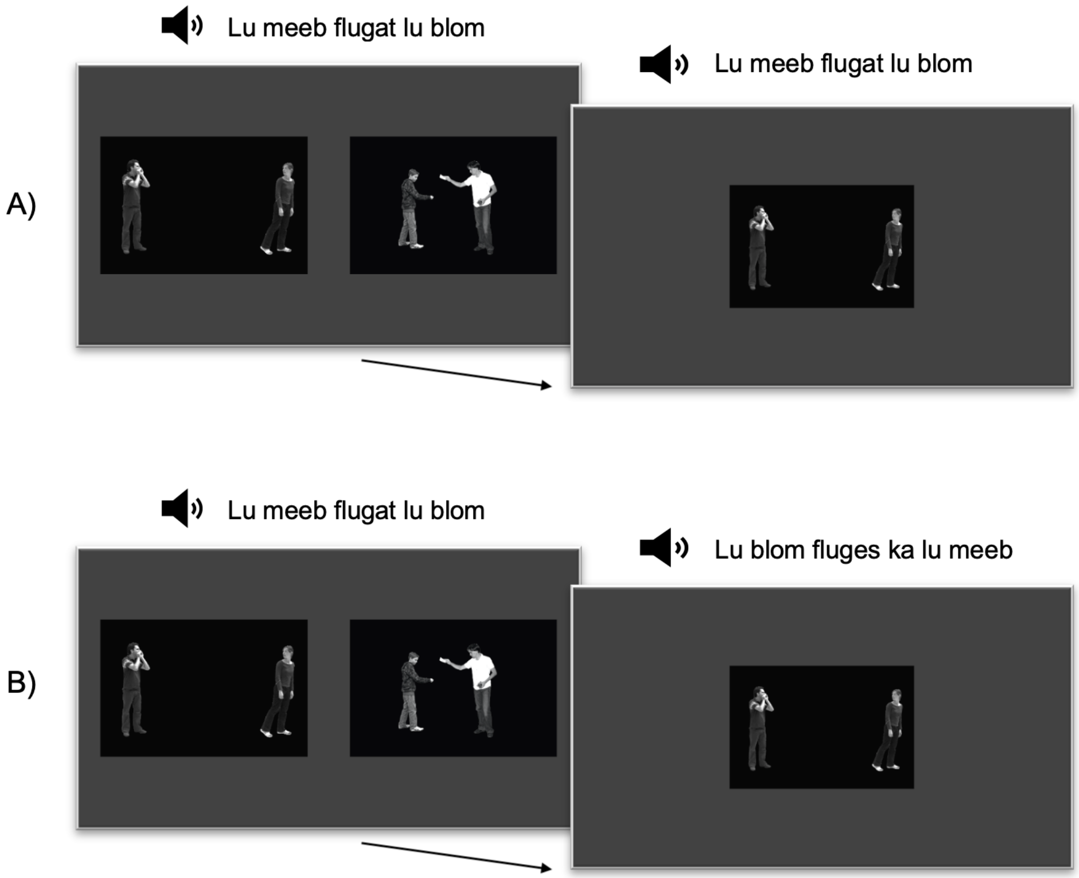


Figure 3. ***Learning trials with feedback, congruent (a) and incongruent (b)***

Language Background and Debriefing Questionnaires

At the end of Day 3, participants filled in a language background and debriefing questionnaire. The anonymised survey data can be downloaded from https://doi.org/10.17605/OSF.IO/NKSU8 and https://www.iris-database.org/ The first part of the questionnaire included questions on the participants' educational and language background, including the amount of formal grammar instruction received in the L1 and in any foreign languages spoken. The second part included specific questions on the experiment itself, aimed at probing participants' awareness of the structures and of the functional distinction between them ('Did you notice that a new type of sentence was introduced on Day 2 (yesterday's session)?', and if Yes, 'What were the two types of sentence you learned, and what do you think the difference was between them?').

Table 2. ***Types of sentences included in the Grammaticality Judgment Task***

Sentence type	Verb inflection	Example
Grammatical	Active	Lu meeb flug**at** lu blom
Grammatical	Passive	Lu blom flug**es ka** lu meeb
Ungrammatical (-at + ka)	Active	Lu meeb flug**at ka** lu blom
Ungrammatical (-es + Ø)	Passive	Lu blom flug**es** lu meeb

Individual Difference Measures Taken on Day 3

Serial Reaction Task.

A Serial Reaction Task (SRT) was administered to measure procedural learning abilities, following the paradigm used by Walker et al. (2020) and Lum, Gelgic, & Conti-Ramsden (2010). Participants saw a white square appear on one of four possible positions on screen (top, bottom, left and right), and had to press the corresponding arrow on their keyboard in response, as quickly and accurately as they could. The positions in which the square appeared followed a set sequence (bottom, top, right, left, right, top, bottom, right, top, left), which was repeated twice per block over five blocks (100 trials in total). The last block (Block 6) followed a different, pseudorandom sequence, repeated twice. Following Lum et al. (2010), the likelihood of the square appearing in any one particular position over the course of the pseudorandom sequence and the transitional probabilities between positions were kept consistent with those of the training sequence. To score the tasks, we followed Walker et al. (2020), subtracting the median RT for Block 5 from that for Block 6 (violation block).

LLAMA B3.

A vocabulary learning task (LLAMA B3) was included to measure verbal declarative memory. We replicated the design of the LLAMA B3 task (Meara & Rogers, 2019) using JavaScript library jsPsych (De Leeuw, 2015). Participants saw 20 drawings of novel fictional entities arranged in a grid on screen. Hovering over a drawing with the mouse cursor revealed a label with the written name of that entity. Participants were given 2 minutes to learn the names. At the end of the study period, the drawings appeared again, arranged in a different sequence. Participants were then given the names and asked to click on the relevant drawing (e.g., 'Click on the *taa*. If you are not sure, just guess'). All the drawings remained unchanged on screen throughout the test phase and could be selected at any time, and participants received no feedback on their answers.

Results

A total of 70 participants were included in the analysis (Table 3). Four participants were excluded for failing to listen to the items before giving their responses (the criterion response time for this exclusion decision was under 1s on at least six trials per block, in any given block). One participant was excluded due to suspect unfair means (such as taking notes, based on response times over 10s and 100% accuracy from Block 1 of the cross-situational learning task on Day 1). One participant was excluded for failing to finish the Day 2 task in one sitting.

Table 3. ***Descriptive statistics for analysed sample***

	Sex	Age	LLAMA B3	SRT
	F	Years	Score	RT (ms)
Group	n	*M (sd)*	*M (sd)*	*M (sd)*
Surprisal (n = 36)	29	30.9 (7.64)	6.77 (4.07)	45.82 (43.08)
Control (n = 34)	25	31.1 (7.63)	6.28 (4.54)	38.86 (33.06)

Cross-situational Learning Task

We analysed accuracy data as binary outcome (correct / incorrect) at the trial level. We used generalized linear mixed-effect models (GLMER) for binomial data, which we implemented in R version 4.0.3 (R Core Team, 2020) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). Following Barr, Levy, Scheepers, & Tily (2013) we used the maximal random structure supported by the model, in order to control for as much variance as possible. For each model, we first created a formula containing the maximal fixed effect structure and the maximal random effect structure (random intercepts by subject and item as well as random slopes for subjects and items by each of the fixed effect predictors, and their interactions). We used the package *buildmer* (Voeten, 2020) to automatically identify the maximal random structure that would allow the model to converge. We then used *buildmer* again on the resulting formula to do stepwise backwards model selection using likelihood-ratio tests, eliminating fixed effect predictors one by one (starting from higher-level interactions) and only retaining them if they significantly improved model fit. All models were checked for overdispersion and none of them showed signs of being overdispersed. We report the coefficients of the mixed-effect models converted to odds ratios (*OR*) to provide a measure of effect size, together with the statistical significance of the effects (*p* values). Full descriptive statistics for the cross-situational task on all three days can be found in Appendix S2. Final statistical models for all tests can be found in Appendix S5.

Learning Blocks

Learning trials were included in the cross-situational task on Day 1 (Figure 4) and on Day 2 (Figure 5). We analysed data from the learning trials on Day 1 (blocks 1 – 5) and Day 2 (blocks 2 – 3 and 5 – 8) in two separate models, entering Group and Block (centred) as fixed effects for each. There were no significant differences in performance between groups on learning blocks on either Day 1 or Day 2. There was a significant effect of Block on both Day 1 (*OR* = 1.28, 95% CI [1.21, 1.36], *p* < .001) and Day 2 (*OR* = 1.12, 95% CI [1.06, 1.18], *p* < .001).

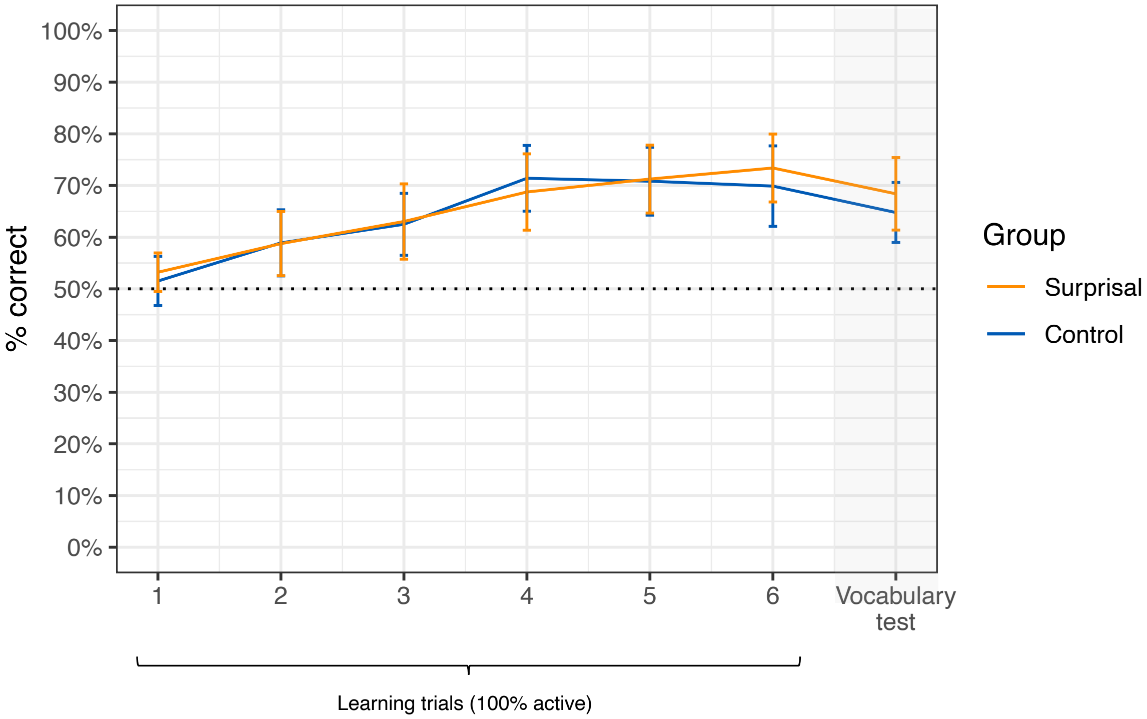


Figure 4. ***Accuracy on Day 1 by block. Error bars represent 95% CIs of group means (computed by averaging over subject means).***

Vocabulary Test Blocks

Participants undertook five vocabulary test blocks over the course of the experiment: one on Day 1 (Test 1), three on Day 2 (Test 2, 3 and 4), and one on Day 3 (Test 5) (Figure 4, Figure 5 & Figure 6). We entered data from all these tests together into *glmer* models with Group and Test (centred) as fixed factors, creating two separate models for verbs and for nouns. For verbs, there was no significant effect of Group, only a main effect of Test (*OR* = 1.33, 95% CI [1.24, 1.41], *p* < .001). For nouns, there was a significant interaction between Test and Group (*OR* = 0.80, 95% CI [0.69, 0.93], *p* = .004). In post-hoc comparisons, a difference emerged in the transition from Test 3 to Test 4 (second and third test blocks on Day 2), which led to a significant improvement in accuracy for the Control group ($\chi^2(1) = 12.28$, *p* = .009) but not for the Surprisal group ($\chi^2(1) = 1.87$, *p* = 1). However, the difference in accuracy between the two groups was not significant at any point (Test 1: $\chi^2(1) = 0.13$, *p* = 1; Test 2: $\chi^2(1) = 0.27$, *p* = 1; Test 3: $\chi^2(1) = 0.005$, *p* = 1; Test 4: $\chi^2(1) = 1.51$, *p* = 1; $\chi^2(1) = 5.20$, *p* = .113).

Structure Test Trials after Feedback

Individual structure test trials were inserted after feedback learning trials to test for any immediate (priming) effects as well as any cumulative effects of the experimental manipulation on structural knowledge. We analysed data from the Structure test trials in a *glmer* model with Group and Trial number (scaled and centred) as fixed factors. We observed a main effect of Trial (*OR* = 1.40, 95% CI [1.17, 1.67], *p* < .001), but no effects of Group.

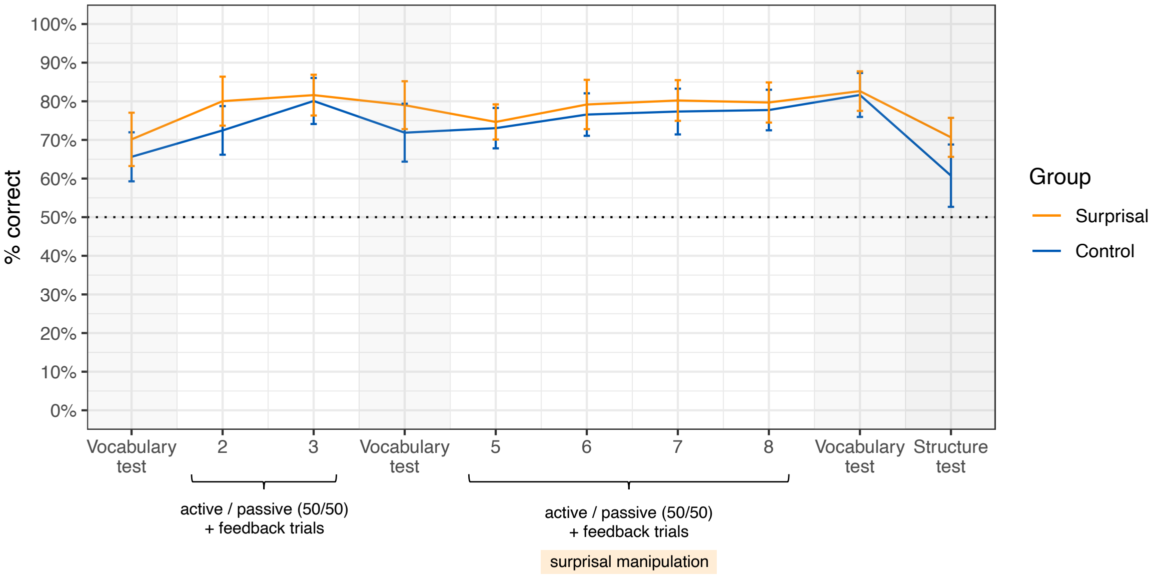


Figure 5. ***Accuracy on Day 2 by block. Error bars as in Figure 4.***

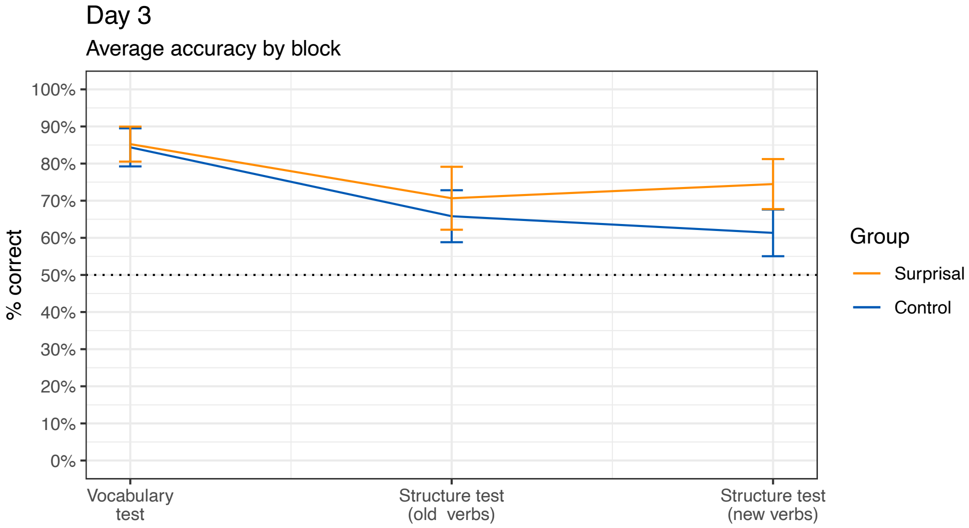


Figure 6. ***Accuracy on Day 3 by block. Error bars as in Figure 4.***

Structure Test Blocks

We analysed data from each of the three structure test blocks in individual *glmer* models, entering Group, Structure (Active vs. Passive) and their interaction as predictors in the initial model for each.

Day 2 (old verbs). In the structure test block on Day 2 (Figure 5), there was a numerical trend towards higher accuracy in the Surprisal group and for Active sentences, but no statistically significant effect of either Structure or Group. There was a great deal of variability between participants (see Appendix S4 for additional figures).
Day 3 (old verbs). In this test block, we found a significant effect of Structure, with higher accuracy for Active relative to Passive (*OR* = 2.80, 95% CI [1.50, 5.23], *p* = .001), but no effect of Group.[2]

Day 3 (new verbs). In the generalisation structure test block, there were significant main effects of Group (*OR* = 2.50, 95% CI [1.25, 4.99], *p* = .009) and Structure (*OR* = 2.82, 95% CI [1.42, 5.59], *p* = .003): Participants in the Surprisal group were more accurate than those in the Control group, and both groups had higher accuracy for active structures compared to passives (Figure 7).

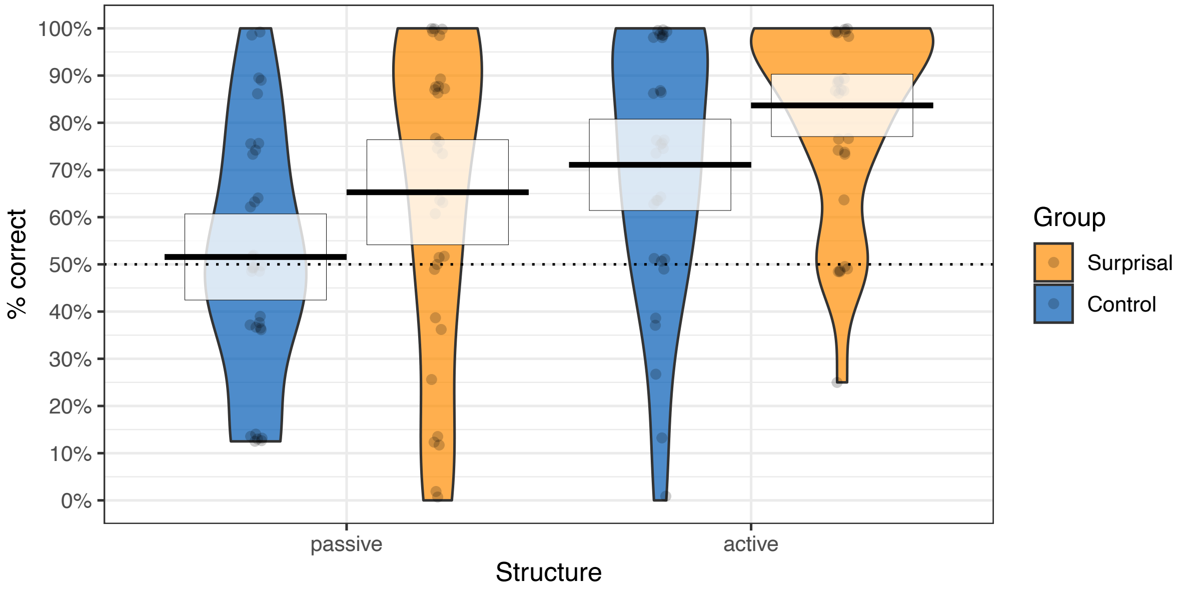


Figure 7. ***Accuracy on Day 3 structure test block (new verbs). Horizontal bars represent group means, shaded rectangles 95% CIs.***

Grammaticality Judgment Task

Descriptive statistics for this task can be found in Appendix S2, and the full statistical models are reported in Appendix S5. We analysed both raw endorsement rates, to capture differences in endorsement bias between groups, and *d'* scores to obtain a measure of sensitivity to grammaticality in the two groups. Endorsement data was collected as a binary response (Yes / No) so we analysed it using a generalized linear mixed-effect model (GLMER) for binomial data, following the same procedure we use

[2] We can only draw limited conclusions from the results of the Day 3 (trained verbs) test block, however, as this block was affected by a counterbalancing error which meant that half of the participants (equally spread among groups) saw the exact same items as in the Day 2 structure test (while the other half saw the same pictures but described using the opposite structure, which was the intended design). This does not affect the following test block (Day 3, new verbs), which used entirely novel Agent – Verb – Patient combinations. See Appendix S4 for a detailed figure of the Day 3 (trained verbs) block, including accuracy by group and counterbalancing status.

to analyse accuracy data from the cross-situational learning task. Group, Verb type (active vs. passive) and Grammaticality were entered as fixed predictors in the model.

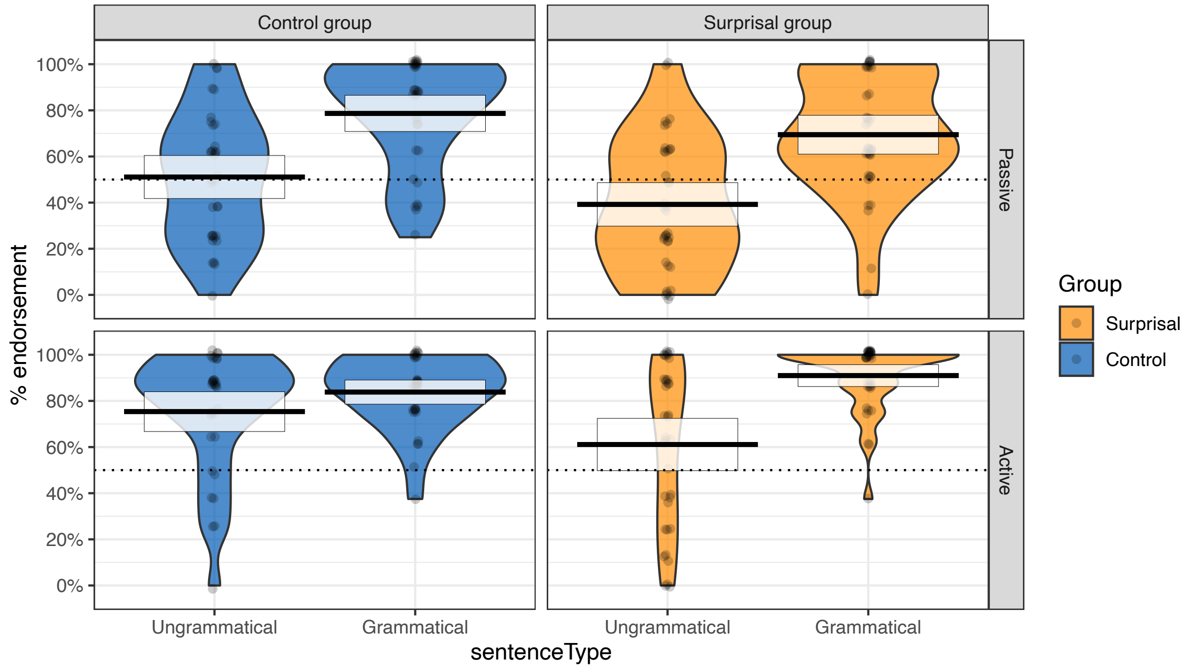


Figure 8. ***Endorsement rates in the Grammaticality Judgment Task (Day 3), by Group, sentence Grammaticality and Verb type. Horizontal bars represent group means, shaded rectangles 95% CIs.***

We found a three-way interaction between Group, Verb type and Grammaticality (*OR* = 4.44, 95% CI [1.85, 10.64], *p* = .001) (Figure 8). Post-hoc comparisons showed that this was driven by the effect of Grammaticality varying across groups, specifically for Active sentences. Participants in the Surprisal group were significantly more likely to endorse grammatical relative to ungrammatical sentences, whether they contained Active ($\chi^2(1) = 29.22$, *p* < .001) or Passive verb forms ($\chi^2(1) = 21.86$, *p* < .001). The Control group, on the other hand, showed the same effect of grammaticality with Passive sentences ($\chi^2(1) = 19.54$, *p* < .001) but not with Active ones ($\chi^2(1) = 1.57$, *p* = .84). The effect of grammaticality was of similar magnitude for Passive sentences in both groups (SG: *OR* = 0.18, 95%CI [0.06, 0.59], *p* < .00; CG: *OR* = 0.19, 95%CI [0.05, 0.66], *p* < .001) and in Active sentences for the Surprisal group (*OR* = 0.12, 95%CI [0.02, 0.62], *p* = .001). The Control group were more likely than the Surprisal group to endorse sentences with a Passive verb in general, regardless of their grammaticality ($\chi^2(1) = 5.74$, *p* = .033). In general, endorsement across groups was higher for Active then for Passive sentences, both grammatical (($\chi^2(1) = 10.30$, *p* = .003) and ungrammatical ($\chi^2(1) = 29.66$, *p* < .001).

To estimate participants' ability to discriminate grammatical from ungrammatical sentences, regardless of endorsement bias, we calculated *d'* scores for different item types and entered them in a mixed ANOVA using package *ez* (Lawrence, 2016), with Group and Verb type as predictors. The ANOVA returned a significant interaction between Group and Verb type (*F*(1,68) = 4.590, *p* = .036) but no significant main effects of either Group ($F(1,68) = 1.809$, $p = .183$) or Verb type ($F(1,68) = 0.230$, $p = .633$). We carried out post-hoc comparisons with the Bonferroni correction using package *emmeans* (Lenth et al., 2021), which showed a significant difference in d' scores between groups for Active sentences ($F(1,68) = 5.505$, $p = .04$) but not for Passive ones ($F(1.68) = 0.028$, $p = 1$).

Cognitive Measures

Basic descriptive statistics for LLAMA B3 and SRT scores can be found in Table 3; see Appendix S1 for detailed statistics. The groups did not significantly differ in their scores, and the effects of Group we reported for the cross-situational learning task and grammaticality judgment task were not affected by the inclusion of these cognitive measures in the analysis. See Appendix S1 for a detailed report of the analyses that included these measures as variables.

Debriefing Questionnaire (RQ2)

21 out of 36 subjects in the Surprisal group and 14 out of 34 subjects in the Control group developed sufficient explicit knowledge of the structures to be able to verbalise their respective functions. To assess whether the experimental manipulation had made participants in the Surprisal group more likely to develop explicit knowledge of the Active / Passive distinction, we constructed a simple logistic regression with explicit knowledge as a binary outcome and Group as predictor. While the Surprisal group had a numerically higher rate of explicit knowledge, the effect was not significant ($OR = 0.50$, 95% CI [0.19, 1.28], $p = .15$).

Discussion

Our first research question concerned the effect of surprisal on structural knowledge. We hypothesised that surprisal at the item level would lead to stronger abstract structural knowledge of the passive structure in the Surprisal group: Our results partially supported this hypothesis. We found that participants in the Surprisal group performed significantly better than those in the Control group in both a structure comprehension test and a grammaticality judgment task. Crucially, the structure test used novel verbs, which shows that the Surprisal group had developed stronger abstract knowledge than the Control group, and were able to use that knowledge to generalise structure to a new lexicon. However, we did not observe an effect of Group on the other structure test blocks or structure test trials in earlier blocks, which all used familiar verbs. Additionally, the effects we observed, were not limited to the passive construction as we had hypothesised, given that the manipulation was only on passive items. In the comprehension test, the advantage for the Surprisal group was found across both structures. In the Grammaticality Judgment Task, against our expectations, the main difference between groups emerged on active sentences, where only the Surprisal group showed a significant ability to distinguish grammatical from ungrammatical sentences.

Our secondary hypothesis was that surprisal would also lead to greater awareness of the functional distinction between active and passive constructions, measured as the ability to verbalise the distinction in retrospective verbal report. While there was a numerical advantage for the Surprisal group, this was not statistically significant. We consider these findings below, offering possible interpretations for the observed pattern of results and discussing the limitations of the current study.

Structural Accuracy in Comprehension

Against our expectations, we did not find an effect of Group in the structure tests which used familiar verbs (on either the structure tests blocks or the structure tests trials following feedback trials). We discuss potential explanations for these findings in the section on 'Study limitations' below. However, in the structure test on Day 3 (new verbs), we found a main effect of Structure, and one of Group: Both groups were better at selecting the correct interpretation of active sentences than they were for passive ones, and the Surprisal group was overall more accurate than the Control group. The effect of structure is compatible with our experimental design: Given that participants had received more and earlier exposure to this structure than to the passive, it is not surprising that they developed higher accuracy on it. We also expected the Surprisal group to perform better than the Control group in the structure test, which was confirmed. However, the effect was found for both Active and Passive structures (and was numerically greater for active ones), whereas we had expected to find an advantage specifically for passive sentences, given that they were the target of our experimental manipulation.

One possible explanation is that the mere presence of surprising trials led to greater attention and therefore better overall learning in the Surprisal group. In a series of cross-situational learning studies of vocabulary learning, Fitneva & Christiansen (2011, 2017) found that experiencing error (i.e., initially forming incorrect label-referent mappings) led to better learning in adults. Crucially, this effect was not limited to the words that participants had initially assigned to the wrong referent, but to the whole set of items, suggesting that experiencing error may have led to greater attention and better encoding of information overall (Fitneva & Christiansen, 2017).

A second possibility is that the effect was due to an interplay between the two structures: better knowledge of the passive construction could have led to higher accuracy on active trials, by providing negative evidence that helped participants rule out the incorrect alternative. In our structure test, the competitor (incorrect) picture always depicted the same action happening with Agent and Patient roles reversed, meaning the two constructions were effectively put in competition against each other. If the sentence was in the active form, e.g., *Lu meeb flug**at** lu prag* ('The girl calls the boy'), then the target picture would depict a girl calling a boy, while the competitor would depict a girl being called by a boy. This means that a sentence with the same nouns in the same positions as the target sentence could be used to describe the competitor picture, but only if it had different morphosyntax, that is, *Lu meeb flug**es ka** lu prag,* ('The girl is called by the boy'). Being sensitive to this distinction would help participants make the correct choice by ruling out the competitor picture, that is, by providing negative evidence of what the active sentence could *not* describe. Crucially, however, this requires specific sensitivity to the morphosyntactic distinction, which would in turn depend on accurate knowledge of the passive construction, as well as the active. Relying only on vocabulary would not be of help in this context, as both pictures could be described by sentences containing the same verb and nouns in the same order.

Yet another potential explanation for our findings is that the surprisal feedback trials did lead to better structural learning, but not in the way we had hypothesised. It is possible that what drove the effect of the surprisal feedback trials was actually the juxtaposition of an active and passive sentence used in sequence to describe the same event, rather than the passive feedback sentence being better encoded due to it being unexpected. This would have showed learners that the two structures could be used to describe the same event, potentially prompting them to pay more attention to the specific form-meaning mappings in the two structures. If learners follow a 'uniqueness principle' and assume that any given meaning can only be encoded by one grammatical form (Pinker, 2009), then the presence of two superficially equivalent forms may trigger a search for functional distinctions that may justify the existence of both forms in the grammar. We have no way to confirm or rule out this explanation given the currently available data. One future development of this research, however, will be to include a measure of item memory, testing for specific memory of the feedback sentences received in the critical feedback trials. If participants do show better memory for passive feedback sentences encountered in the surprising condition, this will lend support to our original hypothesis, that the surprisal manipulation improved memory for specific, individual items, which in turn lead to better generalisation. However, this would not entirely rule out a role for the second potential mechanism just described (i.e., juxtaposition of two structures leading to more accurate representations of structure-meaning mappings). In order to fully investigate this point, further research could include a different way to generate surprisal, that does not result in juxtaposition of an active with a passive sentence describing the same picture. If the same effects are observed, it would suggest that the effect of our experimental manipulation was not primarily driven by an artefact of our experimental design (the juxtaposition of two structures for the same event) but, rather, but the surprisal phenomenon itself.

In sum, further research needs to attempt to identify the explanatory power of these two accounts. However, it might also be worth bearing in mind that these two mechanisms could in fact—at least some of the time—be two sides of the same coin, working reciprocally, in tandem; that is, surprisal may serve to highlight meaning- (or function-) bearing linguistic contrasts, and, in turn, meaning-bearing contrasts may be a cause of surprisal events.

Development of Explicit Knowledge

Our second experimental hypothesis was that the surprisal manipulation may lead participants to develop a higher degree of awareness of the functional distinction between active and passive sentences. The data we collected does not allow us to satisfactorily answer this question, unfortunately. The debriefing questionnaire we used to measure explicit rule knowledge showed a numerical difference between groups, with higher rates of awareness among the Surprisal group; however, this was not significant in a statistical test. It is possible that the questionnaire may simply have been underpowered: due to the lack of previous research using a similar manipulation, we did not have a reference effect size which we could expect to see in this study, with regards to explicit knowledge[3]. Future developments of this research could employ larger sample sizes to address the possibility that our manipulation did have an effect on rule awareness, but that the effect was too small to be detected with sufficient confidence in our sample.

In addition to the potential lack of power, our measure of awareness was admittedly not a fine-grained one. It merely set a threshold based on retrospective verbal report, to divide participants into two categories (aware and unaware). Retrospective verbal report has been criticised for a potential lack of sensitivity to awareness; for instance, lack of confidence may lead to underreporting in some participants (Rebuschat, 2013). It could still be that awareness itself was emerging in a graded manner as structural representations were becoming stronger and more stable in the Surprisal group, in a way that is not captured by our cut-off point (the ability to verbalise the functional distinction between structures). This kind of graded emergence of awareness was the other potential mechanism that we hypothesised may have led to increased awareness in the Surprisal group, compatible with the radical plasticity theory of the relation between implicit and explicit knowledge (Cleeremans, 2008, 2011).

A related question, answers to which can only remain speculative for now, concerns the extent to which accuracy in the structure test may have been driven by explicit knowledge. While the difference between groups in terms of their reported awareness of the structure was not statistically significant, there was a numerical advantage for the Surprisal group. A higher degree of structural awareness may have helped participants in the Surprisal group to perform better in the structure test, once they did become aware (or were 'on their way to' awareness). However, given that awareness was assessed at the end of the study by verbal report, we do not know at which point participants did become sufficiently aware of the distinction to influence accuracy. Knowing that tipping point would be a prerequisite for any analysis aiming to use awareness as a predictor for accuracy. To address both of these points—the gradual emergence of awareness, and the extent to which it contributed to performance in the structure test—future research would need to include more fine-grained measures of awareness administered as the trials progressed, such as source attribution (Dienes & Scott, 2005) or the use of multiple direct and indirect tests to tease apart the contribution of different types of knowledge (Ellis Rod, 2009). The challenge for that line of research is to avoid 'reactivity', whereby the probe of awareness itself promotes, or interferes with, actual awareness (Bowles, 2010).

Grammaticality Judgment Task

In the grammaticality judgment task, we found further evidence that the Surprisal group had developed better structural knowledge than the Control group, broadly

[3] The effect size we observed in the Day 3 Structure Test on new verbs was Cohen's d = .64 (a medium effect size according to Cohen's (1988) benchmark, but a small one in the context of L2 acquisition research (Plonsky & Oswald, 2014)). Based on this effect size, we carried out a post-hoc power analysis in G*Power, which showed that the study had .84 power to detect the effect which we observed in structural comprehension. However, this does not provide an indication of the power the study had to detect a potential effect in awareness, which may be smaller than the one on structural comprehension (see section on Study limitations).

supporting our experimental hypothesis. However, we did not find the exact effect that we had anticipated, that is, greater accuracy in discriminating grammatical from ungrammatical sentences by the Surprisal group on passive items, compared to the Control group. Instead, *both* groups were significantly more likely to endorse grammatical sentences relative to ungrammatical ones in the passive structure, and the magnitude of the effect was the same in both groups. The two groups, however, differed in their overall likelihood to endorse passive sentences—irrespective of grammaticality—in that endorsement of passives was lower in the Surprisal group. This was not predicted by our experimental hypothesis. It may reflect a greater sense in the Surprisal group of the fact that the passive was an entirely new structure, while the Control group was more accepting of all sentences that resembled items they encountered during training.

On the other hand, a significant interaction between grammaticality and group—indicating that the groups differed in their ability to discriminate between grammatical and ungrammatical sentences—did emerge, but only for Active items. Here, the difference was remarkable: The Surprisal group showed a difference in endorsement rates between grammatical and ungrammatical sentences which was statistically significant and comparable in size to the effect observed for passive items. The Control group, by contrast, showed practically no difference in their endorsement of grammatical and ungrammatical items, and were equally likely to endorse any sentence containing an active verbal form (ending in *-es*), regardless of whether it was used in a grammatical way. An analysis of *d'* scores confirmed that the Surprisal and Control group differed significantly in their ability to discriminate grammatical from ungrammatical sentences, but only when they were in the active form.

This pattern may be explained by considering the way in which ungrammatical items were constructed in the grammaticality judgment task. These items mixed morphosyntax from different structures to create sentences that were unattested in the input participants had thus far received. Specifically, ungrammatical active sentences contained the active verbal suffix -*at* followed by the passive agent marker *ka*, while ungrammatical passive sentences contained the passive verbal suffix *-es* without the agent marker *ka* (see Table 2 for example stimuli). It appears that participants in the Control group were equally likely to endorse any sentence that contained chunks they had already encountered in training: either a verb with an active suffix, or a verb with a passive suffix followed by *ka*. When one of these chunks was broken—as happened in the case of ungrammatical passive sentences, which had a passive verb suffix but no *ka*—they were sensitive to this violation, resulting in lower endorsement rates. However, when a chunk was found in its entirety, as previously attested (verb + active suffix), but followed by a novel element—as in ungrammatical active sentences, where the active verb inflection was followed by *ka*—they did not perceive this as violating an established pattern. This could indicate that they were paying less attention to the material that followed the verbal inflection in the sentence, compared to the Surprisal group.

By contrast, the Surprisal groups showed equal sensitivity to ungrammatical usage of both active and passive verb forms, showing that they also paid attention to the material following the verbal inflection, resulting in lower endorsement for active inflections being followed by a novel item (the *ka* marker). This suggests that they had developed a more sophisticated kind of knowledge than the Control group. They had not only acquired the individual forms for active and passive inflection (and its associated marker), but, crucially, they had learned better that the two forms were associated with a different order of Agent and Patient. This suggests that they paid attention to both the material that followed the verb as well as that preceding it. In the grammaticality judgment task, this allowed them to discriminate between grammatical and ungrammatical usage of both verbal suffixes.

The lack of sensitivity shown by the Control group to grammaticality in active items is seemingly at odds with the results of the structure test, where they were able to pick the correct interpretation for active sentences with reasonably good accuracy. However, in the active items in the structure test, participants did not technically need to pay attention to the noun following the verb (the Patient) to answer correctly. Just correctly identifying the first noun as the Agent of the action, in combination with the active inflection, would suffice to answer correctly. Therefore, considering the results of both the structure and the grammaticality judgment task together, it is possible that participants in the Control group had settled on a basic heuristic, namely identifying the first noun as the Agent of the sentence (independently of verbal inflection), which was sufficient to answer correctly to active sentences in the structure test. This is also compatible with the fact that they were essentially at chance level in their responses to passive items in the structure test. By contrast, the Surprisal group could rely on additional cues for determining the correct meaning of the sentences (by attending to the material that came after first noun), resulting in higher accuracy on *both* active and passive items.

Crucially, however, it is not the case that Control participants simply never paid attention to the second noun in sentences. There was a specific task in the study—namely, the vocabulary test blocks—which could only be performed correctly by paying attention to the Patient noun (always in second position, since all sentences in vocabulary test blocks were active). There was no significant difference between groups in vocabulary test blocks, indicating that both groups were attending to the relevant noun. When the task did not specifically demand it (as in the structure test), however, the Control group did not seem to attend to the material following the verbal inflection. This suggests that they had developed little sensitivity to the relation between noun position, verb form, and sentence meaning. The Control group learned that different verbal forms existed, but their knowledge of the structures they were found in, with the relevant form-meaning connections—that is, the different assignment of Patient and Agent roles—was reduced, relative to that of the Surprisal group. In turn, this resulted in lower accuracy in the Control group in the comprehension task, too, for passive sentences.

Study Limitations

Against our expectations, we did not observe an effect of group in structure tests that used previously trained verbs, in either the structure test blocks (on Days 2 and 3) or in the individual test trials following feedback (on Day 2). To some extent, these findings may be explained by limitations in the structure tests themselves. As we previously mentioned, the first structure test on Day 3 (old verbs) was affected by a counterbalancing problem. The lack of an effect on the individual test trials following feedback, too, could be due to limitations in the study setup. These structure test trials were placed after 'critical' learning trials, that is, those which included incongruent passive feedback in the Surprisal group, and their congruent passive counterparts in the Control group. We hypothesised that surprisal may lead to a stronger structural priming effect in the Surprisal group, which may manifest itself as higher accuracy on passive structure test trials. However, congruent trials in the Control group involved passive feedback presented after a passive sentence, meaning that Control participants were exposed to two passive sentences in a row, leading to potential cumulative priming. Therefore, it is possible that even if any effect due to surprisal was present, its effects relative to the Control group may have been obscured by cumulative priming effects in the Control group. This could have potentially cancelled out any differences between groups.

However, these problems do not affect the first structure test block using old verbs, which was administered at the end of Day 2. Why did we observe an effect of Group in the generalisation test on Day 3, but not in the old verbs test on Day 2? One possibility is that the surprisal effect, which we hypothesised to affect memory formation for critical passive sentences, may have required overnight sleep for memory consolidation and abstraction to take place[4]. The possible need for overnight consolidation was one of the reasons behind the decision to add tests on Day 3, in addition to the test at the end of Day 2. Under this interpretation, we should have observed an effect on first Day 3 structure test (old verbs), too. However, the technical error affecting this test blocks means that we have no conclusive evidence on this point. Further research replicating this design would be needed to provide evidence in support of this hypothesis about the role of sleep consolidation.

There are, however, other indicators from the study suggesting that the surprisal manipulation did not work as intended, i.e., by generating stronger memories for passive sentences when presented in surprising feedback. One point, which we already raised in the discussion, was that the effect on structural comprehension was found for both structures, not just the passive. Since only the passive was meant to be affected by our surprisal manipulation, it seems that the manipulation did not have the effect it was meant to have. We mentioned in the discussion the possibility that juxtaposition between structures in incongruent trials may have caused the effect we observed, by leading to higher awareness of the rule. The results of the debriefing questionnaire are not conclusive in this respect: they show a numerical difference between groups, which, however, is not significant. We have discussed the possibility that, while the study appeared sufficiently powered to detect the effect on structural comprehension, it may not have been sufficient to detect potentially smaller effects on awareness (footnote on p. 25). Indeed, sensitivity to differences between structures as a result of juxtaposition could have been stronger in the Surprisal group than in the Control group, leading to better performance in comprehension, but still not strong enough to lead to the level of explicit awareness needed to verbalise the distinction.

[4] While we are not aware of any research on the consolidation of syntactic structure, work using novel (artificial) L2 morphology shows an effect of overnight consolidation on the acquisition of new systematic patterns (Mirković et al., 2019; Tamminen et al., 2015).

Similarly, in the Grammaticality Judgment Task we observed an effect of Group but only for active sentences, not for passive ones, which is at odds with the fact that our manipulation was intended to target passive sentences. In the discussion, we offered a potential explanation of the results of the Grammaticality Judgment Task based on different patterns of attention in the two groups: we hypothesised that the Surprisal groups had developed stronger sensitivity to the fact that different morphosyntax on the verb correlated with different orderings of Agent and Patient, while the Control group relied on an 'Agent first' heuristic which accounted for their low performance in structural comprehension of passive sentences. However, the juxtaposition of active and passive sentences in the feedback, a potential limitation (confound) in the design, could plausibly have caused such an effect, too.

Finally, we should point out that the manipulation we used, whatever its effects, was quite subtle: there were only 4 critical trials per block, for a total of 16 over the whole experiment. Additionally, the expectation for congruent feedback—which was necessary for participants to experience surprisal at incongruent feedback—was only set up over the course of the first two learning blocks on Day 2 (a total of 24 trials with congruent feedback), which may have been insufficient to set up sufficiently strong expectations for congruent feedback to influence some of the dependent variables examined. In sum, some of the limitations and incongruities in this study may also be the result of a relatively weak manipulation. Future developments of this study could use a stronger surprisal manipulation, which may shed more light on some of the issues raised in this section.

Conclusion

In this study, we examined the effect of expectation violation on the acquisition of novel syntactic structures. Specifically, we examined the acquisition a minority syntactic structure (passive) introduced after the default structure (active) had been consolidated. We hypothesised that presenting instances of the passive structure in a way that violated expectations (surprisal) would lead to better acquisition of the passive structure itself, and greater awareness of its function. Our predictions with regards to accuracy were mainly supported: Although the pattern of results did not support the prediction of an isolated effect on only the passive structures, it clearly demonstrated that the Surprisal group developed stronger and more accurate structural representations than the Control group, for both constructions. In contrast, the experimental manipulation did not lead to statistically significantly sufficient levels of awareness to lead to knowledge that could be articulated explicitly, despite a numerical trend in that direction. The lack of statistical significance could be due to a number of design and methodological limitations, however, and the role of explicit knowledge should be investigated further. Nevertheless, it seems intriguing to us that a very simple manipulation, on a relatively small number of trials, had quite significant consequences for the representations developed by the two groups, and seemed to lead to different patterns of attention, too. Further research will be needed to investigate the effects we found, and to pinpoint their exact origin, among the different explanations we offered.

References

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54*(3), 218–250. doi:10.1016/j.cogpsych.2006.07.001

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013-4). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). doi:10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355–387. Retrieved from http://www.linguisticsnetwork.com/wp-content/uploads/Syntactic-Persistance_Bock-1986-.compressed.pdf

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition, 104*(3), 437–458. doi:10.1016/j.cognition.2006.07.003

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: transient activation or implicit learning? *Journal of Experimental Psychology. General, 129*(2), 177–192. doi:10.1037//0096-3445.129.2.177

Bowles, M. A. (2010). *The think-aloud controversy in second language research* [X, 172 p. : ill. ; 24 cm.]. New York ; Abingdon: Routledge.

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction, 55,* 22–31. doi:10.1016/j.learninstruc.2018.01.013

Bybee, J. L., & Hopper, P. J. (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins Publishing.

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures: An analysis of learning outcomes and processes. *Studies in Second Language Acquisition, 38*(2), 265–291. doi:10.1017/S0272263116000139

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. doi:10.1037/0033-295X.113.2.234

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass, 6*(5), 259–278. doi:10.1002/lnc3.337

Cintrón-Valentín, M. C., & Ellis, N. C. (2016). Salience in Second Language Acquisition: Physical Form, Learner Attention, and Instructional Focus. *Frontiers in Psychology, 7,* 1284. doi:10.3389/fpsyg.2016.01284

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. *Progress in Brain Research, 168,* 19–33. doi:10.1016/S0079-6123(07)68003-0

Cleeremans, A. (2011). The Radical Plasticity Thesis: How the Brain Learns to be Conscious. *Frontiers in Psychology, 2,* 86. doi:10.3389/fpsyg.2011.00086

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. doi:10.3758/s13428-014-0458-y

De Loof, E., Ergo, K., Naert, L., Janssens, C., Talsma, D., Van Opstal, F., & Verguts, T. (2018). Signed reward prediction errors drive declarative learning. *PloS One, 13*(1), e0189212. doi:10.1371/journal.pone.0189212

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*(5), 338–351. doi:10.1007/s00426-004-0208-3

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (2016). SALIENCE, COGNITION, LANGUAGE COMPLEXITY, AND COMPLEX ADAPTIVE SYSTEMS. *Studies in Second Language Acquisition, 38*(2), 341–351. doi:10.1017/S027226311600005X

Ellis, N. C. (2017). Salience in usage-based SLA. In S. M. Gass, P. Spinner, & J. Behney (Eds.), *Salience in Second Language Acquisition* (1st ed., pp. 21–40). doi:10.4324/9781315399027

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of construction grammar.* Hoboken, NJ: Wiley.

Ellis Rod. (2009). Implicit and explicit knowledge in second language learning, testing and teaching. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42).

Farmer, T., Fine, A., Yan, S., Cheimariou, S., & Jaeger, F. (2014). Error-Driven Adaptation of Higher-Level Expectations During Reading. *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*(36). Retrieved from https://escholarship.org/uc/item/00t2m3wr

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science, 7*(11), 180877. doi:10.1098/rsos.180877

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16*(1), 88–92. doi:10.3758/PBR.16.1.88

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes, 21*(7–8), 1011–1029. doi:10.1080/016909600824609

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 42*(9), 1362–1376. doi:10.1037/xlm0000236

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PloS One, 8*(10), e77661. doi:10.1371/journal.pone.0077661

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science, 35*(2), 367–380. doi:10.1111/j.1551-6709.2010.01156.x

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental Changes in Cross-Situational Word Learning: The Inverse Effect of Initial Accuracy. *Cognitive Science, 41 Suppl 1,* 141–161. doi:10.1111/cogs.12322

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language, 94,* 149–165. doi:10.1016/j.jml.2016.11.001

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology. General, 148*(2), 325–341. doi:10.1037/xge0000498

Hartsuiker, R. J., Kolk, H. H. J., & Huiskamp, P. (1999). Priming Word Order in Sentence Production. *The Quarterly Journal of Experimental Psychology Section A, 52*(1), 129–147. doi:10.1080/713755798

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition, 75*(2), B27-39. doi:10.1016/s0010-0277(99)00080-3

Jackson, C. N. (2018). Second language structural priming: A critical review and directions for future research. *Second Language Research, 34*(4), 539–552. doi:10.1177/0267658317746207

Jackson, C. N., & Ruf, H. T. (2017). The priming of word order in second language German. *Applied Psycholinguistics, 38*(2), 315–345. doi:10.1017/S0142716416000205

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*(1), 57–83. doi:10.1016/j.cognition.2012.10.013

Kaan, E., & Chun, E. (2018a). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition, 21*(2), 228–242. doi:10.1017/S1366728916001231

Kaan, E., & Chun, E. (2018b). Syntactic Adaptation. In K. D. Federmeier & D. G. Watson (Eds.), *Psychology of Learning and Motivation* (Vol. 68, pp. 85–116). doi:10.1016/bs.plm.2018.08.003

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition, 35*(5), 925–937. doi:10.3758/bf03193466

Kaschak, M. P., & Borreggine, K. L. (2008). Is long-term structural priming affected by patterns of experience with individual verbs? *Journal of Memory and Language, 58*(3), 862–878. doi:10.1016/j.jml.2006.12.002

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: cumulative priming effects and individual differences. *Psychonomic Bulletin & Review, 18*(6), 1133–1139. doi:10.3758/s13423-011-0157-y

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition, 99*(3), B73-82. doi:10.1016/j.cognition.2005.07.002

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments.* Retrieved from https://cran.r-project.org/package=ez

Ledoux, K., Traxler, M. J., & Swaab, T. Y. (2007). Syntactic priming in comprehension: evidence from event-related potentials. *Psychological Science, 18*(2), 135–143. doi:10.1111/j.1467-9280.2007.01863.x

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., & Singmann, H. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* Retrieved from https://cran.r-project.org/package=emmeans

Leow, R. P. (2015). *Explicit Learning in the L2 Classroom: A Student-Centered Approach.* Retrieved from https://play.google.com/store/books/details?id=rKzABgAAQBAJ

Lum, J. A. G., Gelgic, C., & Conti-Ramsden, G. (2010). Procedural and declarative memory in children with and without specific language impairment. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists, 45*(1), 96–107. doi:10.3109/13682820902752285

McDonough, K., & Trofimovich, P. (2015). Structural priming and the acquisition of novel form-meaning mappings. In T. Cadierno & S. Wind Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 105–123).

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA Tests v3. LLAMA B3 v3.00*. Cardiff: Lognostics.

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science, 22*(9), 1173–1182. doi:10.1177/0956797611418347

Monaghan, P., Ruiz, S., & Rebuschat, P. (2020). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research,* 0267658320927741. doi:10.1177/0267658320927741

Montero-Melis, G., & Jaeger, F. T. (2020). Changing expectations mediate adaptation in L2 production. *Bilingualism: Language and Cognition, 23*(3), 602–617. doi:10.1017/S1366728919000506

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. doi:10.3758/s13428-018-01193-y

Peter, M. S., & Rowland, C. F. (2019). Aligning Developmental and Processing Accounts of Implicit and Statistical Learning. *Topics in Cognitive Science, 11*(3), 555–572. doi:10.1111/tops.12396

Pinker, S. (2009). *Language Learnability and Language Development.*

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research, 26*(4), 583–621. doi:10.1177/0267658319828413

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Retrieved from R Foundation for Statistical Computing website: https://www.R-project.org/

Rebuschat, P. (2013). Measuring Implicit and Explicit Knowledge in Second Language Research: Measuring Implicit and Explicit Knowledge. *Language Learning, 63*(3), 595–626. doi:10.1111/lang.12010

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition, 206,* 104475. doi:10.1016/j.cognition.2020.104475

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension--an FMRI study. *Cerebral Cortex , 22*(7), 1662–1670. doi:10.1093/cercor/bhr249

Shin, J.-A., & Christianson, K. (2012). Structural Priming and Second Language Learning. *Language Learning, 62*(3), 931–964. doi:10.1111/j.1467-9922.2011.00657.x

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

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition, 163,* 1–14. doi:10.1016/j.cognition.2017.02.008

Voeten, C. C. (2020). *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression.* Retrieved from https://cran.r-project.org/package=buildmer

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning, 70*(52). doi:10.1111/1467-923X.12837

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2019). Primed From the Start: Syntactic Priming During the First Days of Language Learning. *Language Learning, 69*(1), 198–221. doi:10.1111/lang.12327

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 56*(3), 165–209. doi:10.1016/j.cogpsych.2007.04.002

Yu, C., & Smith, L. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18*(5), 414–420. doi:10.1111/j.1467-9280.2007.01915.x

Data, code and materials availability statement

All data, analysis code and materials used in this study can be found from the OSF repository for this study at https://doi.org/10.17605/OSF.IO/NKSU8 and from the IRIS database at https://www.iris-database.org/.

Authorship and contributorship statement

Both authors of this paper meet the criteria for authorship set out by the International Committee of Medical Journal Editors and listed in Author Guidelines for *Language Development Research.* Specifically, the authors contributed to the following aspects: Giulia Bovolenta: experiment design, data collection, data analysis and interpretation, manuscript preparation. Emma Marsden: experiment design, data analysis and interpretation, manuscript preparation.

Appendix S1 – Individual differences

We report here the results of the two cognitive tests that were administered: Serial Reaction Task and LLAMA B3. In order to determine whether the observed effects of our experimental manipulation were independent of any individual differences captured by these cognitive measures, we ran a series of *glmer* models where we added the z-transformed scores from cognitive measures as fixed predictors, in addition to the factors already entered in the main analysis. As random effects structure, we used the same structure that was originally used for the corresponding models in the main analysis. As in the main analysis, we used *buildmer* to simplify the models to only retain predictors that significantly improved model fit.

Serial Reaction Task

There was no significant difference in mean SRT score between the two groups (t(65.32) = -0.7601, p = 0.45). The distribution of scores is shown in Figure 9. Both groups deviated from the normal distribution to some extent, which was significant in a Shapiro-Wilk test for the Control group (W = 0.929, p = 0.03) but not for the Surprisal group (W = 0.953, p = 0.13). When adding SRT scores to the model for accuracy on the Day 3 structure test block (new verbs), the effects of Group and Condition were still observed, in addition to a negative effect of SRT score (Table 4 & Figure 10). When adding SRT scores to the model for item endorsement in the GJT, SRT score was removed as a predictor during model selection as it had no significant effect on model fit, while the original Group x Verb Type x Grammaticality remained significant. We report the output of the initial model (with maximal fixed structure) for reference (Table 5).

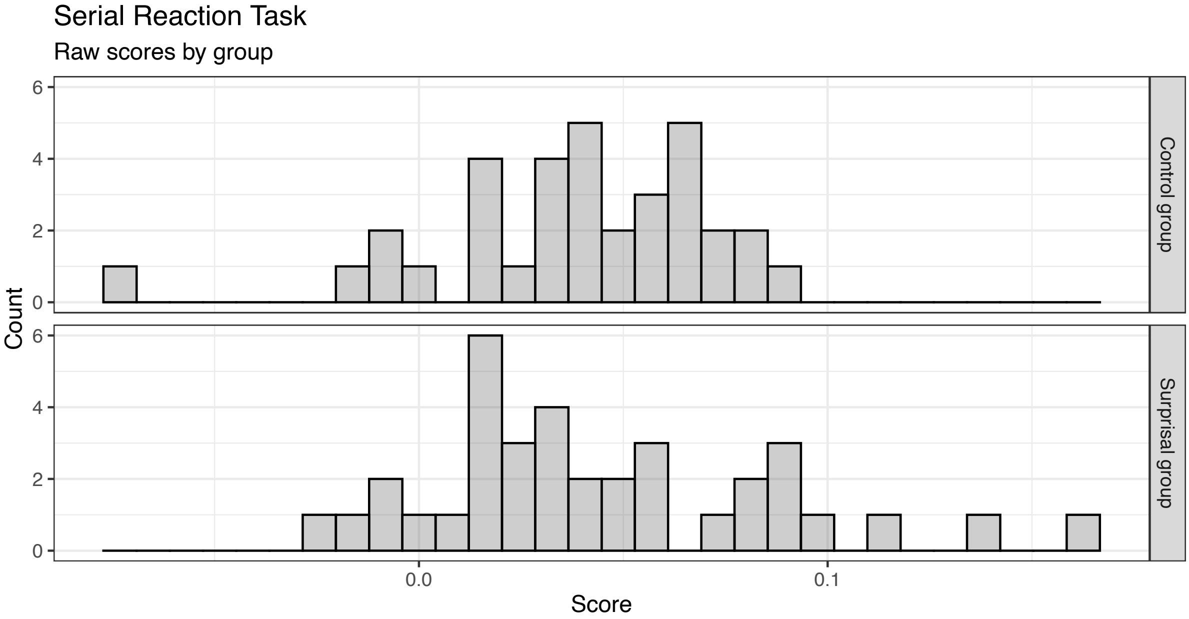


Figure 9. ***SRT scores by group***

Table 4. ***Final model for accuracy on Day 3 structure test (new verbs) with SRT score added to fixed effects structure***

	Accuracy		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.14	0.67 – 1.94	0.622
Structure (Active)	3.09	1.57 – 6.05	**0.001**
Group (Surprisal)	2.65	1.42 – 4.95	**0.002**
SRT score	0.70	0.52 – 0.95	**0.023**
Observations	1120		
Marginal R^2 / Conditional R^2	0.102 / 0.483		
Random effects: (1 + Structure | subject)			

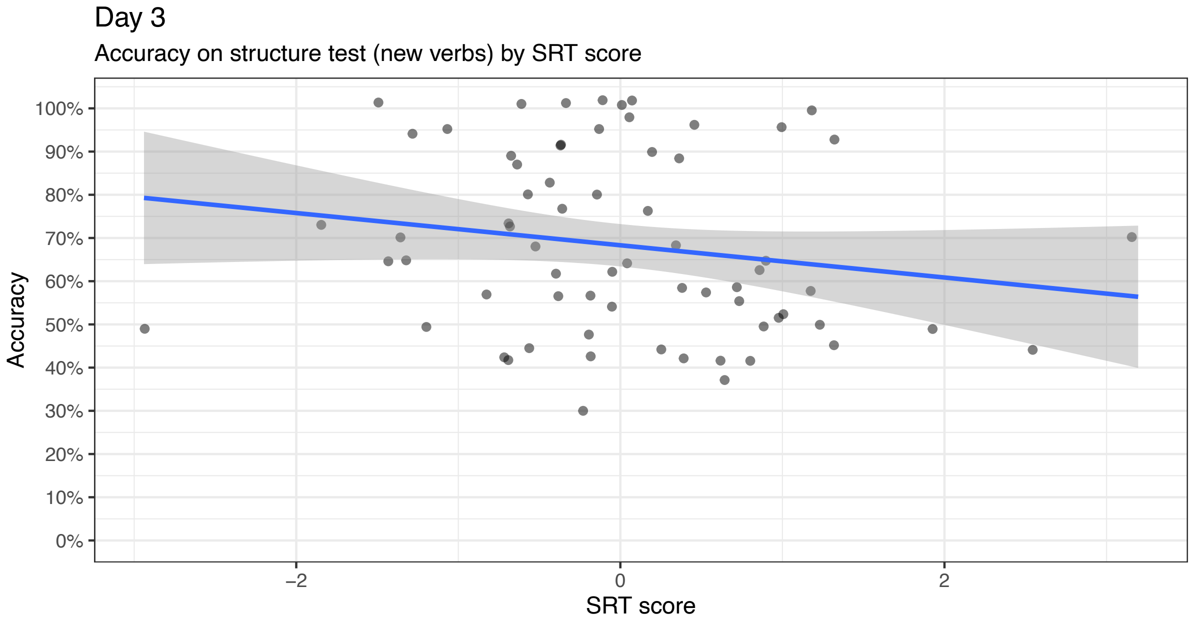


Figure 10. ***Accuracy on Day 3 structure test (new verbs) plotted against z-transformed SRT score***

Table 5. ***Initial model for endorsement in GJT, with SRT score added to fixed effects structure***

	Endorsement		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.09	0.66 – 1.78	0.740
cond [SG]	0.51	0.25 – 1.02	0.056
sentenceType [Grammatical]	5.25	2.38 – 11.58	**<0.001**
verbType [Active]	4.60	2.23 – 9.50	**<0.001**
SRT_z_score1	1.21	0.68 – 2.15	0.513
cond [SG] * sentenceType [Grammatical]	1.07	0.37 – 3.11	0.896
cond [SG] * verbType [Active]	0.75	0.28 – 2.01	0.571
sentenceType [Grammatical] * verbType [Active]	0.27	0.10 – 0.77	**0.015**
cond [SG] * SRT_z_score1	0.89	0.44 – 1.82	0.750
sentenceType [Grammatical] * SRT_z_score1	0.85	0.34 – 2.13	0.726
verbType [Active] * SRT_z_score1	1.15	0.51 – 2.60	0.734
(cond [SG] * sentenceType [Grammatical]) * verbType [Active]	5.77	1.39 – 23.89	**0.016**
(cond [SG] * sentenceType [Grammatical]) * SRT_z_score1	0.92	0.30 – 2.85	0.890
(cond [SG] * verbType [Active]) * SRT_z_score1	0.96	0.35 – 2.65	0.939
(sentenceType [Grammatical] * verbType [Active]) * SRT_z_score1	1.18	0.36 – 3.79	0.787
(cond [SG] * sentenceType [Grammatical] * verbType [Active]) * SRT_z_score1	0.64	0.15 – 2.76	0.548
Observations	2240		
Marginal R^2 / Conditional R^2	0.179 / 0.464		

Random effects: (1 + sentenceType + verbType + sentenceType:verbType | subject)

LLAMA B3

Due to a technical problem, we lacked LLAMA B3 scores for two participants. For remaining participants, there was no significant difference in mean LLAMA B3 scores between the two groups (*t*(62.769) = -0.472, *p* = 0.64). t = -0.47254). The distribution of scores is shown in Figure 11. Scores tended to deviate from normality, although this was only significant in the Control group (*W* = 0.909, *p* = 0.01) and not for the Surprisal group (*W* = 0.970, *p* = 0.42).

When adding z-transformed LLAMA B3 scores to the model for accuracy on the Day 3 structure test block (new verbs), all interactions were removed as they did not significantly improve model fit. The effects of Group and Condition were still observed, and the effect of LLAMA B3 score was not significant (Table 6). When adding LLAMA B3 scores to the model for item endorsement in the GJT, LLAMA B3 score was removed as a predictor during model selection as it had no significant effect on model fit, while the original Group x Verb Type x Grammaticality remained significant. We report the output of the initial model (with maximal fixed structure) for reference (Table 7).

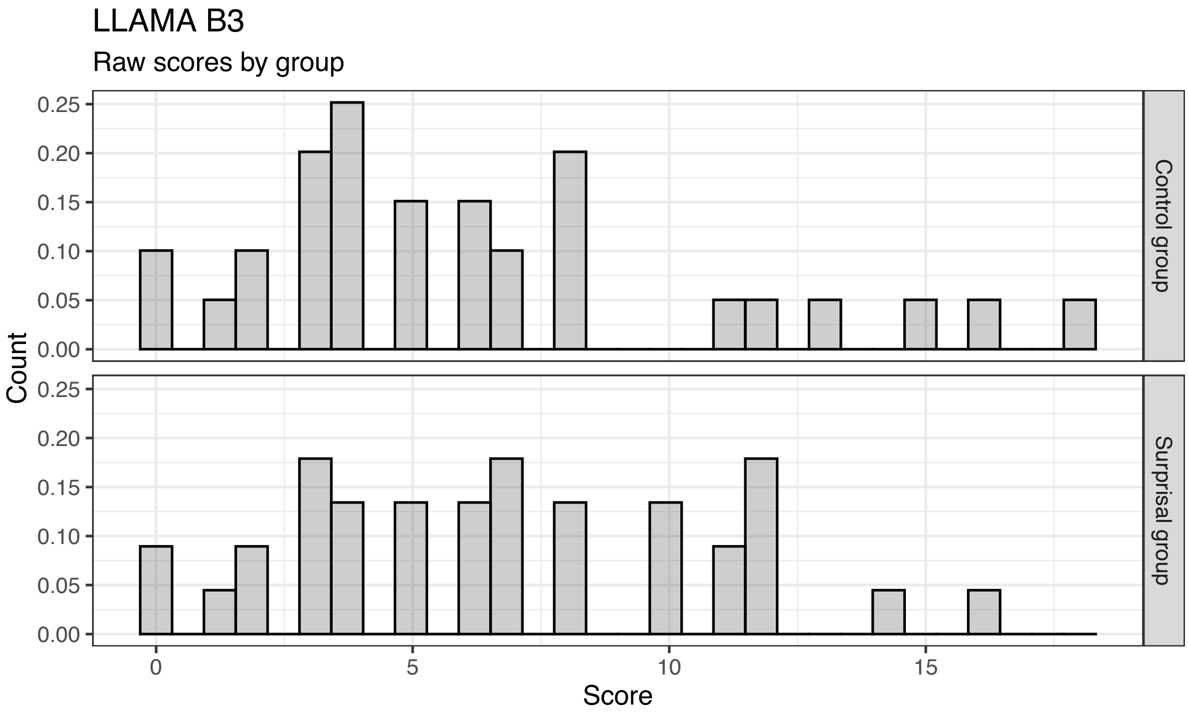


Figure 11. ***LLAMA B3 scores by group***

Table 6. ***Final model for accuracy on Day 3 structure test (new verbs) with LLAMA B3 score added to fixed effects structure***

	Accuracy		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.08	0.64 – 1.83	0.779
LLAMA B3 score	1.10	0.79 – 1.53	0.569
Structure (Active)	3.48	1.84 – 6.58	**<0.001**
Group (Surprisal)	2.45	1.30 – 4.63	**0.006**
Observations	1088		
Marginal R^2 / Conditional R^2	0.098 / 0.464		
Random effects: (1 + Structure | subject)			

Table 7. ***Initial model for endorsement in GJT, with LLAMA B3 score added to fixed effects structure***

	Endorsement		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	1.07	0.65 – 1.76	0.790
cond [SG]	0.52	0.26 – 1.05	0.067
sentenceType [Grammatical]	5.28	2.36 – 11.80	**<0.001**
verbType [Active]	4.11	1.99 – 8.47	**<0.001**
LLAMA_B3_z_score	0.99	0.62 – 1.60	0.979
cond [SG] * sentenceType [Grammatical]	1.02	0.35 – 2.98	0.971
cond [SG] * verbType [Active]	0.86	0.32 – 2.28	0.757
sentenceType [Grammatical] * verbType [Active]	0.30	0.11 – 0.88	**0.028**
cond [SG] *LLAMA_B3_z_score	0.57	0.28 – 1.15	0.116
sentenceType [Grammatical] * LLAMA_B3_z_score	0.77	0.37 – 1.61	0.489
verbType [Active] * LLAMA_B3_z_score	1.20	0.61 – 2.36	0.600
(cond [SG] * sentenceType [Grammatical]) * verbType [Active]	4.93	1.18 – 20.54	**0.029**
(cond [SG] * sentenceType [Grammatical]) * LLAMA_B3_z_score	2.64	0.90 – 7.70	0.076
(cond [SG] * verbType [Active]) * LLAMA_B3_z_score	1.17	0.44 – 3.14	0.753
(sentenceType [Grammatical] * verbType [Active]) * LLAMA_B3_z_score	0.76	0.29 – 1.97	0.574
(cond [SG] * sentenceType [Grammatical] * verbType [Active]) * LLAMA_B3_z_score	0.93	0.23 – 3.85	0.924
Observations	2176		
Marginal R^2 / Conditional R^2	0.185 / 0.466		

Random effects: (1 + sentenceType + verbType + sentenceType:verbType | subject)

Appendix S2 – Additional descriptive statistics

Cross-situational learning task

Day 1, accuracy by block:

Group	Block	Mean	Sd
Control	1	0.51	0.13
Control	2	0.59	0.19
Control	3	0.62	0.18
Control	4	0.71	0.19
Control	5	0.71	0.19
Control	6	0.70	0.23
Control	7	0.65	0.17
Surprisal	1	0.53	0.11
Surprisal	2	0.59	019
Surprisal	3	0.63	0.22
Surprisal	4	0.69	0.22
Surprisal	5	0.71	0.20
Surprisal	6	0.73	0.20
Surprisal	7	0.68	0.21

Day 2, accuracy by block:

Group	Block	Mean	Sd
Control	1	0.66	0.18
Control	2	0.73	0.18
Control	3	0.80	0.17
Control	4	0.72	0.22
Control	5	0.73	0.15
Control	6	0.77	0.16
Control	7	0.77	0.17
Control	8	0.78	0.15
Control	9	0.82	0.16
Control	10	0.61	0.23
Surprisal	1	0.70	0.21
Surprisal	2	0.80	0.19
Surprisal	3	0.82	0.16
Surprisal	4	0.79	0.19
Surprisal	5	0.75	0.14
Surprisal	6	0.79	0.20
Surprisal	7	0.80	0.16
Surprisal	8	0.80	0.16

Surprisal	9	0.83	0.16
Surprisal	10	0.71	0.15

Day 3, accuracy by block:

Group	Block	Mean	Sd
Control	1	0.84	0.15
Control	2	0.66	0.20
Control	3	0.61	0.18
Surprisal	1	0.85	0.14
Surprisal	2	0.71	0.26
Surprisal	3	0.74	0.20

Day 3, accuracy on structure test block (New verbs), by structure:

Group	Structure	Mean	Sd
Control	Passive	0.52	0.26
Control	Active	0.71	0.28
Surprisal	Passive	0.65	0.34
Surprisal	Active	0.84	0.20

Vocabulary tests, accuracy by vocabulary item type:

Group	Vocab type	Test	Mean	Sd
Control	Nountest	1	0.65	0.27
Control	Nountest	2	0.70	0.23
Control	Nountest	3	0.74	0.25
Control	Nountest	4	0.85	0.23
Control	Nountest	5	0.90	0.17
Control	Verbtest	1	0.65	0.17
Control	Verbtest	2	0.62	0.27
Control	Verbtest	3	0.72	0.26
Control	Verbtest	4	0.80	0.18
Control	Verbtest	5	0.80	0.21
Surprisal	Nountest	1	0.70	0.23
Surprisal	Nountest	2	0.75	0.28
Surprisal	Nountest	3	0.81	0.25
Surprisal	Nountest	4	0.85	0.19
Surprisal	Nountest	5	0.86	0.19
Surprisal	Verbtest	1	0.67	0.26
Surprisal	Verbtest	2	0.65	0.23
Surprisal	Verbtest	3	0.77	0.20

Surprisal	Verbtest	4	0.80	0.20
Surprisal	Verbtest	5	0.84	0.17

Grammaticality Judgment Task

Grammaticality Judgment Task, endorsement by grammaticality and structure (verb type):

Group	Sentence type	Error type	Verb type	Mean	Sd
Control	Ungrammatical	Actka	Active	0.75	0.26
Control	Ungrammatical	Passnoka	Passive	0.51	0.28
Control	Grammatical	None	Passive	0.79	0.23
Control	Grammatical	None	Active	0.84	0.16
Surprisal	Ungrammatical	Actka	Active	0.61	0.35
Surprisal	Ungrammatical	Passnoka	Passive	0.39	0.29
Surprisal	Grammatical	None	Passive	0.69	0.26
Surprisal	Grammatical	None	Active	0.91	0.15

Grammaticality Judgment Task, *d'* scores (sensitivity to grammaticality) by structure (verb type):

Group	Verb type	Mean	Sd
Control	Active	0.44	1.24
Control	Passive	0.07	1.34
Surprisal	Active	0.33	1.37
Surprisal	Passive	0.01	1.54

Appendix S3 – Additional figures (Day 2)

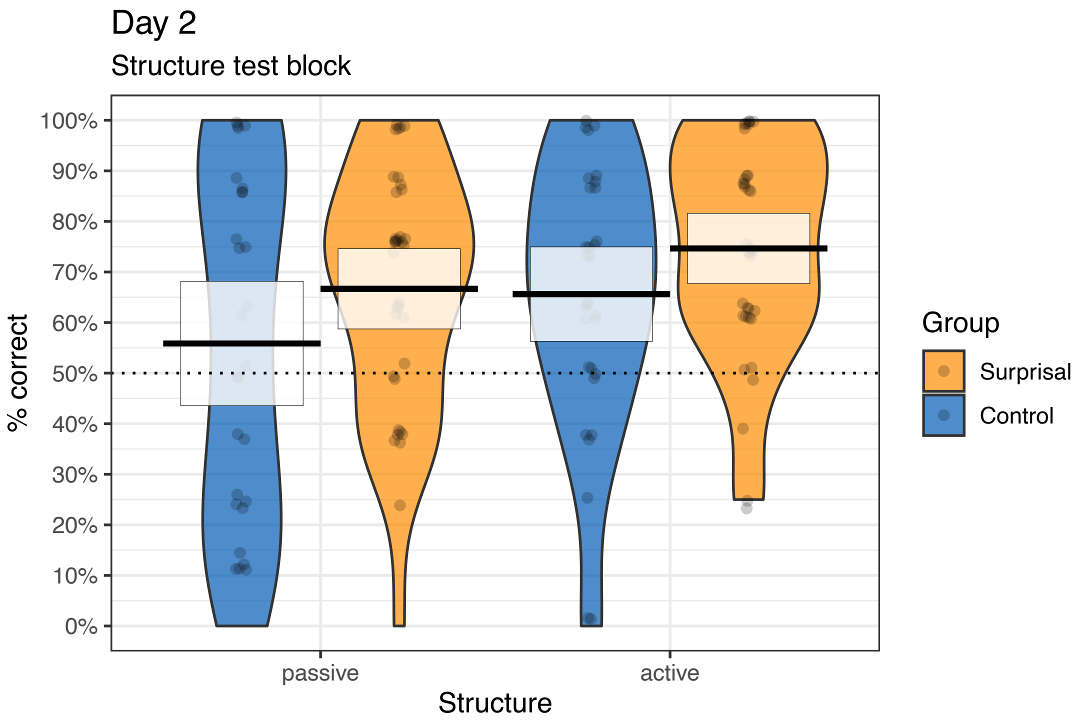


Figure 12. ***Accuracy on Day 2 structure test block, group means. Horizontal bars represent group means, shaded rectangles 95% CIs.***

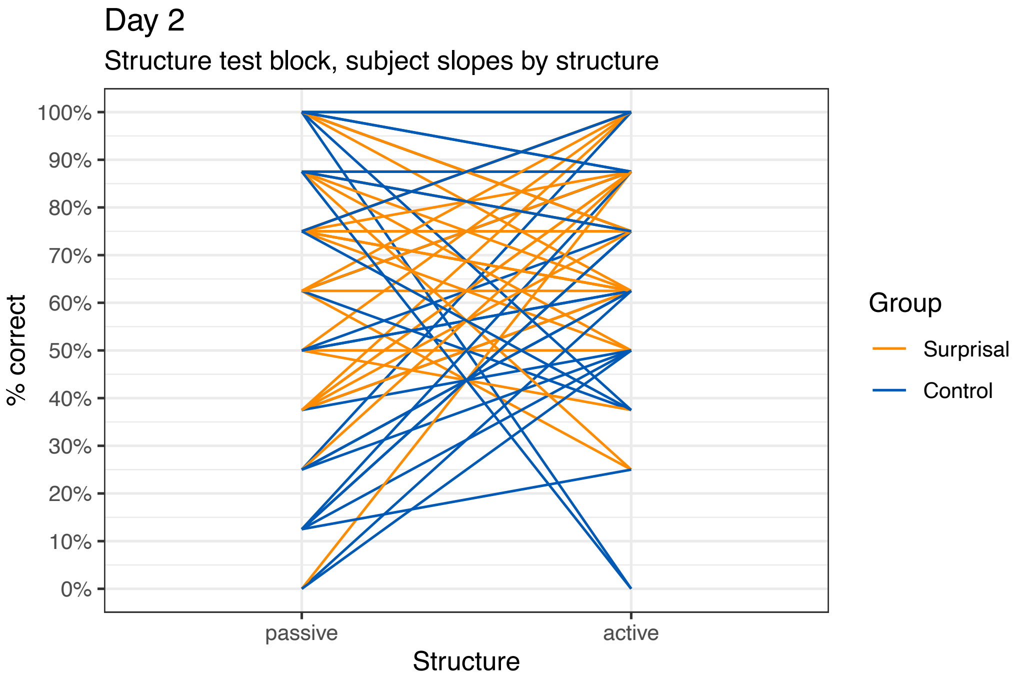


Figure 13. ***Accuracy on Day 2 structure test block by subject. Horizontal bars represent group means, shaded rectangles 95% CIs.***

Appendix S4 – Details of Day 3 Structure Test block (Old verbs)

The Structure Test (Old verbs) block on Day 3 was affected by a counterbalancing problem, which meant roughly half of participants saw pictures described with the same structure they had encountered them with in the Day 2 Structure Test block, while the other half saw the pictures described with the other structure. This distinction was orthogonal to Group, although participants in the Surprisal group were numerically more likely to be exposed to the opposite structure, compared to the Control group (Table 1).

Table 8. ***Structure counterbalancing between Day 2 and Day 3 Structure Tests (Old verbs)***

	Structure encountered on Day 3	
	Same as Day 2	Different
Control group	19	15
Surprisal	15	21

Figure 1 shows the mean accuracy scores obtained by participants in the Structure Test (Old verbs) block on Day 3, broken down by whether the item had been seen in the Day 2 structure test block with the same structure. Figure 2 shows the same data, further broken down by whether participants had answered correctly (i.e., picked the correct structural interpretation) to the item on Day 2.

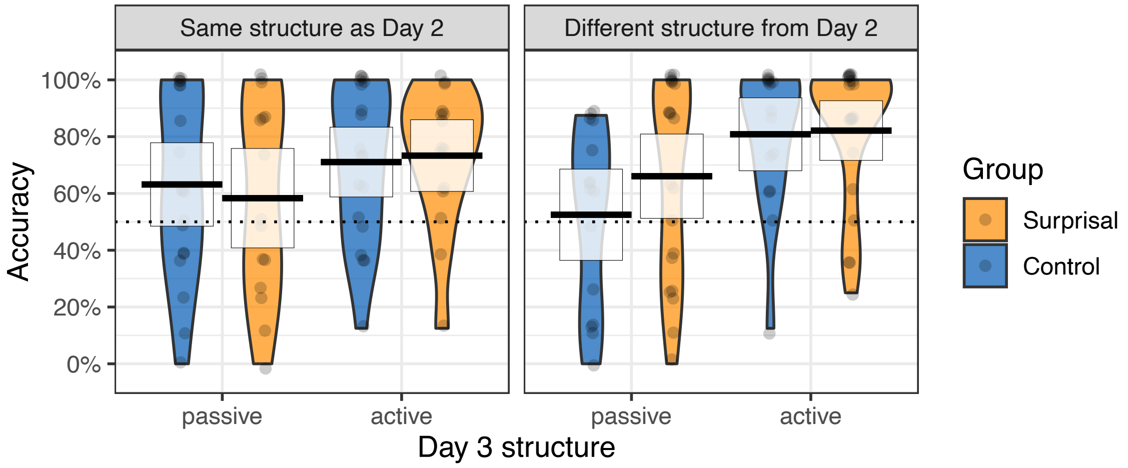


Figure 14. ***Mean accuracy on Day 3 Structure Test (Old verbs), divided by whether items used the same structure as in the Day 2 Structure Test.***

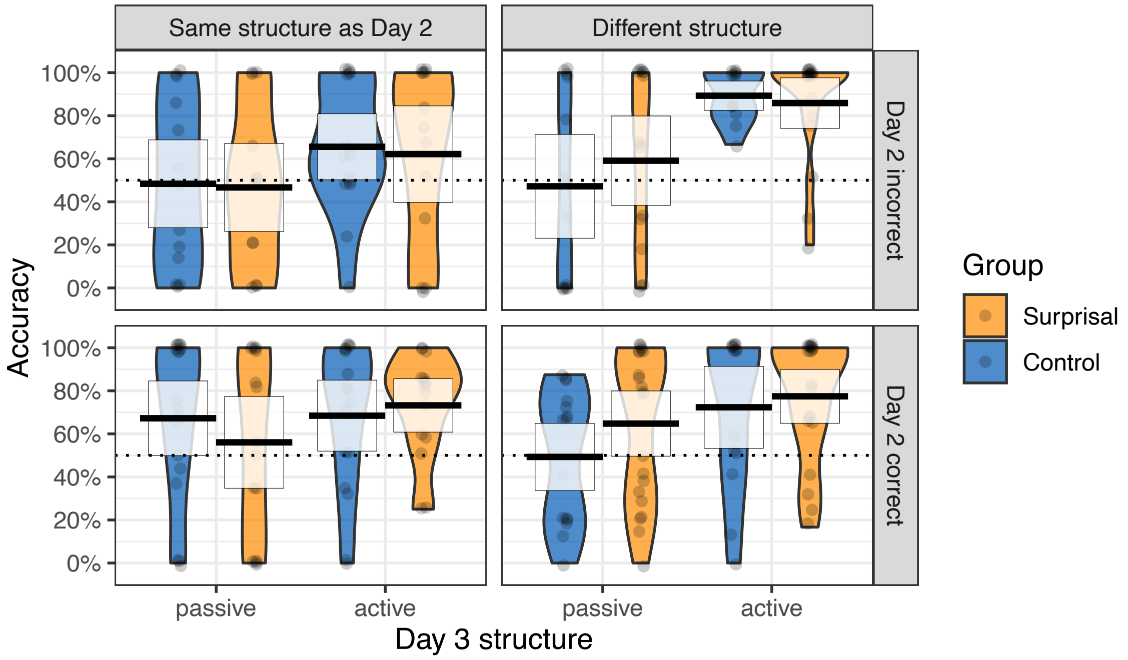


Figure 15. ***Mean accuracy on Day 3 Structure Test (Old verbs), divided by whether items used the same structure as in the Day 2 Structure Test, and by response accuracy to those items on Day 2.***

Appendix S5 – Final statistical models for all tests

Day 1, learning blocks

		Accuracy		
Predictors		*Odds Ratios*	*95% CI*	*p*
(Intercept)		2.36	1.90 – 2.94	**<0.001**
Block		1.28	1.21 – 1.36	**<0.001**
Observations		6720		
Marginal R^2 / Conditional R^2		0.043 / 0.221		
Random effects: (1 + Block ǀ Subject)				
Groups			*SD*	
Subject	(Intercept)		0.90	
	Block		0.21	

Day 2, learning blocks

		Accuracy		
Predictors		*Odds Ratios*	*95% CI*	*p*
(Intercept)		6.90	5.34 – 8.90	**<0.001**
Block		1.12	1.06 – 1.18	**<0.001**
Observations		5600		
Marginal R^2 / Conditional R^2		0.013 / 0.256		
Random effects: (1 + Block ǀ Subject)				
Groups			*SD*	
Subject	(Intercept)		1.01	
	Block		0.13	

Vocabulary test blocks: Verbs

		Accuracy	
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.45	1.08 – 1.95	**0.013**
Test	1.33	1.24 – 1.41	**<0.001**
Observations	2800		
Marginal R^2 / Conditional R^2	0.037 / 0.230		
Random effects: (1 + Subject)			
Group		*SD*	
Subject (Intercept)		0.91	

Vocabulary test block: Nouns

		Accuracy	
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.43	0.84 – 2.45	0.186
Test	1.69	1.51 – 1.88	**<0.001**
Group (Surprisal)	1.48	0.74 – 2.96	0.269
Test x Group (Surprisal)	0.80	0.69 – 0.93	**0.004**
Observations	2800		
Marginal R^2 / Conditional R^2	0.068 / 0.405		
Random effects: (1 + Subject)			
Group		*SD*	
Subject (Intercept)		1.36	

Day 2, Structure Test trials after feedback trials

		Accuracy	
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.43	1.11 – 1.85	**0.005**
Trial	1.40	1.17 – 1.67	**<0.001**
Observations	1120		
Marginal R^2 / Conditional R^2	0.024 / 0.265		
Random effects: (1 + Trial | Subject)			
Groups		*SD*	
Subject (Intercept)		0.92	
Trial		0.47	

Day 2, Structure Test block

		Accuracy	
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	2.03	1.34 – 3.07	**0.001**
Structure (Active)	1.44	0.86 – 2.39	0.162
Observations	1120		
Marginal R^2 / Conditional R^2	0.006 / 0.351		
Random effects: (1 + Structure | Subject)			
Groups		*SD*	
Subject (Intercept)		1.50	
Structure		1.72	

Day 3, Structure Test block (Old verbs)

	Accuracy		
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	2.15	1.28 – 3.61	**0.004**
Structure (Active)	2.80	1.50 – 5.23	**0.001**
Observations	1120		
Marginal R^2 / Conditional R^2	0.040 / 0.506		
Random effects: (1 + Structure l Subject)			
Groups		*SD*	
Subject (Intercept)		1.89	
Structure		1.87	

Day 3, Structure Test block (New verbs)

	Accuracy		
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.25	0.82 – 1.90	0.303
Structure (Active)	2.82	1.42 – 5.59	**0.003**
Group (Surprisal)	2.50	1.25 – 4.99	**0.009**
Observations	1120		
Marginal R^2 / Conditional R^2	0.073 / 0.496		
Random effects: (1 + Structure + Group l Subject)			
Groups		*SD*	
Subject (Intercept)		1.12	
Structure		2.30	
Group		1.59	

Grammaticality Judgment Task: Endorsement

	Endorsement		
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.05	0.66 – 1.67	0.845
Sentence Type (Grammatical)	4.45	2.19 – 9.03	**<0.001**
Verb Type (Active)	3.68	2.45 – 5.52	**<0.001**
Group (Surprisal)	0.52	0.27 – 1.00	0.050
Sentence Type (Grammatical) x Group (Surprisal)	1.15	0.43 – 3.05	0.781
Verb Type (Active) x Group (Surprisal)	0.85	0.49 – 1.48	0.562
Sentence Type (Grammatical) x Verb Type (Active)	0.40	0.22 – 0.73	**0.003**
Sentence Type (Grammatical) x Verb Type (Active) x Group (Surprisal)	4.44	1.85 – 10.64	**0.001**
Observations	2240		
Marginal R^2 / Conditional R^2	0.184 / 0.396		
Random effects: (1 + Sentence Type | Subject)			
Groups		*SD*	
Subject (Intercept)		1.14	
Sentence Type		1.70	

Grammaticality Judgment Task: *d'*

Effects	*DFn*	*DFd*	*SSn*	*SSd*	*Generalised η^2*	*F*	*p*
(Intercept)	1	68	0	169.677	0	0	1
Group	1	68	4.514	169.677	0.017	1.809	0.183
Verb Type	1	68	0.301	89.125	0.001	0.230	0.633
Group x Verb Type	1	68	6.016	89.125	0.023	4.590	**0.036**

Debriefing questionnaire

	Awareness		
Predictors	*Odds Ratios*	*95% CI*	*p*
(Intercept)	1.43	0.73 – 2.89	0.306
Group (Surprisal)	0.50	0.19 – 1.28	0.153
Observations	70		
R^2 Tjur	0.029		

License

Children’s language abilities predict success in remote communication contexts

Karla K. McGregor
Ronald Pomper
Nichole Eden
Timothy Arbisi-Kelm
Nancy Ohlmann
Shivani Gajre
Erin Smolak
Boys Town National Research Hospital, USA

Abstract: Remote communicative contexts are part of everyday social, familial, and academic interactions for the modern child**.** We investigated the ability of second-graders to engage in remote discourse, and we determined whether language ability, theory of mind, and shy temperament predicted their success. Fifty 7-to-9-year-old monolingual English speakers with a wide range of language abilities participated in standardized testing and an expository discourse task in which they taught two adults to solve the Tower of London, one in an audiovisual condition to simulate video chat and a second in an audio-only condition to simulate phone communication. The discourse was scored with a rubric of 15 items deemed relevant to the explanation. Children included 27% to 87% of the items, with more items communicated via gesture than spoken word in both conditions. Gesture scores and spoken scores were highly correlated. Children specified more rubric items overall in the audio condition and more rubric items in the spoken modality when in the audio condition than the audiovisual condition. Performance in both conditions was positively associated with scores on independent measures of language ability. There was no relationship between performance and theory of mind, shy temperament, ability to solve the Tower of London, age, or sex. We conclude that 7-to-9-year-olds adjust the modality and content of their message to suit their remote partner's needs, but their success in remote discourse contexts varies significantly from individual to individual. Children with below-average language skills are at risk for functional impairments in remote communication.

Keywords: remote communication; pragmatics; gesture; phone; video chat

Corresponding author(s): Karla McGregor, Center for Childhood Deafness, Language & Learning, Boys Town National Research Hospital, 425 N. 30th St., Omaha, NE, 68131, USA. Email: karla.mcgregor@boystown.org.

ORCID ID(s): https://orcid.org/0000-0003-0612-0057

Citation: McGregor, K.K., Pomper, R., Eden, N., Arbisi-Kelm, T., Ohlmann, N., Gajre, S., & Smolak, E. (2021). Children’s language abilities predict success in remote communication contexts. *Language Development Research,* 1(1), p 245-282. DOI: 10.34842/8jgf-r802

Introduction

Children modify their communication according to audience and context (Akhtar et al., 1996; Mori & Cigala, 2016; Nadig & Sedivy, 2002; Nilsen & Fecica, 2011; Shatz & Gelman, 1973). For example, four-year-olds speak in shorter sentences and use more utterances to draw the listener's attention when addressing toddlers than peers or adults (Shatz & Gelman, 1973). However, this early emerging skill must be honed over developmental time as children encounter new contexts. In this paper, we considered remote communication, a context that is prominent in the everyday lives of children (McClure et al., 2015; MPFS, 2018) and particularly so since the first quarter of 2020 when many families were quarantined in response to the COVID19 pandemic (Koeze & Popper, 2020). We investigated second-graders' ability to engage in discourse via an audio-only channel to simulate phone communication and via an audiovisual channel to simulate video chat, and we determined whether language ability, theory of mind, and shy temperament predicted their success.

Children’s Phone and Video Communications

From birth, children have access to various technologies with which they learn to share meaning with others (Erstad et al., 2020). Children’s phone talk has long been recognized as a developmental step between the ‘here and now’ audible and instantaneous exchange characteristic of face-to-face talk and the decontextualized nature of written communication (Cameron & Lee, 1997; Gillen, 2002). Children talk on the phone long before they can read and write yet, to do so successfully, they must heed the needs of listeners who do not share their immediate context, just as the successful writer must.

Children’s phone communication skills emerge early but follow a protracted developmental course. Take, for example, a study by Cameron and Lee (1997). They had three to eight-year-olds solve the Tower of London (Shallice, 1982), a task in which the child must move beads, one at a time, onto pegs to match an array demonstrated by the examiner in as few moves as possible. Afterward, they asked the children to explain the task to listeners face-to-face or on the phone. The older children gave more complete directions than the younger children. Children of all ages provided more detail and used more specific vocabulary while speaking on the phone than in person (Cameron & Lee, 1997). However, their communication was not necessarily worse when in person. The children used frequent visual checking of the listener's performance, presumably to adapt the directions to the listener's needs; what they accomplished with shared visual contexts when face-to-face, they accomplished with more extensive, detailed instructions when on the phone.

An early ability to adjust to a remote listener's needs is also evident in narrative discourse. For example, Pinto et al. (2016) found that 5-to7-year-olds make more mention of characters' mental states when narrating a pictured event during phone talk than when face-to-face, presumably because they realize that the listener cannot intuit the characters' mental states without seeing the pictures.

Children's communication on video-chat platforms has been less extensively researched than their phone talk. However, here too, we find very young children demonstrating some success. For example, McClure et al. (2018) observed families with babies ages 6 to 24 months as they engaged in video chats with the babies' grandparents. They were particularly interested in times when the baby initiated joint attention with the grandparent across the screen by, for example, showing a toy. Whereas only 8% of the babies under 15 months ever did so, 46% of the babies older than 15 months engaged their grandparents in this way. This extremely early performance reflects an essential difference between video chat and phone talk, shared visual context.

In complicated communicative exchanges, even adults may benefit from shared visual context. For example, Veinott et al. (1999) asked native and non-native English speakers to explain a mapped route to a listener who joined them via video or audio-only. Although the native-speaking pairs were equally successful in the video and audio conditions, the non-native speakers benefitted from the video context. Specifically, they were faster and more accurate at communicating the route in the video condition. This success was engendered by more talk devoted to instruction, more checks on mutual understanding, and more frequent gesturing. However, gesture was not considered in the Cameron and Lee (1997) comparison of children's phone and in-person communication; therefore, we do not know the extent to which children's communication might be enhanced by gesture in remote communication contexts that involve shared visual access.

The Development of Discourse in the Gesture and Speech Modalities

Gesture and speech are integrated systems of communication. They work together to convey meaning and affect, and they do so in ways that benefit both the listener and the speaker. When compared to speech alone, speech plus gesture reliably enhance listeners' comprehension. The benefits are greater for gestures that convey actions (e.g., how to do something) rather than abstract concepts (e.g., how something feels) and for gestures that complement rather than reproduce the meanings conveyed by words (Hostetter, 2011). In addition, children benefit from gestured input more than adults (Hostetter, 2011). Not surprisingly then, children tend to learn more readily from teachers who frequently gesture in ways that accurately convey new concepts than from teachers who do not (Alibali et al., 2013; Ovendale et al., 2018).

For the speaker, gesture assists with memory and planning the information to be conveyed (Alibali et al., 2000). Moreover, gesture supports thought. Under high cognitive load, like solving a problem, counting, or reasoning out load, children's gestures reflect their understanding (Alibali & Goldin-Meadow, 1993; Broaders, et al., 2007; Church & Goldin-Meadow, 1986; Ehrlich et al., 2006; Garber & Goldin-Meadow, 2002; Göksun et al., 2010; Pine, et al., 2004). When on the cusp of understanding, children often convey a more accurate grasp with their gestures than with their words, and this gesture-speech mismatch indicates learning readiness (Alibali & Goldin-Meadow, 1993).

Children communicate with gestures before they speak their first word (Capone & McGregor, 2004). However, gesture, like spoken language, continues to develop well into the school years. Alamillo et al. (2013) observed six- and 10-year olds during monologic narrative and dialogic explanation tasks. The older children used not only more complex spoken language but also more complex gestures. Both groups of children used more gestures during dialogue than monologue, suggesting some awareness of their partner's needs during the dialogic exchange.

Roth (2002) argues that gesture provides a stepping stone along the path of discourse development and in the expression of new knowledge during discourse. He observed tenth-grade students over multiple physics lessons as they conducted hands-on experiments and then explained their findings. The result was a robust developmental progression. First, the students spoke as they repeated their actions on the objects used in the experiments and, soon after, as they simulated the actions on other arbitrary objects. Gradually, they supplemented their spoken explanations with gestures produced without the support of objects. In these early attempts, the students tended to be more accurate in their actions on objects and their gestures in general than in their spoken explanations. Finally, the students arrived at a mature understanding of the physics problems they expressed in written or spoken words. At this mastery stage, spoken language and gesture continued to co-occur, although the gesture frequency was lower than in the more novice stages. In this way, the developmental course of these much older children faced with a new and complex task recapitulated the early communicative development of toddlers. Both demonstrate a progression from actions on objects to gesture to speech while never abandoning any of these highly functional modalities (Capone & McGregor, 2004).

The Contribution of Language Ability, Theory of Mind, and Shy Temperament to Successful Remote Communication

The success of any communicative interaction and the extent to which speakers can adapt their gestured and spoken communication to contextual demands will depend, in part, on the language abilities of the interlocutors. When communicating on the phone, the lack of shared visual context means that the speaker must provide information via words rather than gestures and that the words and their organization must be specific enough to enable comprehension. In these decontextualized exchanges, language *becomes* the context, and thus the communication partners must have the lexical, syntactic, and discourse skills necessary for creating clarity and common ground (Uccelli et al., 2019). When communicating via video chat, the shared visual context lessens these demands. However, challenges remain relative to face-to-face communication where the extent of the shared visual context is greater, the signal-to-noise ratio is higher, and physical interaction is viable.

Another factor influencing the success of communicative exchange is the interlocutors' ability to engage in theory of mind (Hughes & Leekham, 2004; Miller, 2006). Theory of mind refers to perceiving one's own and others' emotions, beliefs, desires, thoughts, and knowledge systems. Theory of mind develops from early childhood well into adolescence (Dorval et al., 1984). A speaker with a mature theory of mind will recognize the listener's need for more or less information.

Nevertheless, remote communication may present some challenges to perspective-taking, even for those with a strong theory of mind. Keeping track of the listener's perspective without a shared visual context, as during phone conversations, may impose a high memory load (Zhao et al., 2018). Moreover, primate work demonstrates that face, hand, and body movements provide essential cues to accurate social perception (Allison et al., 2000). Adult humans will even infer mental states from the movement of animated abstract shapes (Castelli et al., 2002). During remote communication, movements that cue the extent of the listener's understanding are limited (in video chat) or missing (in phone interactions).

The temperament of the interlocutors may influence the success of communication as well. Temperament is a stable trait that is highly heritable (Buss & Plomin, 1984; see Henderson & Wachs, 2007, for a review). Children who possess a shy temperament are inclined to withdraw from social interaction, particularly in unfamiliar social situations (Schmidt & Tasker, 2000). Shy children are reticent to talk; they talk less, make fewer spontaneous remarks, and are more likely to be unresponsive to strangers than their outgoing peers (Prior et al., 2000; Smith Watts et al., 2014). Thus, the high verbal demands of remote communication contexts may be especially challenging for shy children. That said, remote communication may be more comfortable than face-to-face communication for some shy children, in which case they may be less reticent than documented in previous (in-person) research.

Although they are separate constructs, language ability, theory of mind, and shy temperament are interrelated. The relation between language and theory of mind is well studied. A meta-analysis of 104 studies with English-speaking children below age seven revealed a significant positive relationship between language and theory of mind abilities (Milligan et al., 2007). For example, vocabulary and grammar skills at two (Farrar & Maag, 2002) and grammar skills at three (Astington & Jenkins, 1999) predict theory of mind skills two years later. Language and shyness are also related. Specifically, shy children tend to score lower on formal tests of pragmatics (Copelan & Weeks, 2009), receptive vocabulary, and phonemic awareness (Spere et al., 2004) than their more outgoing peers, although not so low to be of clinical concern.

In contrast, the relation between shyness and theory of mind is not fully understood. Some investigators find that shy children perform more poorly on theory of mind tests than outgoing children (Banerjee & Henderson, 2001; DeRosnay et al., 2014; Walker, 2005), leading them to hypothesize that a shy temperament limits their social-communicative interactions, and thus their opportunities to learn more about reading another's mind. On the other hand, others report that shy children demonstrate an advantage on theory of mind tests (Mink et al., 2014; Wellman et al., 2011), leading them to hypothesize that shy children sharpen their theory of mind by observing others' social-communicative interactions.

The Current Study

In the current study, we examined the expository discourse of second graders in two remote communication contexts, one that simulated phone communication by providing an audio channel only and the other that simulated video chat communication by including both audio and visual channels. We were particularly interested in second graders because they are in the throes of language and theory of mind development and, by second grade, their temperament is highly stable (Neppl et al., 2010). Moreover, they are still early in their formal reading and writing instruction years, a time when individual differences in bridging the fully contextualized nature of face-to-face talk and the fully decontextualized nature of formal writing are likely to be high.

Expository discourse typically involves more complex syntax and more specific or sophisticated vocabulary choices than conversational discourse. We selected an expository discourse task because it is more likely than conversational exchange to reveal individual differences between children. For example, adolescents with language impairments do not differ significantly from their typical age-mates in conversation but, during exposition, they tend to use shorter, less syntactically complex sentences (Nippold et al., 2008). Expository discourse is also high in ecological validity given that mastery of expository discourse is recommended as an instructional goal in the academic curriculum (CCSS; Common Core State Standards Initiative, 2015), and it is the

type of discourse required for everyday communicative goals such as giving directions or explaining the rules of a game (Lundine & McCauley, 2016).

Like Cameron and Lee (1997), we used the Tower of London to elicit the expository discourse. The task requires problem-solving and planning (visualizing several moves ahead) and other aspects of executive function such as attention, memory, and inhibition. Language, either internalized or externalized, is helpful for scaffolding performance on the Tower of London. Perhaps, as a result, children with developmental language disorder tend to perform poorly on the task (Larson et al., 2019; Marton, 2008; Roello et al., 2015, and verbal suppression impairs performance in children with and without language disorder (Lidstone et al., 2012). After completing the Tower of London, we asked the children to explain the game to two naive adults, one who was not present but could hear them (audio condition) and one who was not present but could hear and see them (audiovisual condition).

Questions and Hypotheses

We preregistered the study (McGregor et al., 2019, available at this link: OSF Registries | Children's Voabulary Project; Remote Communication) as a comparison between second graders with and without developmental language disorder, a prevalent neurodevelopmental condition characterized by limitations in language learning, comprehension, and use. Unfortunately, because of the COVID-19 pandemic, we were forced to close the study before recruiting enough participants with developmental language disorder. Nevertheless, we had an excellent distribution of language abilities represented in the sample and adequate power to investigate language ability as a continuous predictor of remote communication performance. Thus, we modified our predictions to be:

Children would provide more complete directions in the audiovisual than audio condition because they would more frequently supplement their verbal message with gestures in the latter than the former.

There would be individual differences across children such that those who have more robust vocabularies, who are less shy, and who have a more highly developed theory of mind would be more successful on the task than those who scored lower in these domains. We also determined the effect of age, sex, and success on the Tower of London itself.

Finally, we took this opportunity to explore the relationships between vocabulary, theory of mind, and shyness to address incomplete or conflicting reports in the literature.

Methods

Participants

The project was conducted in compliance with protocols approved by the Internal Review Board of Boys Town National Research Hospital to ensure the protection of human subjects. Participants were 50 second graders (29 girls), ages 7 to 9 years (median = 100 months, min-max = 88 to 109). Two additional children were tested but are not included here, one because of a subsequently diagnosed seizure disorder and the other because of attrition.

All participants were monolingual English speakers from Iowa or Nebraska in the United States recruited from a larger longitudinal study of language development (Research Registry 3425, 2017). According to parents' reports of ethnicity, one participant was Hispanic or Latino, 41 were neither Hispanic nor Latino. Eight parents did not report ethnicity. According to parents' reports of race, one participant was African American, 43 were Caucasian, and six were more than one race.

The children presented with a range of spoken language abilities, with standard scores from 72 to 127 on the *Test of Narrative Language-second edition* (TNL-2, Gillam & Pearson, 2017). Eleven were receiving special support for language in or outside of school. To ensure that neither intellectual disability nor hearing loss contributed to variability in task performance, we limited enrollment to participants who earned a perceptual index score of 70 or higher on the *Wechsler Abbreviated Scale of Intelligence* (Wechsler, 1999) and passed a pure tone audiometric screening.

Procedure

We administered standardized tests to determine the abilities that predict expository discourse performance, and then we administered the expository discourse task itself. Data collection occurred over two or three sessions scheduled within two weeks.

NIH Toolbox Picture Vocabulary Test

The *NIH Toolbox Picture Vocabulary Test* (Gershon et al., 2013) measures receptive single-word vocabulary. The participant is instructed to touch the image in a 4-alternative forced-choice array that they believe is most closely associated with the word they heard. The difficulty level of each trial is automatically adjusted by the software program, contingent on the participant's previous response's accuracy. Raw scores on this task were converted to normally distributed standard scores (scaled scores) that were not age-corrected. Specifically, the raw scores were ranked and then transformed to create a standard normal distribution, which was then re-scaled to have a mean of 10 and a standard deviation of 3.

Temperament in Middle Childhood Questionnaire (TMCQ)

The TMCQ (Simonds & Rothbart, 2006) taps caregiver judgments of the emotional temperament of children between 7 and 10 years of age. Although the TMCQ measures a broad range of temperament traits, we only used the items that tapped shyness. The temperament dimension "shyness" is operationally defined in the TMCQ as "slow or inhibited approach in situations involving novelty uncertainty." The shyness score was calculated by summing the ratings for the five questions categorized within the shyness temperament dimension.

Theory of Mind Inventory-2 (ToMI-2)

On the ToMI-2 (Hutchins & Prelock, 2016), caregivers mark along a 5-point continuum ranging from “Definitely Not” to “Definitely” to describe the most likely way their child would mentalize in 60 different scenarios. The questionnaire is scored by plotting each response along a 20-centimeter scale and rounding the score to the nearest tenth. In the present analysis, these scores were then summed and divided by 60 to derive a mean score ranging from 1 to 20.

Tower of London (ToL)

The Tower of London Drexel University -2nd Edition (Culbertson & Zillmer, 2005) measures problem-solving and planning. In the ToL, both the children and the examiner use boards containing three pegs of varied sizes holding one to three colored beads. The examiner demonstrates beads stacked in 10 different arrangements on wooden pegs. The goal is for the child to move their beads one at a time and in as few moves as possible from a start position to match the examiner’s array.

We administered the task according to the directions in the test manual. The critical directions given to the children were: 1) the two pegboards must be alike, 2) as few moves as possible must be used to copy the design on the examiner's board, 3) no peg may contain more beads than it can hold, and 4) only one bead at a time can be moved, in other words, two or more beads cannot be taken off the board at one time. The score we derived was the number of moves needed to solve each. Four of the participants skipped one (N = 3) or 5 (N = 1) items on the ToL due to experimenter error or participant fatigue; therefore, we transformed scores by dividing the total number of extra moves (total moves – minimum moves) by the minimum number of moves for each child. A child who completed each arrangement in the minimum number of moves would have a proportion of 0, whereas a child who completed each arrangement with twice the number of the minimum possible moves would have a proportion of 1.

Expository Discourse

After completing the ToL, we asked the child to explain its procedures and five example problems to two naive adults who were not present, one who could hear them (audio condition), and one who could hear and see them (audiovisual condition). By simulating phone and video chat rather than engaging the children in these actual contexts, we ensured that the performance was the child's own, not the result of more or less scaffolding from a communicative partner. The exact instructions are included in the Appendix. All children participated in both conditions with order counterbalanced across participants.

In the audio condition, we showed the child a photograph of an unfamiliar adult woman. We said that she did not have access to video technology but would hear the child's instructions when she called later. A phone was included in the photograph. The children were then asked to explain to their listeners what the game looks like, what the rules are, and exactly how to play.

In the audiovisual condition, we showed the child a photograph of a second unfamiliar adult woman. We said that she had access to a computer (a computer was included in the photo), so she would hear and see the child's instructions when she logged in later.

The entirety of the data collection session was video recorded via a laptop camera for later scoring. For the audiovisual condition only, we also recorded with a camera on a tripod to illustrate more clearly to the child that their remote partner would be able to not only hear them but also see them. An example of a child participating in the aduio and audiovisual conditions is available at OSF | Example of child completing the discourse task.

Discourse Scoring

A 15-item rubric was created to capture the pragmatic and semantic content of each child's discourse (see Appendix). The child could receive one point for each of the first 11 items in the rubric, regardless of whether they expressed that item in gesture, spoken words, or both. Gestures could be representational (e.g., making a circle shape to indicate a bead), deictic (e.g., pointing or showing), or demonstration (e.g., moving the bead from one peg to the next). The items scored were: item 1, introducing the discourse (e.g., saying or gesturing hello); items 2 – 5, explaining each of four rules well enough for a naive listener to apply the rule successfully; items 6 – 10, explaining each of five trials well enough for a naive listener to complete the trial successfully; and item 11, closing the discourse (e.g., saying or gesturing goodbye, we're done). In addition, we were interested in the children's use of vocabulary deemed essential to the explanation. Thus, they could also earn one point for each of four spoken vocabulary items. These were at least one mention of 1) the bead/ball, 2) the peg/stick/stand, 3) the location (e.g., here, long peg), and 4) the sequential order (e.g., next, last). A second coder independently scored 22% of the discourse samples. The point-to-point agreement was 91.94%.

When using the overall score as a dependent variable, the highest possible score was 15: 11 spoken and/or gestured items + 4 spoken words. Because some children skipped items due to experimenter error or participant fatigue (in the audio condition, 10 participants skipped one item and one participant skipped two; in the audiovisual condition, six participants skipped one item), we transformed scores into proportions (points received/total maximum points possible given items administered) and applied a logit transformation.

Statistical Analysis

The preregistered data analysis plan was to use a linear mixed-effects model with the explanation score as the dependent variable and independent variables including a fixed within-subjects factor of condition (audio or audiovisual) and between-subjects effects of diagnosis (DLD or TD), sex (M, F), age, diagnosis x sex, and scores from the ToL, vocabulary, theory of mind, and shy temperament assessments. The random-effects structure was specified as a random intercept for subject. In the modified version presented here, we ran the model without the effects of diagnosis and the interaction between diagnosis and sex. Before conducting the analysis, we simulated 1000 datasets to determine the power of the study using a random intercept mixed model with a 2-factor within-group variable (condition was a 2-factor within-group variable), with 50 total participants and an intraclass correlation of 0.33. We found approximately 89% power to detect the difference between conditions with an effect size of 0.50, a moderate effect.

We also ran two exploratory models. To anticipate, we had predicted better scores in the audiovisual condition but, instead, obtained better scores in the audio condition. To explore this finding, we split the omnibus score into one for the gestured modality (maximum possible = 11) and one for the spoken modality (maximum possible = 11). We then ran a linear mixed model that included the original model variables plus the additional vocabulary x condition, modality, and modality x condition variables.

In a second exploration, we asked whether overall language ability predicted the expository discourse score. In effect, this is the same question we posed in the registered version of the project but abandoned because we were unable to recruit a sizeable cohort of children with developmental language disorder before COVID-19. Instead of including the diagnostic category—language disorder or typical language development—in the model, we regressed children's average explanation score (across audio and audiovisual conditions) on their scores on the TNL-2. The TNL-2 was used to group children into DLD or TD categories in the larger longitudinal project. There is evidence that developmental language disorder is a spectrum condition, not a categorical one (Lancaster & Camarata, 2019; Dollaghan, 2004); thus, considering the scores of children who potentially have developmental language disorder on a continuum with those of children who have typical language development is a valid approach.

Finally, we ran a confirmatory model, retesting our primary hypotheses with a linear model.

Results

Descriptive Data

The children's performance on the measures that served as independent variables in the statistical models appears in Table 1. Note from the min-max information that there was a reasonable range of scores on all measures for use in the statistical models. The exact distributions are plotted in the Supplemental Materials (Figures S8 through S18) available at this link OSF | Children's Vocabulary Project; Remote Communication).

Table 1. ***Summary statistics for scores that serve as predictors of expository discourse performance***

Construct	Measure	Score	Mean (sd)	Median	Min-Max
Receptive & Expressive Language	TNL-2	Omnibus Standard Score	104.74 (14.96)	108	72-127
Receptive Vocabulary	NIH PVT	Uncorrected Standard Score	76.76 (6.97)	76.5	60-89
Planning & Problem Solving	ToL	Proportion Extra Moves Score	0.91 (0.32)	0.93	0.14-1.71
Theory of Mind	ToMI-2	Composite Mean	16.95 (1.94)	16.84	12.42-19.89
Shyness	TMCQ	Shyness Total	13.28 (3.41)	13	5-20

Note: TNL-2 = *Test of Narrative Language-2^nd^ edition,* NIH PVT = *NIH Toolbox Picture*

Vocabulary Test, ToL = *Tower of London*, ToMI-2 = *Theory of Mind Inventory-2nd edition*, TMCQ = *Temperament in Middle Childhood Questionnaire*.

Before proceeding, we examined the relationships between the independent variables. The univariate correlations appear in Figure 1. As expected for two language measures, the TNL-2 and NIH Toolbox PVT scores were highly and positively correlated. The TNL-2 scores were also moderately correlated with the ToMI-2 scores. Higher language scores were associated with better theory of mind. Higher language scores were weakly correlated with the shyness scores. Shyer children had lower language scores on the TNL-2 than more outgoing children. The NIH Toolbox PVT scores were moderately correlated with age; higher vocabulary scores were associated with older ages. There were also weak correlations between vocabulary scores and scores on the TMCQ-shy and the ToMI-2. Children with larger vocabularies tended to be less shy and to have a stronger theory of mind.

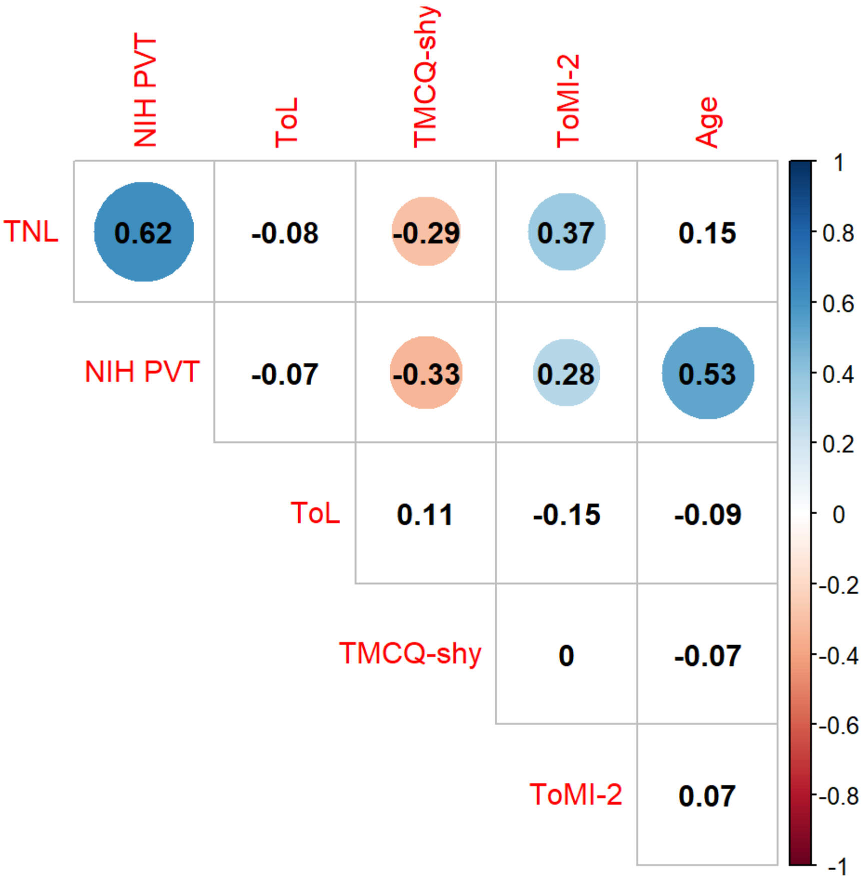


Figure 1. ***Matrix of univariate correlations between Predictor Variables. Cells with circles indicate a significant correlation (p < .05). Figure created via code in Wei and Simko (2017).***

We derived variance inflation estimates (VIF) to determine multicollinearity (Choueiry, 2021; Fox & Weisberg, 2019). The VIF is equal to 1 when a given independent variable is orthogonal to the other independent variables. VIF values between 5 to 10 are considered large and indicative of multicollinearity. To anticipate, we ran models with language measured by the NIH Toolbox PVT or the TNL-2. In either case, VIF estimates were ≤1.775; thus, we could proceed with the models as planned. The detailed results of the VIF analysis appear in the Supplemental Materials.

Discourse

The expository discourse scores are plotted in Figure 2. Scores ranged widely, from as low as 27% of total possible points in the audiovisual condition to as high as 87% in the audio condition.

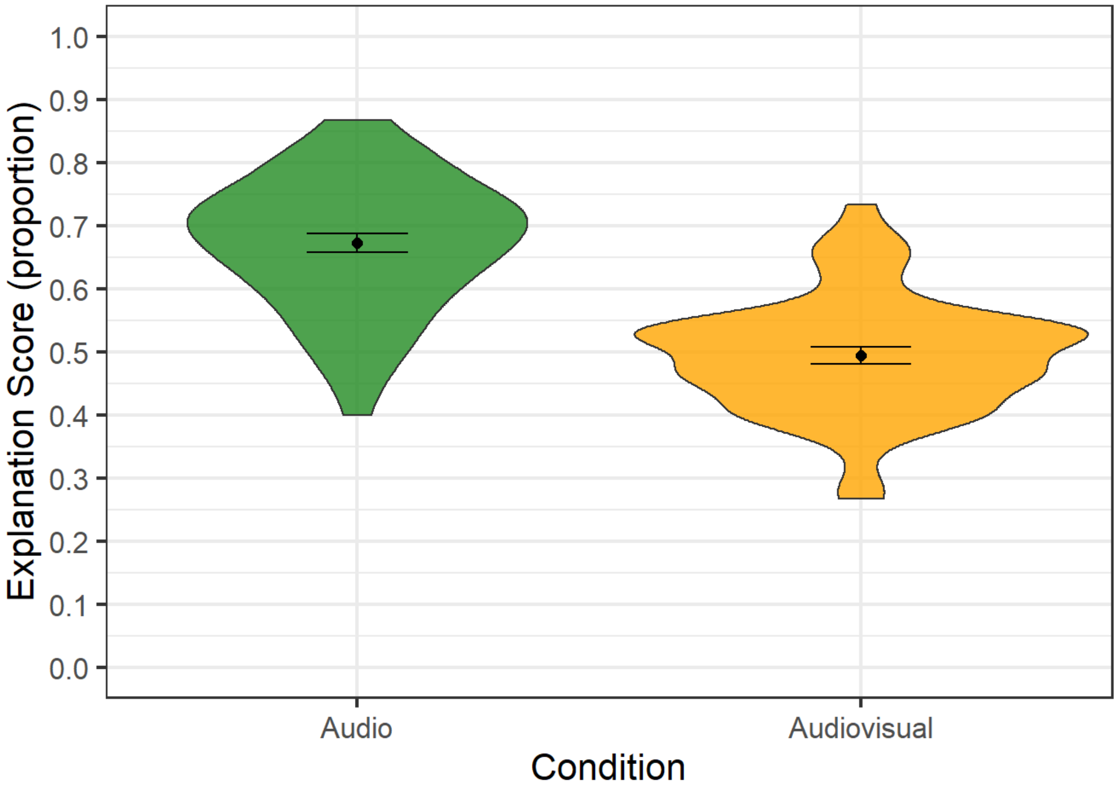


Figure 2. ***Explanation scores (proportion) as a function of Condition (Audio vs. Audiovisual). Diamonds represent the group average and error bars +/- 1 SE. Violins show the distribution of Explanation scores across children.***

The outcome of the model predicting expository discourse performance indexed by mean explanation scores appears in Table 2. Performance varied with condition; however, the effect was the opposite of our prediction. We found that children's explanation scores were significantly *lower* in the audiovisual (b=0.493) than in the audio (b=0.671) condition. As predicted, there was a significant effect of vocabulary. Children with larger vocabularies had significantly higher explanation scores. For instance, the average explanation score for a child with below-average vocabulary (i.e., 69.79; -1 SD below mean) was 0.55. The average explanation score for a child with an above-average vocabulary (i.e., 83.73; +1 SD above mean) was 0.62.

Table 2. ***Results of Linear Mixed Model Evaluating Predictors of Discourse Performance***

Variable	Estimate	SE	df	t	Pr(>|t|)
Intercept	0.582266	0.011192	43	52.023	< 0.00000001
Condition	-0.177751	0.013307	49	-13.358	< 0.00000001
Sex	-0.009997	0.023173	43	-0.431	0.6683
Age	-0.001828	0.002686	43	-0.680	0.4999
Vocabulary	0.005246	0.002130	43	2.463	0.0179
Tower of London	- 0.013447	0.035539	43	-0.378	0.7070
Theory of Mind	0.010406	0.006326	43	1.645	0.1073
Shyness	-0.002423	0.003554	43	-0.682	0.4990

Note: SE = standard error, df = degrees of freedom, Pr = probability

The remaining predicted effects were not obtained. Among these, it is notable that the explanation scores did not vary with performance on the ToL, the same task the participants were attempting to explain. Thus, any problems in explaining the task were *not* the result of an inability to understand its rules or goals.

To further understanding of the discourse performance, we examined variation by item (Table 3). The children seldom framed their discourse with salutations, openings, or closings. At the other extreme, nearly all children specified spatial and temporal information when conveying how to solve the ToL.

Exploratory analyses

The planned analyses revealed, contrary to prediction, that the children performed better in the audio condition than in the audiovisual condition. Before drawing any conclusions, we examined the data anew. First, we asked whether the condition effect varied with vocabulary knowledge to determine whether the decrease in performance in the video compared to the audio condition was greater for children with better vocabularies. A significant condition x vocabulary interaction could suggest that the lower explanation scores in the audiovisual condition indicate *more* mature behavior than higher scores(i.e., children with better vocabularies may be more aware of the increased common ground in the audiovisual condition and therefore decrease how much they explain). Next, we asked whether the condition effect varied with modality to determine whether the more robust overall performance in the audio condition was primarily the result of children using more spoken communication when on the phone. The results appear in Table 4.

Table 3: ***The proportion of participants who conveyed each item.***

Item	Audiovisual	Audio
Opening salutation	.22	.10
Closing salutation	.16	.04
Rule 1: boards must look alike	.42	.66
Rule 2: use fewest moves possible	.22	.20
Rule 3: number of beads must not exceed height of peg	.26	.32
Rule 4: move one bead at a time	.14	.30
Problem 1	.20	.54
Problem 2	.22	.60
Problem 3	.24	.50
Problem 4	.20	.46
Problem 5	.18	.38
Specific word for bead	.46	.66
Specific word for peg	.36	.66
Specific word for spatial information	.82	.96
Specific word for sequential information	.78	.84

Table 4: ***Results of Linear Mixed Model Evaluating Predictors of Discourse Performance***

Variable	Estimate	SE	df	t	Pr(>|t|)
Intercept	0.582266	0.011192	43	52.023	<0.0001
Condition	-0.177751	0.013132	48	-13.536	<0.0001
Sex	-0.009997	0.023173	43	-0.431	0.6683
Age	-0.001828	0.002686	43	-1.680	0.4999
Vocabulary	0.005246	0.002130	43	2.463	0.0179
Tower of London	-0.013447	0.035539	43	-0.378	0.7070
Theory of Mind	0.010406	0.006326	43	1.645	0.1073
Shyness	-0.002423	0.003554	43	-0.682	0.4990
Condition x Vocabulary	-0.002895	0.001902	48	-1.522	0.1346

Note: SE = standard error, df = degrees of freedom, Pr = probability

Does the Condition Effect Vary with Vocabulary Knowledge?

The significant effect of Vocabulary, b=0.005, t(43)=2.463, p=0.018 did not vary significantly between conditions, b=-0.003, t(48)=-1.522, p=0.135. In other words, the effect of condition, b=-0.178, t(48)=-13.536, p=0 (i.e., higher explanation scores in the audio than the audiovisual condition) was similar for children with smaller vocabularies and children with larger vocabularies.

Does the Condition Effect Vary with Modality?

We credited the child one point per the first 11 rubric items in the original analysis, whether produced in words, gestures, or both. Here, to discern any differential responses to condition according to the modality of the response, we instead awarded the child one point for each item conveyed with words (11 points maximum) and one point for each item conveyed with gesture(11 points maximum). Children responded primarily by demonstration gestures, scoring on average 5.66 points (SD=1.12) in the audio condition and 6.37 points (SD=1.13) in the audiovisual condition. Children only occasionally responded using deictic gestures, scoring on average 1.35 (SD=1.73) in the audio condition and 1.56 out (SD=2.05) in the audiovisual condition. In addition, children rarely (if ever) responded using representational gestures, scoring on average 0.34 (SD=1.02) in the audio condition and 0.14 (SD=0.4) in the audiovisual condition. Given the relative infrequency of deictic and representational gestures, we did not repeat our analyses separately for each type of gesture. These distributions, however, indicate that children's gesture explanation scores primarily reflect their ability to demonstrate actions on the objects used to complete the task.

Note that, when broken apart by modality, the children's explanation scores were highly correlated, b=0.724, t(48)=2.887, p=0.006. For every one-point increase in children's gestured explanation score, there is a 0.724 increase in their spoken explanation score.

We ran a linear mixed model that included the original model variables plus the modality and modality x condition variables. The results appear in Table 5.

There was a significant effect of condition, b=-0.037, t(97.997)=-2.134, p=0.035. As before, the explanation scores were *lower* in the audiovisual (b=0.4205) than the audio (b=0.4575) condition. Also as before, vocabulary was a significant predictor, b=0.007, t(42.999)=2.747, p=0.009. There was a significant effect of modality. Children's gesture explanation scores were *higher* (b=0.5755) than their spoken explanation scores (b=0.3025). The effect of condition was qualified by a condition x modality interaction, b=0.232, t(97.997)=6.614, p=0. This captures a reversal in the effect of condition.

Children's spoken explanation scores are *lower* in the audiovisual than the audio condition (b=-0.081), while their gestured explanation scores are *higher* in the audiovisual than the audio condition (b=0.151). In both conditions, however, children's gestured explanation scores are *higher* than their spoken explanation scores. This difference is larger in the audiovisual condition (b=0.389) than the audio condition (b=0.157) (Figure 3).

Table 5: ***Results of Linear Mixed Model Evaluating Predictors of Discourse Performance***

Variable	Estimate	SE	df	t	Pr(>|t|)
Intercept	0.439048	0.015628	43.152702	28.094	<0.0001
Condition	-0.037414	0.017528	97.996544	-2.134	0.03530
Sex	-0.002562	0.027175	42.998779	-0.094	0.92531
Age	-0.003715	0.003150	42.998779	-1.179	0.24484
Vocabulary	0.006863	0.002498	42.998779	2.747	0.00874
Tower of London	-0.020329	0.041675	42.998779	-0.488	0.62816
Theory of Mind	0.004530	0.007418	42.998779	0.611	0.54464
Shyness	-0.003506	0.004168	42.998779	-0.841	0.40497
Modality	0.272788	0.024900	48.998357	10.955	<0.0001
Condition x Modality	0.231879	0.035057	97.996544	6.614	<0.0001

Note: SE = standard error, df = degrees of freedom, Pr = probability

Language Ability as a Predictor of Expository Discourse

We already have some evidence that language ability influences expository discourse, given that vocabulary scores were a significant predictor. Next, we determined whether the effect was limited to vocabulary or extended to language ability more broadly defined. Because performance on the NIH PVT and the TNL-2 were significantly correlated, r = .62, we removed vocabulary from the model and replaced it with the TNL-2 scores. The results appear in Table 6.

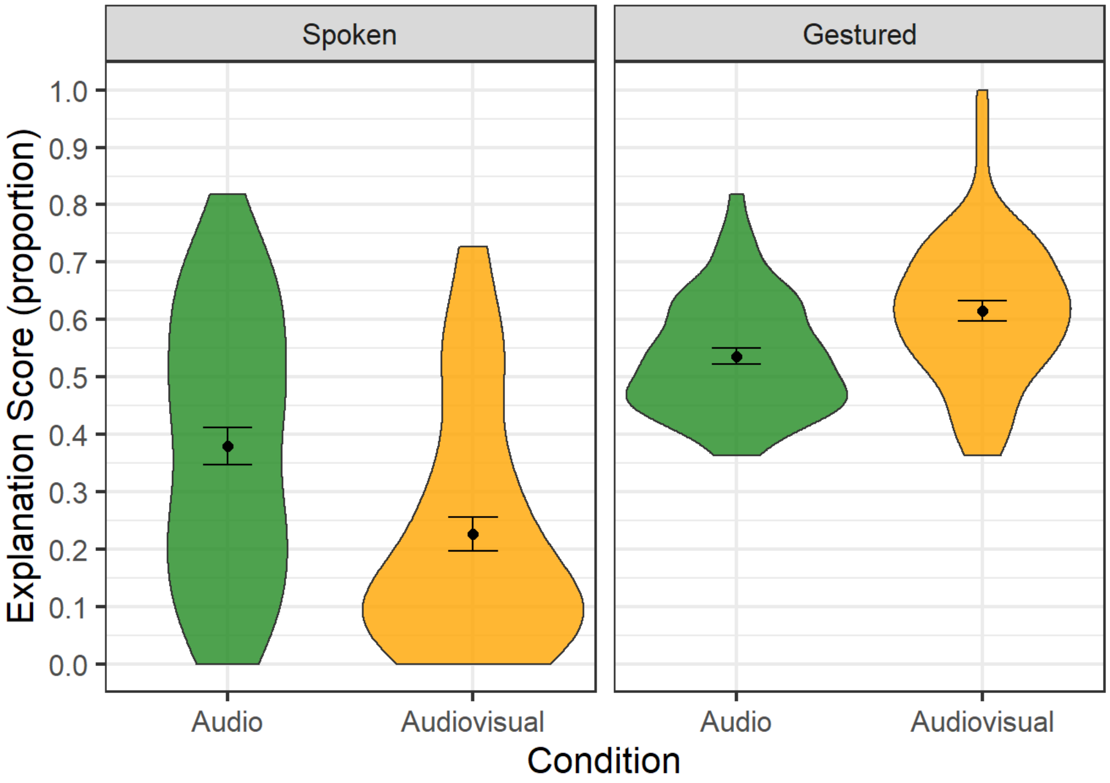


Figure 3: ***Explanation scores (proportion) as a function of Condition (Audio vs. Audiovisual) and modality (Spoken vs. Gestured). Diamonds represent the group average and error bars +/- 1 SE. Violins show the distribution of Explanation scores across children.***

Table 6: ***Results of Linear Mixed Model Evaluating Predictors of Discourse Performance***

Variable	**Estimate**	**SE**	**df**	**t**	**Pr(>|t|)**
Intercept	0.5819619	0.0095073	42	61.212	<0.00000002
Condition	-0.1777509	0.0130155	48	-13.657	<0.00000002
Sex	-0.0082890	0.0195446	42	-0.424	0.6737
Age	0.0002123	0.0019940	42	0.106	0.9157
Language	0.0035786	0.0007266	42	4.925	0.0000136
Tower of London	-0.0108946	0.0301276	42	-0.362	0.7195
Theory of Mind	0.0043402	0.0055416	42	0.783	0.4379
Shyness	-0.0009602	0.0029443	42	-0.326	0.7459
Sex x Language	-0.0011283	0.0013768	42	-0.819	0.4171

Note: SE = standard error, df = degrees of freedom, Pr = probability

As expected by now, there was a significant effect of condition. Children's explanation scores were significantly lower in the audiovisual (b=0.493) than in the audio (b=0.671) condition. There was a significant effect of language ability: Children with a higher score on the TNL-2 had significantly higher explanation scores. For instance, the average explanation score for a child with a below-average TNL-2 score (i.e., 89.78; -1 SD below mean) was 0.52. The average explanation score for a child with an above-average TNL score (i.e., 119.7; +1 SD above mean) was 0.64. There were no other significant effects.

Confirmatory Analyses

In our planned analysis, we found a significant effect of condition (audio vs. audiovisual). This analysis, however, controlled for individual differences in our predictor variables (i.e., it was the effect of condition for a child with average performance on each measure). Moreover, the model did not include the full random effects structure (Barr et al., 2013). We could not include a random slope for condition because the model would overfit the data (the number of random effects would match the number of data points per participant - 2). Therefore, we repeated our analyses using the difference between children's explanation scores in each condition and without the predictor variables. This linear model allowed a more direct test of Hypothesis 1 - testing whether explanation scores differ for children in general, and the use of a linear model provided a more transparent way to calculate p-values and effect sizes without the need for random effects. Consistent with the results of our planned analysis, children's total explanation score was significantly lower in the audiovisual condition than the audio condition, b=-0.178, t(49)=-13.358, p=<.001. The condition effect accounted for 78.8% of the variance in children's explanation scores. We fit a second linear model in which we regressed children's explanation scores (averaged across Conditions) on sex, vocabulary, ToL performance, TMCQ-Shyness, ToMI-2, and age. The significant effect of vocabulary was confirmed, b=0.007, t(43)=2.457, p=0.018. Children with larger vocabularies have significantly higher explanation scores. For instance, the average explanation score for a child with below-average vocabulary (i.e., 69.79; -1 SD below mean) was 0.75. The average explanation score for a child with above-average vocabulary (i.e., 83.73; +1 SD above mean) was 0.85. Controlling for the other predictor variables, the size of children's vocabulary accounted for 12.3% of the variance in children's explanation scores. No other predictors were significant, p's > 0.112. The details appear in the Supplemental Materials.

Discussion

Remote communicative contexts are part of everyday social, familial, and academic interactions for the modern child. However, much of what we know about the child's ability to meet the informational needs of a communicative partner is based on data collected during face-to-face interactions. This study adds to the sparse extant literature on communicative success in remote contexts.

Thanks to the rise in video-chat apps available in home and school settings, remote communication increasingly involves a shared visual context. Traditional phone communication does not. Thus, in this study, we were interested in the completeness of children's remote discourse, its variation with the presence or absence of visual context, the extent to which children modified their use of gestural communication when moving between contexts, and the characteristics that predict children's communicative success. Below we organize our findings into two main categories: how remote communication varies with context and how it varies with child characteristics.

Variation Associated with Context Demands

Remote communication places a high demand on speakers. They must infer and then meet their listeners' need for information without many of the cues available in face-to-face communication. Given that the audiovisual condition reinstates some of these cues, we predicted that the children's overall explanation scores would be higher in the audiovisual condition than in the audio condition. Specifically, we anticipated that the children would use gestures to supplement their spoken messages. Relative to the audio condition, the children did gesture more information in the audiovisual condition, but, at the same time, they provided less information in the spoken modality; thus, the overall rubric score was higher in the audio condition, contrary to prediction. This finding is consistent with Cameron and Lee (1997), who reported that 3-to-8-year-olds provided more detail and specificity while speaking on the phone than in person. The children responded to their listeners' needs by adjusting to the listener's need for spoken input.

The particular items that the children most often included in their discourse also illustrate their sensitivity to listeners' needs. Consider, for example, the low rate of salutations and the high rate of spatial and temporal content. The pragmatic framing of the discourse with salutations was uncommon, but such niceties are not necessary for explaining the task (and perhaps awkward given that the partner was not present). In contrast, when and where to move the beads were details that nearly every child included, and this information was essential for solving the task.

Thus, just as they do when face-to-face (e.g., Akhtar et al., 1996; Mori, & Cigala, 2016; Nadig & Sedivy, 2002; Nilsen, & Fecica, 2011; Shatz & Gelman, 1973), children demonstrate adaptations in content and modality according to their listener's needs when communicating remotely. When the remote context lacked shared visual reference, they enhanced the clarity and completeness of their spoken messages. When their partner was able to see them, they offloaded some of their verbal explanation into gestures. In her nuanced description of discourse in a year-five classroom, Taylor

(2014, p. 416) wrote:

> “It is important to emphasize that modes other than language are not simply additional contextual information but part of an enmeshed nexus of many modes used in conjunction with one another for the purpose of making meaning. All modes are potentially available for making meaning, within the constraints of our social world. The mode selected by the communicator is the one judged by them to be the most apt and expedient at that moment in time.”

Gesture was, of course, an apt and expedient means of communication in the audiovisual condition. The types of gestures the children used were well suited to the partners’ needs. All of the participants used demonstration gestures. These were hand gestures that resulted from manipulating the materials, in other words, moving the beads. As the primary goal was to teach the partner how to move the beads, demonstration gestures were an effective means of explanation.

Of course, the demonstration gestures were apt and expedient for the partner in the audiovisual condition only. Nonetheless, the children gestured more items than they presented in words even in the audio condition. We do not take this as a counter to the conclusion that they were sensitive to the partners' needs. Instead, gestures can be apt and expedient for the speaker as well as the listener (Goldin-Meadow, 2003). Speakers use gestures with exceptionally high frequency when communicating spatial information (Alibali, 2005). When explaining a visual-spatial task, working out the problem by moving the hands through space is an excellent strategy for thinking through the steps one must convey. The children likely gestured in the audio condition (and to some extent in the audiovisual condition) because the gestures helped them explain the task. Had the children been given repeated practice with the ToL, we would predict less reliance upon gestures that involved demonstrations on objects and more free-handed gestures. As it were, their high use of demonstration gestures was consistent with their status as novice ToL solvers (Roth, 2002). We turn now to other characteristics that were related to the success of their remote discourse.

Variation Associated with Language Ability

Motivated by previous work on the influence of language, theory of mind, and temperament on communicative success, we tested the predictive utility of language scores, theory of mind ratings, and temperament—specifically shyness—ratings in our models of remote discourse success. However, we first measured the inter-dependence of these predictors. Given the equivocal reports of relationships between shyness and theory of mind, some reporting a negative relationship (Banerjee & Henderson, 2001; DeRosnay et al., 2014; Walker, 2005) and others a positive relationship (Mink et al., 2014; Wellman et al., 2011), it is noteworthy that we found neither in the current sample.

That said, we did find a relationship between shyness and language; specifically, the shyer children in our sample tended to have lower vocabulary and receptive/expressive language scores than more outgoing children. Spere et al. (2004) compared the receptive vocabulary scores of four-year-olds grouped as shy or not shy and found the shyer children to have significantly lower scores. Here we extend these findings to an older cohort. Some have speculated that their reticence to speak masks the language competency of shy children (see summary in Coplan & Evans, 2009). Although this could be the case, we found a negative relationship between shyness and performance on a receptive vocabulary test (which does not require spoken responses), a finding at odds with the masked language competency hypothesis. Another possibility is that the lower test scores on both the TNL-2 and the NIH-PVT reflect more test anxiety on the part of the shyer children, but we think this is unlikely given that previous work has established the validity of standardized tests administered to shy children. Specifically, shy children did not perform better on language tests administered in the home by a familiar adult than on those same tests administered at school by an unfamiliar adult (Spere et al., 2009).As has been previously proposed (Spere et al., 2004), we think it likely that children who are shy limit their opportunities for language learning by refraining from social-communicative interactions. Shyness (or temperament more broadly measured) may be a source of individual differences in children's language outcomes.

Language, as measured by the TNL-2 and, to a lesser extent, by the NIH-PVT, was also correlated with theory of mind. Milligan et al. (2007) also reported a positive relationship between language and theory of mind with an overall effect size of .43 among children below seven. Here we extend that finding to children who are seven to nine and report a similar effect size of .37 (on the TNL-2). The relation between language and theory of mind is likely bidirectional. Children who participate frequently and competently in communicative exchange access multiple opportunities for learning about others' mental states, and conversely, children who are skilled at mind-reading may learn mental state vocabulary and hone their social language skills upon realizing that their listener is confused, skeptical, interested or bored (De Rosnay et al., 2014). Language ability in the form of complex sentence construction may also aid thought about others' mental states, especially at the relatively older ages tested here, years during which children may be progressing from first order (I suspect he is hungry) to second-order observations (He knows that I suspect he is hungry) (de Villiers, 2007).

Our analyses allowed us to examine the potential effects of language, theory of mind, and shy temperament on discourse, each after controlling for the others. We also examined the effects of sex, ability to solve the ToL, and age. Whether measured as receptive vocabulary or a receptive and expressive narrative ability, language was the only predictor. Children with stronger language abilities produced more complete explanations of the ToL during the discourse task, as evident by their rubric scores.

There were no significant interactions between language and condition or between language and modality. In other words, regardless of their vocabulary knowledge, children tended to offload more information onto gestures when the visual context allowed. The high positive correlation between gesture scores and spoken language scores further supports the conclusion that children's gesture use was a sign of the integrity of their overall communicative competence rather than a way of compensating for communicative weaknesses. These findings are consistent with age-related differences in communicative competence. Older children not only use more complex spoken language but also more complex gestures than younger children (Alamillo et al., 2013). Like teachers who package relevant information into their gestures during classroom lectures (Alibali et al., 2013; Ovendale et al., 2018), children who frequently gesture when sharing a visual context with their interlocutor likely maximize the effectiveness of their message.

We do not dismiss the potential influence of theory of mind or shy temperament in other discourse contexts. Recall that, in the discourse task used here, we told the children that their communication partner did not know how to solve the ToL; thus, we likely reduced the need for mind reading. Moreover, the children did not interact with their partner but, instead, were recorded for later listening or viewing. This situation may have lessened the burden that shyer children may have felt had they been part of an actual exchange. The decision to simulate phone and video chat rather than engage the children in these actual contexts was purposeful. We wanted to control the amount of feedback a listener would provide, but we could not imagine how to do so in a pragmatically appropriate way. By recording the children's discourse, we got around this problem. That said, this strength is also a limitation of the work. The child was at a remove from an actual communicative exchange. Moreover, the child did not receive the scaffolding that the verbal and gestural responses of a listener would have provided, which surely made the discourse task more difficult than usual, perhaps especially so for those with weaker language abilities. Observations of naturalistic remote discourse would be a valuable complement to the work reported here. Also, a comparison between the two types of remote discourse studied here and actual face-to-face discourse would be helpful if we are to understand fully the challenges involved in remote communication.

Finally, we turn to the implications of the language as a predictor of remote discourse skills. This remote discourse task was difficult. None of the children provided 100% of the information we deemed essential. That said, some of the children had particularly poor performance. On average, those whose receptive/expressive language scores fell one standard deviation below the mean provided only 30% of the essential information. Real-world remote communication is likely to be challenging for these children unless their partner provides ample scaffolding in the form of feedback and questions. Given the ubiquity of remote communication in children's lives, it is essential to document how the estimated 9% of children with language disorders (Norbury et al., 2016) fare in remote contexts and determine the supports needed to ensure adequate remote discourse function.

Conclusions

Among the limitations of this study were our inability to complete it as registered and the lack of ecological validity inherent in a partnerless simulation. That said, we provided evidence of positive relationships between language and theory of mind and negative relationships between language and shyness, extending the extant literature to older children. We also confirmed that, as a group, seven-to-nine-year-olds adjust their discourse to the needs of their remote communication partners. They include essential semantic information, and when they know that their partner does not share their visual context, they still gesture frequently, but they increase their reliance upon the spoken modality. Perhaps the primary contribution of this work is the finding that remote discourse is challenging, even for children as old as nine, and especially so for children who have below-average receptive and expressive language abilities. This finding has important practical implications given that children's communication partners—their friends, families, teachers, and health care providers—are frequently remote. Children with low language abilities may experience functional limitations during remote communication, a context that is increasingly necessary in today's world.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development, 67(2),* 635-645. https://doi.org/10.1111/j.1467-8624.1996.tb01756.x

Alamillo, A.R., Colletta, J., & Guidetti, M. (2013). Gesture and language in narratives and explanations: the effects of age and communicative activity on late multimodal discourse development. *Journal of Child Language, 40,* 511-538. https://doi.org/10.1017/S0305000912000062

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. *Spatial Cognition and Computation, 5*(4), 307-331. https://doi.org/10.1207/s15427633scc0504_2

Alibali, M. W., & Goldin-Meadow, S. (1993). Modeling learning using evidence from speech and gesture. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 203-208). Hillsdale, NJ: Erlbaum.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*(6), 593-613. https://doi.org/10.1080/016909600750040571

Alibali, M. W., Young, A.G., Crooks, N.M., Yeo, A., Wolfgram, M.S., ... Knuth, E.J. (2013). Students learn more when their teacher has learned to gesture effectively. *Gesture, 13* (2), 210–233. https://doi.org/10.1075/gest.13.2.05ali

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences, 4*(7), 267-278. https://doi.org/10.1016/S1364-6613(00)01501-1

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology, 35*(5), 1311. https://doi.org/10.1037/0012-1649.35.5.1311

Banerjee, R. & Henderson, L. (2001). Social-cognitive factors in childhood social anxiety. *Social Development*, 10(4), 558-572. https://doi.org/10.1111/1467-9507.00180

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3),* 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General, 136*(4), 539. doi:10.1037/0096-3445.136.4.539

Buss, A. H., & Plomin, R. (1984). *Theory and measurement of EAS, temperament: early developing personality traits,* pp. 84-104. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Cameron, C. A., & Lee, K. (1997). The development of children's telephone communication. *Journal of Applied Developmental Psychology, 18*(1), 55-70. https://doi.org/10.1016/S0193-3973(97)90014-9

Capone, N. C., & McGregor, K. K. (2004). Gesture development. *Journal of Speech, Language, and Hearing Research, 47*, 173-186. https://doi.org/10.1044/1092-4388(2004/015)

Castelli, F., Frith, C., Happé, F., & Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain, 125*(8), 1839-1849. https://doi.org/10.1093/brain/awf189

CCSS (Common Core State Standards). (2015). English Language Arts Standards » Speaking & Listening » Grade 4 » 6 | Common Core State Standards Initiative (corestandards.org). Downloaded 1/10/2021.

Choueiry, G. (2021). Correlation vs Collinearity vs Multicollinearity. Correlation vs Collinearity vs Multicollinearity – Quantifying Health Downloaded 3/11/21.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*(1), 43-71. https://doi.org/10.1016/0010-0277(86)90053-3

Coplan, R. J., & Evans, M. A. (2009). At a loss for words? Introduction to the special issue on shyness and language in childhood. *Infant and Child Development, 18,* 211–215. https://DOI: 10.1002/icd.620

Coplan, R. J., & Weeks, M. (2009). Shy and soft-spoken: shyness, pragmatic language, and socio-emotional adjustment in early childhood. *Infant and Child Development: An International Journal of Research and Practice, 18*(3), 238-254. https://doi.org/10.1002/icd.622

Culbertson, W.C., & Zillmer, E.A. (2005). *Tower of London Drexel University, 2nd Edition,* North Tonawanda, NY: Multi-Health Systems

de Rosnay, M.C., Fink, E., Begeer, S., Slaughter V., & Peterson C. (2014). Talking theory of mind talk: young school-aged children's everyday conversation and understanding of mind and emotion. *Journal of Child Language,* 41(5), 1179-1193. https://doi.org/10.1017/S0305000913000433

de Villiers J. (2007). The Interface of Language and Theory of Mind. *Lingua. International review of general linguistics. Revue internationale de linguistique generale, 117*(11), 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006

Dollaghan, C. A. (2004). Taxometric analyses of specific language impairment in 3- and 4-year-old children. *Journal of Speech, Language & Hearing Research, 47*(2). https://doi.org/10.1044/1092-4388(2004/037)

Dorval, B., Eckerman, C. O. & Ervin-Tripp, S. (1984). Developmental trends in the quality of conversation achieved by small groups of acquainted peers. *Monographs of the Society for Research in Child Development,* 49(2), Serial No. 206. https://doi.org/10.2307/1165872

Ehrlich, S. B., Levine, S. C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology, 42*(6), 1259. https://doi.org/10.1037/0012-1649.42.6.1259

Erstad O, Flewitt R, Kümmerling-Meibauer B, et al. (eds) (2020) *The Routledge Handbook of Digital Literacies in Early Childhood.* Abingdon: Routledge.

Farrar, M. J., & Maag, L. (2002). Early language development and the emergence of atheory of mind. *First Language, 22*(2), 197-213. https://doi.org/10.1177/014272370202206504

Fox, J., & Weisberg, S (2019). *An R Companion to Applied Regression,* Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem-solving in adults and children. *Cognitive Science, 26*(6), 817-831. https://doi.org/10.1207/s15516709cog2606_5

Gershon, R. C., Gleason, J. B., Michnick Golinkoff, R., Jager Adams, M., Schnipke, D., Hirsh-Pasek, K., Slotkin, J., Manly, J. J., Blitz, D., Beaumont, J. L., Wallner-Allen, K., Weintraub, S. (2013). NIH Toolbox - Cognition - Early Childhood Picture Vocabulary Test (NIH TPVT - Early Childhood) National Institute of Health and Northwestern University. NIH TPVT - Early Childhood - NIH Toolbox - Cognition - Early Childhood Picture Vocabulary Test (mapi-trust.org)

Gillam, R. B., & Pearson, N. A. (2017). *TNL-2: Test of Narrative Language*. Pro-ed.

Gillen, J. (2002). Moves in the territory of literacy? The telephone discourse of three- and four-year-olds. *Journal of Early Childhood Literacy, 2*(1), 21-43. https://doi.org/10.1177/14687984020021002

Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2010). How do preschoolers express cause in gesture and speech?. *Cognitive Development, 25*(1), 56-68. https://doi.org/10.1016/j.cogdev.2009.11.001

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Belknap Press of Harvard University Press.

Henderson, H. A., & Wachs, T. D. (2007). Temperament theory and the study of cognition–emotion interactions across development. *Developmental Review, 27*(3), 396-427. https://doi.org/10.1016/j.dr.2007.06.004

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*(2), 297. https://doi.org/10.1037/a0022128

Hughes, C., & Leekam, S. (2004). What are the links between theory of mind and social relations? Review, reflections and new directions for studies of typical and atypical development. *Social Development, 13*(4), 590-619. https://doi.org/10.1111/j.1467-9507.2004.00285.x

Hutchins, T. and Prelock, P., 2016. Technical manual for the theory of mind inventory-2. Unpublished Copyrighted Manuscript. Available at: theoryofmindinventory.com.

Koeze, E., & Popper, N. (April 7, 2020). The virus changed the way we internet. The New York Times. Downloaded on 1/10/2021 from https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html.

Lancaster, H. S., & Camarata, S. (2019). Reconceptualizing developmental language disorder as a spectrum disorder: Issues and evidence. *International Journal of Language & Communication Disorders, 54*(1), 79-94. https://doi.org/10.1111/1460-6984.12433

Larson, C., Gangopadhyay, I., Kaushanskaya, M., & Weismer, S. E. (2019). The relationship between language and planning in children with language impairment. *Journal of Speech, Language, and Hearing Research,* 62(8), 2772-2784. https://doi.org/10.1044/2019_JSLHR-L-18-0367

Lidstone, J. S., Meins, E., & Fernyhough, C. (2012). Verbal mediation of cognition in children with specific language impairment. *Development and Psychopathology., 24*(2), 651-660. http://dx.doi.org/10.1017/S0954579412000223

Lundine, J. P., & McCauley, R. J. (2016). A tutorial on expository discourse: Structure, development, and disorders in children and adolescents. *American Journal of Speech-Language Pathology,* 25(3), 306-320. https://doi.org/10.1044/2016_AJSLP-14-0130

Marton, K. (2008). Visuo-spatial processing and executive functions in children with specific language impairment. *International Journal of Language & Communication Disorders,* 43(2), 181-200. https://www.tandfonline.com/doi/citedby/10.1080/16066350701340719?scroll=top&needAccess=true

McClure, E. R., Chentsova-Dutton, Y. E., Barr, R. F., Holochwost, S., & Parrott, W. G. (2015). "Facetime doesn't count": Video chat as an exception to media restrictions for infants and toddlers. *International Journal of Child-Computer Interaction,* 6, 1–6. https://doi.org/10.1016/j.ijcci.2016.02.002

McClure, E. R., Chentsova-Dutton, Y. E., Holochwost, S. J., Parrott, W. G., & Barr, R.

(2018). Look at that! Video chat and joint visual attention development among babies and toddlers. *Child Development, 89*(1), 27-36. https://doi.org/10.1111/cdev.12833

McGregor, K. K., Eden, N., Arbisi-Kelm, T., Foody, M., & Oleson, J. (2019, September 3). Children's Vocabulary Project; Remote Communication. Retrieved from osf.io/cuhy 10.17605/OSF.IO/CUHYX

McGregor, K. K., Eden, N., Arbisi-Kelm, T., Oleson, J., & Pomper, R. (2020, November 6). Children's Vocabulary Project; Remote Communication. Retrieved from osf.io/we8am

Miller, C. A. (2006). Developmental relationships between language and theory of mind. *American Journal of Speech-Language Pathology, 15(2), 142-154.* https://doi.org/10.1044/1058-0360(2006/014)

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*(2), 622-646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Mink, D., Henning, A., & Aschersleben, G. (2014). Infant shy temperament predicts preschoolers theory of mind. *Infant Behavior and Development, 37*(1), 66-75. https://doi.org/10.1016/j.infbeh.2013.12.001

Mori, A., & Cigala, A. (2016). Perspective taking: Training procedures in developmentally typical preschoolers. Different intervention methods and their effectiveness. *Educational Psychology Review, 28*(2), 267-294. https://doi.org/10.1007/s10648-015-9306-6

MPFS: Medienpädagogischen Forschungsverbundes Südwest (Media Educational ResearchAssociationnSouthwest) (2018). Kindheit, Internet, Medien Studie (Childhood, Internet, Media Study). Downloaded 3/11/2021 from KIM-Studie 2018 - Kindheit, Internet, Medien - Initiativbüro Gutes Aufwachsen mit Medien (gutes-aufwachsen-mit-medien.de)

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science, 13(4),* 329-336. https://doi.org/10.1111/j.0956-7976.2002.00460.x

Nilsen, E. S., & Fecica, A. M. (2011). A model of communicative perspective-taking for typical and atypical populations of children. *Developmental Review, 31*(1), 55-78. https://doi.org/10.1016/j.dr.2011.07.001

Neppl, T. K., Donnellan, M. B., Scaramella, L. V., Widaman, K. F., Spilman, S. K., Ontai, L. L., & Conger, R. D. (2010). Differential stability of temperament and personality from toddlerhood to middle childhood. *Journal of Research in Personality, 44*(3), 386-396. https://doi.org/10.1016/j.jrp.2010.04.004

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology*, 17(4), 356-366. https://doi.org/10.1044/1058-0360(2008/07-0049)

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

Ovendale, A., Brookes, H., Colletta, J. M., & Davis, Z. (2018). The role of gestural polysigns and gestural sequences in teaching mathematical concepts: The case of halving. *Gesture, 17*(1), 128-157. https://doi.org/10.1075/gest.00013.ove

Pine, K. J., Lufkin, N., & Messer, D. (2004). More gestures than answers: children learning about balance. *Developmental Psychology, 40*(6), 1059. https://doi.org/10.1037/0012-1649.40.6.1059

Pinto, G., Tarchi, C., Gamannossi, B. A., & Bigozzi, L. (2016). Mental state talk in children's face-to-face and telephone narratives. *Journal of Applied Developmental Psychology, 44*, 21-27. https://doi.org/10.1016/j.appdev.2016.02.004

Prior, M., Smart, D., Sanson, A. N. N., & Oberklaid, F. (2000). Does shy-inhibited temperament in childhood lead to anxiety problems in adolescence?. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(4), 461-468. https://doi.org/10.1097/00004583-200004000-00015

Research Registry 3425 (2017). The Dynamics of Word Learning, McGregor, P.I. Browse the Registry - Research Registry

Roello, M., Ferretti, M. L., Colonnello, V., & Levi, G. (2015). When words lead to solutions: Executive function deficits in preschool children with specific language impairment. *Research in Developmental Disabilities,* 37, 216-222. https://doi.org/10.1016/j.ridd.2014.11.017

Roth, W. M. (2002). From action to discourse: The bridging function of gestures. *Cognitive Systems Research, 3*(3), 535-554. https://doi.org/10.1016/S1389-0417(02)00056-6

Schmidt, L. A., & Tasker, S. L. (2000). Childhood shyness: Determinants, development and ‘depathology’. In W.R. Crozier, (Ed.) *Shyness: Development, consolidation and change*, 30-46. Routledge.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London*, 298(1089), 199-209. https://doi.org/10.1098/rstb.1982.0082

Shatz, M, & Gelman, R. (1973). The development of communication skills:Modifications in the speech of children as a function of listener. *Monographs of the Society for Research in Child Development*, 38(5), Serial No. 152. https://doi.org/10.2307/1165783

Simonds, J., & Rothbart, M. K. (2006). Temperament in Middle Childhood Questionnaire. *Downloaded in http://www. bowdoin. edu/~ sputnam/rothbart-temperament-questionnaires.*

Smith Watts, A. K., Patel, D., Corley, R. P., Friedman, N. P., Hewitt, J. K., Robinson, J. L., & Rhee, S. H. (2014). Testing alternative hypotheses regarding the association between behavioral inhibition and language development in toddlerhood. *Child Development, 85*(4), 1569-1585. https://doi.org/10.1111/cdev.12219

Spere, K. A., Evans, M. A., Hendry, C. A., & Mansell, J. (2009). Language skills in shy and non-shy preschoolers and the effects of assessment context. *Journal of Child Language, 36*(1), 53. https://doi:10.1017/S0305000908008842

Spere, K. A., Schmidt, L. A., Theall-Honey, L. A., & Martin-Chang, S. (2004). Expressive and receptive language skills of temperamentally shy preschoolers. *Infant and Child Development: An International Journal of Research and Practice, 13(2),* 123-133. https://doi.org/10.1002/icd.345

Taylor, R. (2014). Meaning between, in and around words, gestures and postures–multimodal meaning-making in children's classroom discourse. *Language and Education, 28*(5), 401-420. https://doi.org/10.1080/09500782.2014.885038

Uccelli, P., Demir-Lira, Ö. E., Rowe, M. L., Levine, S., & Goldin-Meadow, S. (2019). Children's early decontextualized talk predicts academic language proficiency in mid adolescence. *Child Development, 90*(5), 1650-1663. https://doi.org/10.1111/cdev.13034

Veinott, E. S., Olson, J., Olson, G. M., & Fu, X. (1999, May). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 302-309). https://doi.org/10.1145/302979.303067

Walker, S. (2005). Gender differences in the relation between children's peer-rated social competence and individual differences in theory of mind. *Journal of Genetic Psychology,* 166(3), 297-312. https://doi.org/10.3200/GNTP.166.3.297-312

Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: The Psychological Corporation.

Wei, T. & Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Wellman, H. M., Lane, J. D., LaBounty, J., & Olson, S. L. (2011). Observant, nonaggressive temperament predicts theory-of-mind development. *Developmental Science, 14*(2), 319-326. https://doi.org/10.1111/j.1467-7687.2010.00977.x

Zhao, L., Wang, J. J., & Apperly, I. A. (2018). The cognitive demands of remembering a speaker's perspective and managing common ground size modulate 8-and 10-year-olds' perspective-taking abilities. *Journal of Experimental Child Psychology, 174,* 130-149. https://doi.org/10.1016/j.jecp.2018.05.013

Data, Code and Materials Availability Statement

The raw data, analysis code, plots of data distributions, videotaped examples, and a link to the registration appear in McGregor et al. (2020, OSF | Children's Vocabulary Project; Remote Communication)

Ethics Approvals and Consent

Ethics approval was obtained from the ethics committee of the Boys Town National Research Hospital. All participants and their parents gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

KKM secured funding for the study, conceived of the study, designed the study, supervised data collection and analysis, wrote the introduction, results, and discussion section and revised the methods section. RP contributed to the planning of the statistical analysis, conducted the statistical analysis and prepared the figures. NE collected data, coded the discourse data, and wrote the methods section as it pertained to the discourse data. TA-K collected data and wrote the participants section of the Methods. NO collected data and wrote the Tower of London description in the Methods. SG conducted reliability checks on the discourse data and assisted with the literature search. ES assisted with statistical analysis and wrote the statistical analysis section of the Methods. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

We thank the children and their families for participation and Jacob Oleson for consultation on the statistical analyses. We give additional thanks to the family who agreed to share the video-taped examples of the discourse task. This work was supported by the National Institutes of Health-NIDCD, grant number 2R01DC011742-06. The funder had no part in the study design; the collection, analysis, and interpretation of data; the writing of the report; or the decision to submit the article for publication.

Appendix

Discourse Instructions for Participants

"Now that you know how to play this game, I'm going to ask you to teach two of my friends."

Audio

"This is X (*photo of woman A or B on phone*). She couldn't be here today but she told me that we could record the directions. Then she will borrow the game later and try it out. She doesn't have a computer screen so she won't be able to see you but she DOES have her cell phone, so she can listen to your recording. I'll start the recording (*make a big fuss about the microphone*). You explain to her how to play the game, and she will listen to what you say. Be very careful to tell her what the game looks like, what the rules are, and exactly how to play."

Now that they know about the game, you can help them solve five of the problems. I'll get the game set up for you each time. "You are the teacher. Tell X how to make this one (*pt to child's board*) look like this one (*pt to target board further away*)."

Repeat for items p, p, 1, 2, 3.

"That was great! Okay, I'm going to ask you to do that one more time for a different friend."

Audiovisual

"This is X (*photo of woman A or B facing screen*). She couldn't be here today but she told me that we could video the directions. Then she will borrow the game later and try it out. She has a computer so she WILL be able to hear you AND see you. I'll start the recording (*make a big fuss about the video camera*). You explain to her how to play the game. She will watch and listen to what you say. Be very careful to tell her what the rules are, and exactly how to play."

Now that they know about the game, you can help them solve five of the problems. I'll get the game set up for you each time. "You are the teacher. Tell X how to make this one (*pt to child's board*) look like this one (*pt to target board further away*)."

Repeat for items p, p, 1, 2, 3.

Debriefing

You did so well being the teacher. Sometimes it is harder to be the teacher when the person you are talking to can't see you. My friends are going to study your recordings to see how children talk to people who can't see them, like when you talk on a phone. They won't really be watching to learn the game, that part was just pretend.

Discourse Scoring

One point was awarded for each item that the child completed successfully. Except for the final four items, the child was credited for successful completion via words OR gestures. For the five trials (items P1, P2, 1, 2, 3 on the ToL), we assumed that the listener had the visual context of the two boards, one in start position and one in target position. Total scores could range from 0 – 15.

Greets the listener (2 points):

- Hello
- Goodbye

Verbal credit was given if the child offered the listener a stated "hello" or "goodbye." Iconic gestures were credited if the child waved to greet or bid farewell to the listener.

***Rules (4 points)*:**

- Explains that the two boards must be alike
- Explains that you must use as few moves as possible
- Explains that no peg can have more beads than it can hold
- Explains that you can move only one bead at a time

Children earned one point per rule if they provided an adequate explanation with words and/or gestures.

Examples for *both boards must look alike*:

- o “You have to copy the teacher’s board.”
- o “Try to do the same pattern.”
- o The child set-up the teacher’s board with an example pattern, and then showed the listener how to manipulate the beads in order to make the other board match.

Examples for *complete the pattern in as few moves as possible*:

- o “Solve in the least amount of moves”
- o “Use less moves as possible”
- o “Whoever has less of the movings wins.”
- o The child used a hand gesture to illustrate moving the beads, while using the word ‘moves,’ and holding up two fingers to show that the example problem required only two moves.

Examples for *no peg can contain more beads than it can hold:*

- o “The small one can hold one *bead,* the middle one can hold two, and the tall one can hold three,”
- o “The large *peg* can hold three, the middle peg can hold two and the small one can hold one.” The child manipulated the beads on the board to show the listener the maximum bead amounts for each peg.
- o The child pointed to a peg as they stated how many beads could be placed on it.

Examples for *move only one bead at a time.*

- o “You can only move one bead at a time.”
- o “You can only take one off the peg at a time.”
- o Children often explained this rule via demonstration gestures by showing the various ways in which this rule can be broken, in the same way it was presented to them prior to completing the standardized portion of the test (i.e., lifting two beads off the pegs with one hand/lifting two beads off the pegs with two hands/lifting one bead off the peg and placing it on the table, and then proceeding to take another bead off a peg).

Example Problems (5 points):

- Gives enough information for listener to correctly solve trial P1, minimum number of moves is not required
- Gives enough information for listener to correctly solve trial P2, minimum

number of moves is not required

- Gives enough information for listener to correctly solve trial 1, minimum number of moves is not required
- Gives enough information for listener to correctly solve trial 2, minimum number of moves is not required
- Gives enough information for listener to correctly solve trial 3, minimum number of moves is not required

To earn verbal points in the absence of gesture, the child was required to provide the listener with sufficient information to solve the problem accurately using only spoken language. Credited instructions included referents and/or descriptive words for each move within the problem. Acceptable descriptors for identifying the target peg included distinctions in size (e.g., short/tall/little/long/middle) and body-oriented directional terms (e.g., left/middle/right.) Credit was also given if the child identified where a bead should be placed by stating the color of the bead already atop the target peg (e.g., Put the green one on the red one), or by stating the color of the bead that was in the target position prior to the previous move (e.g., "Put the red one where the blue one used to be"). In the absence of gesture, environmental-oriented directional terms (e.g., front/back, first/last) were not considered specific enough to describe target pegs (e.g., "Put the blue bead on the front peg and move the red bead to the back one" was too vague to receive credit because it is not clear which is front and which is back).

To earn credit via demonstration gestures, the child needed to move the beads from the starting position to the target position without breaking any rules. Credit was given for the successful completion of the problem, despite move count.

Children received credit if they combined verbal description and gesture to clearly convey content:

- "The red one goes on this peg" (pointing to peg)
- "See this blue bead [participant moves bead closer to the listener], it goes on this peg".
- "I move the red one to the small peg, and that's one move [holds up one finger]", and "I'm going to switch these around [moves hand from left to right to indicate switching]."

Specific content words (4 points):

- Refers to the beads with a relevant word (e.g., bead, ball) at least once
- Refers to pegs with a relevant word (e.g., peg, stick, stand) at least once

- Uses a relevant directional term (e.g., here/there, right/left) at least once
- Uses a relevant sequential term (e.g., first, next, then) at least once

The words ‘bead’ and ‘peg’ were credited with one point each if used correctly to identify the corresponding item, at least once, during the discourse task. Acceptable synonyms were also given credit, and included words such as, *ball* and *block* for the target vocabulary word *bead,* and the words ‘*stick* and *stand ’* for the target word *peg*. Spatial/Location terms (e.g., here/there, middle, tall) and sequential words (e.g., first, next, then) were credited, with one point each, if the participant used one or more of these during the problem-solving portion of the task.

License

Sleep behaviour in children with language disorder

Victoria C. P. Knowland
Mohreet Rauni
M. Gareth Gaskell
Sarah Walker
Elaine van Rijn
Lisa-Marie Henderson
University of York, UK

Courtenay Norbury
University College London, UK

Abstract: Sleep and language are intimately linked over childhood, yet objective measurements of sleep behaviour have never been compared between children with developmental disorders of language and their language-typical peers. The aim of this two-study series was to assess an emergent hypothesis that children with poor structural language development may also exhibit poor sleep. In Study 1, 196 parents of 4-10 year-old children completed the Children's Sleep Habits Questionnaire and the Children's Communication Checklist-2, including the parents of 61 children with reported language disorder. Parent-reported sleep behaviour and language ability showed a positive correlation, with children who scored more highly on the language measure showing better sleep behaviour. Interestingly, parental estimates of sleep duration showed an unexpected reverse pattern, with children who scored lower on the language measure being reported to go to bed earlier and sleep for longer than their peers. In Study 2, a subsample of 20 4-to-6 year-old children with language disorder and 20 language-typical age-matched peers contributed objective, actigraphy-derived estimates of sleep duration, efficiency and onset latency. Mirroring parental estimates in Study 1, actigraphy data showed the language disordered group slept for longer and more efficiently than their language-typical peers. We consider that parental perception of poor sleep behaviour in children with language difficulties may result from a history of poor sleep and/or from observed difficulties in sleep parameters that are not possible to assess with actigraphy. The data suggest that subjective reports of sleep behaviour and objective estimates of children's sleep be thought of as complementary.

Keywords: Sleep; Language Disorder; Actigraphy; CCC-2; CSHQ

Corresponding author: Victoria Knowland, Department of Psychology, University of York, YO10 5DD, UK; now at School of Education, Communication and Language Sciences, Newcastle University, NE1 7RU, UK. Email: vic.knowland@newcastle.ac.uk

ORCID ID: 0000-0003-4367-3689

Citation: Knowland, V.C.P., Rauni, M., Gaskell, M.G., Walker, S., van Rijn, E., Norbury, C.F., & Henderson, L-M. (2021). Sleep behaviour in children with language disorder. *Language Development Research*, 1(1), p 283-318. DOI: 10.34842/fs09-ga12

Introduction

A developmental relationship between sleep and the emergence of linguistic skill in typically developing children has been repeatedly demonstrated. Observationally, unfragmented night-time sleep is positively associated with language performance in pre-school (Lam et al., 2011; Touchette et al., 2007; Quach et al., 2009), and school-age children (Buckhalt et al., 2009). Experimentally, daytime naps enhance new word learning in pre-school children (see Axelsson, Williams & Horst, 2016; Hubach et al., 2009; Kurdziel et al., 2013), and such behavioural gains are positively correlated with expressive vocabulary skill (Horváth et al., 2015). In school-aged children, sleep, compared to equivalent time awake, has been found to support the consolidation of declarative word learning and lexical integration (Henderson et al., 2012) and benefit the learning of word-pair associations (Backhaus et al., 2008).

Decades of work with adults has built a picture of the neural mechanisms by which sleep promotes the consolidation of new memories, including linguistic material, through a process of hippocampal re-activation (see Paller et al., 2021 for a recent review). While equivalent work on the mechanisms on memory consolidation during sleep is still to emerge in children, behavioural research converges on the importance of sleep for the acquisition of language. Despite this, children's behavioural sleep habits, such as duration and efficiency, have never been objectively measured in those with language disorder.

Sleep in Language Disorder

The term *language disorder* refers to a neurodevelopmental disorder characterised by a deficit in the acquisition of language over childhood at any level of language description and in both receptive and expressive modalities. This definition covers idiopathic developmental language disorder (DLD), but extends more broadly to include any children who may not meet the criteria for DLD but 'who are likely to have language problems enduring into middle childhood and beyond' (Bishop et al., 2017; p.1070).

Describing behavioural sleep habits in the language disordered population is of considerable theoretical and clinical interest. Data from electroencephalography recording suggest that around half of children with DLD show atypical electrophysiological activity such as epileptiform discharges during sleep (Dlouha et al., 2020; Echenne et al., 1992; Fabbro et al., 2000; Overvliet et al., 2011; Picard et al., 1998). Initial behavioural evidence also exists to suggest that sleep may not support language learning in individuals with DLD to the same extent or in the same way as in typically developing peers, with adults who have language disorder showing reduced overnight consolidation of new phonemic learning (Earle et al., 2017). A description of sleep behaviour in children with language disorders is currently a missing link in understanding the association between sleep and language development. By 'sleep behaviour' here we mean habitual patterns of behaviour, cognition and emotion which occur during and proximate to sleep. Measurements of sleep behaviours includes subjective estimates, as well as objectively measured sleep parameters such as duration and timing of sleep activity.

Botting and Baraka recently explored subjectively measured sleep habits of 3-18 year old children with language disorders or typical development, using parent report (Botting & Baraka, 2017). Children with language disorders were reported to experience longer sleep-onset latencies (the time it takes to fall asleep after lights out) than their typically developing peers, and were more likely to wake early. Across the sample, sleep-onset latency was found to correlate with both syntax and semantic/pragmatic ability as measured by the Children's Communication Checklist (CCC; Bishop, 1998). One other study using parent-report found that children with clinically meaningful delays in receptive vocabulary at 60 months showed less mature sleep patterns (i.e., less consolidated night-time sleep) at 6 and 18 months of age compared to children with typical language development or transient delays (Dionne et al., 2011)[1].

Sleep in Autism

The work of Botting and Baraka represents an important step forward in understanding the sleep behaviours of children with language disorders. However, nearly a third of the language disordered participants in this study also had an autism spectrum condition. This is an important limitation as children with autism are already known to show extended sleep-onset latency according to parental report (see Díaz-Román et al., 2018 for a review). Indeed links between sleep difficulties and autism have been fairly consistently demonstrated. According to parent report, sleep problems co-occur with early autism symptoms and worsen over development (Verhoeff et al., 2018), with children who have autism going to bed later and getting up earlier than their peers from around 30 months of age (Humphreys et al., 2014). Over the pre-school years, sleep problems as defined by the Children's Sleep Habits Questionnaire (CSHQ; Owens et al., 2000) are more than twice as common in children with a diagnosis of autism, with group differences emerging on every subscale of the questionnaire (Reynold et al., 2019). Actigraphy data from pre-schoolers with autism and general developmental delay have also shown greater night-to-night variability in sleep measures for both groups compared to typically developing peers (Anders et al., 2011). Overall, sleep difficulties in this group are seen in objectively recorded global measures such as total sleep time (Elrod et al., 2015) but are more consistently observed in subjective, parent-reported measures (Díaz-Román et al., 2018).

The Current Study

The current paper aims to describe basic sleep behaviour in relation to language development over childhood. The extant literature is suggestive of a link between impoverished sleep behaviour and the disordered development of structural language; however this is currently based on subjective parent-report. The nature of objectively measured sleep behaviour in children with clinically significant language deficits is yet to be described. Furthermore, it is not yet clear whether an association between

[1] While this study considered a linguistic domain relevant to clinical language disorders (receptive vocabulary), diagnoses were not reported.

sleep problems and language disorder can be explained by the inclusion of children with autism in previous studies of language disorder (Marini et al., 2020; Williams et al., 2008). We therefore focused on the relationship between sleep in early-mid childhood (primary school age) and the acquisition of oral language independent of social communication skills over two studies. Study 1 employed parent-report to consider subjective relationships between language ability and sleep behaviour, while Study 2 employed actigraphic recording to look at the duration and efficiency of children's sleep along with objectively measured language ability.

Study 1 utilised two questionnaires to replicate and extend the work of Botting and Baraka (2017). The aim of Study 1 was to describe basic, parent-reported sleep behaviour and estimates of sleep quantity (as described by the Children's Sleep Habits Questionnaire) in relation to language development (as described by parent report and the Children's Communication Checklist-2) in primary-school aged children without autism. We hypothesised that children whose parents reported better language ability would also have fewer parent-reported sleep problems.

Study 1

Method

Measures

Parents were asked to fill out two well-established questionnaires, the CSHQ (Owens et al., 2000), and the Children's Communication Checklist-2 (CCC-2; Bishop, 2003), along with their child's age, sex, and a description of any developmental disorders and/or diagnoses. The study was granted ethical approval from the Department of Psychology's Departmental Ethics Committee at the University of York.

The CCC-2 is a 70 item parent-rated questionnaire, which asks respondents to quantify their children's strengths and weaknesses in communication on a scale of 0 ("less than once a week") to 3 ("every day"). The questionnaire is split into 10 sub-scales, which generate a General Communication Composite (hereafter referred to as CCC General) and a Social Interaction Deviance Composite (hereafter referred to as CCC Social). The CCC General describes structural language ability and is composed of the subscales: A-Discourse, B-Syntax, C-Semantics, D-Coherence, E-Inadequate initiation, F-Stereotyped language, G-Use of context and H-Non-verbal communication. The CCC Social describes whether or not pragmatic aspects of communication are in line with a child's general communication skill and is calculated by subtracting the age-normed scores for the grammatical/semantic sub-scales (A + B + C + D) from the age-normed scores for the pragmatic sub-scales (E + H +I-Social relations + J-Interests), a score of 0 suggests that structural language and social language are exactly in line. A score below 55 on the CCC General, in conjunction with a CCC Social score of 9 or more is consistent with a profile characteristic of DLD. A CCC General score below 55 with a negative CCC Social score is suggestive of autism, as is a CCC Social score of -15 or below with any CCC General score.

The CSHQ is a 33 item sleep screening instrument which asks parents about their child's sleep habits over the last week (or the most recent typical week). The CSHQ is concerned with sleep behaviour, that is, behavioural habits, emotions and cognitions about sleep, night-time activity including sleep-walking (an example of a parasomnia), as well as medical aspects such as sleep-disordered breathing as indicated by snoring. Scores are given for eight subscales plus a total score (CSHQ Total), with high scores indicating more difficulty in that domain. The subscales are as follows: Bedtime Resistance; Sleep Onset Delay; Sleep Duration; Sleep Anxiety; Night Wakings; Parasomnias; Sleep Disordered Breathing; and Daytime Sleepiness. Each item is scored on a scale of 1-3, such that the minimum score is 33 and the maximum 99; the clinical threshold for concern on the CSHQ is a total (sum) score of 41. Parents are also asked to estimate their child's 'bed time', 'waking time' and their 'usual amount of sleep each day'. Test-retest reliability for subscales ranges from r= .62-.79, while sensitivity for distinguishing between clinical and control groups is .80, and specificity .72. In addition to these two published questionnaires, parents were asked to fill out a descriptive Sleep History questionnaire devised by the research team to give an overall impression of how parents viewed their child's sleep. This questionnaire is available in Supplementary Materials.

Participants

In total, 273 datasets were available for analysis. 242 datasets were collected from parents completing the questionnaires online; these parents were recruited through social media, parent groups, schools and the University of York newsletter. A link to the questionnaire was sent out with a brief description of the aims of the study which mentioned the team's interest in all children, particularly those with developmental disorders of language. An additional 31 datasets were included from a previous study in the lab (Fletcher et al., 2019; Knowland et al., 2019), to which participants were recruited either as typically developing controls, or on the basis of parental concerns about language development (n = 9). Methods of recruitment were the same in this latter case, but parents completed the questionnaires on paper.

Thirty datasets were removed as parents reported that their child had a diagnosis or suspected diagnosis of autism (to address whether any differences in sleep behaviours are apparent in language disorder independent of autism); a further 14 were removed because parents did not complete the CSHQ. As the CSHQ was developed to assess the sleep behaviour of 4-to-10 year old children (Owens et al., 2000), and the CCC-2 was developed to assess the language profiles of 4-to-16 year olds (Bishop, 2003), children outside the age range of the CSHQ were removed from the analysis. This process left 196 participants whose parents reported no developmental concerns (n=135) or whose parents described a developmental language difficulty but no other biomedical condition (n=61).

If parents reported that their child had a difficulty with language development they were asked to describe it and to provide any diagnoses their child had been given. Sixty one parents described their child as having a difficulty with language development that extended beyond pronunciation. Although a smaller number (n=38) used the term DLD or similar, we included all 61 children in a Language Disordered (LD) group as language disorder of unknown origin is understood to extend beyond the group of children who meet criteria for a diagnosis of DLD (see Norbury et al., 2016).

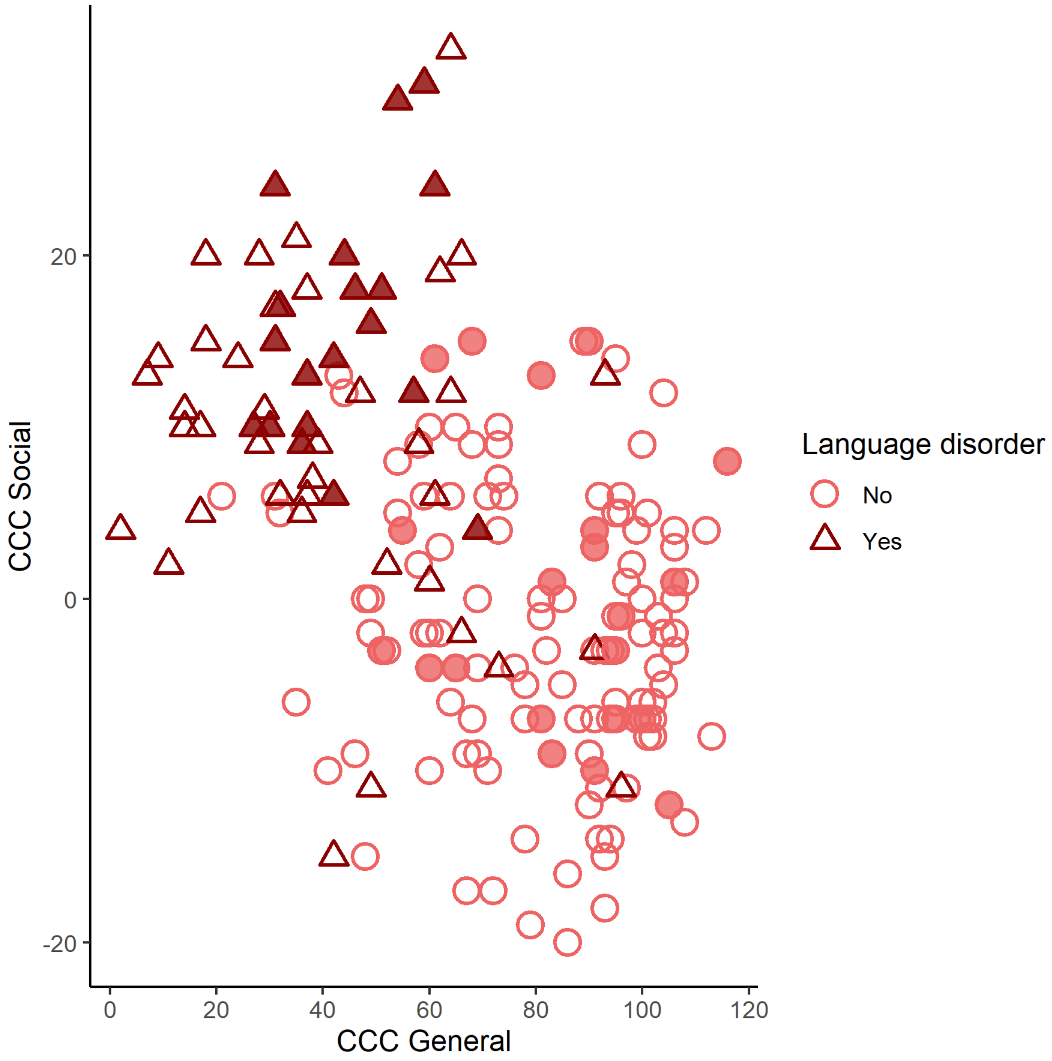


Figure 1. ***Scores on the CCC General (CCC-2 GCC subscale) and CCC Social (CCC-2 SIDC subscale). A CCC Social score of 0 suggests that social and pragmatic skills are exactly in line with general language skill, while scores above this indicate better social and pragmatic skill compared to language. 188 participants in Study 1 are shown who either did or did not have a Language Disorder according to parent report. (Those participants represented by filled shapes also participated in Study 2.)***

Of the sample of children in the LD group, 38 were male and 23 female, with an average age of 80.44 months (6 years, 8 months; *SD* = 23.58 months); while in the comparison group (Typically Developing; TD), 76 were male and 59 female, with an average age of 86.09 months (7 years, 2 months; *SD* = 23.58 months). CSHQ profiles for the whole sample are given in Table 1. The Language Disordered group scored significantly lower on CCC General (group mean = 42.11 (*SD* = 21.26), compared to the TD group (group mean = 81.51, *SD* = 20.72; *t*(104.2)=11.770, *p* <0.001), but had higher scores on the CCC Social as this measure is relative to language ability (LD group mean = 10.98, *SD* = 9.52; TD group mean = -1.43, *SD* = 8.24; *t*(94.2) = -8.547, *p* <0.001). The relationship between CCC General and CCC Social is illustrated in Figure 1 for each group.

Table 1. ***CSHQ profiles for the sample N = 196. SDB = Sleep Disordered Breathing.***

CSHQ subscale	# of items	Mean	*SD*	Min	Max	Skew	Kurtosis
Bedtime resistance	6	7.76	2.49	2	16	1.31	1.16
Sleep onset delay	1	1.62	0.76	1	3	0.76	-0.86
Sleep duration	3	4.10	1.52	2	9	1.26	0.64
Sleep anxiety	4	5.63	1.89	2	12	1.18	0.81
Night wakings	3	4.06	1.48	2	9	1.44	1.39
Parasomnias	7	8.80	1.73	7	15	0.85	0.17
SDB	3	3.39	0.85	2	8	2.75	9.79
Daytime sleepiness	8	11.01	2.51	8	18	0.82	-0.06
Total	33	45.78	8.10	33	73	0.75	-0.17

Results

Exploratory Analyses

Analysis of the questionnaire data was undertaken in an exploratory manner to allow a focused, pre-registered analysis of the objective, actigraphy data in Study 2. Correlations were assessed between CCC General and each of the CSHQ subscales (see Table 2). High scores on the CSHQ subscales indicate poor sleep, while high scores on the CCC-2 indicate better language ability. The negative correlations evident in Table 2 therefore suggest that those with better language have better sleep habits.

Table 2. ***Correlations between each subscale of the CSHQ and the general language score from the CCC-2 (CCC General). Italic font indicates those p-values that do not survive Bonferroni-Holm correction. 188 parents completed both the CCC-2 and the CSHQ. SDB = Sleep Disordered Breathing. *p<0.05, **p<0.01, ***p<0.001.***

			CCC General		*CCC Social*	
		df	*r*	*p*	*r*	*p*
CSHQ subscale	Bedtime resistance	186	-0.10	0.194	0.06	0.400
	Sleep onset delay	186	-0.12	0.111	-0.04	0.580
	Sleep duration	186	-0.25	<0.001***	0.10	0.174
	Sleep anxiety	186	-0.23	0.002**	0.08	0.270
	Night wakings	186	-0.22	0.003**	0.21	0.004**
	Parasomnias	186	-0.26	<0.001***	-0.03	0.725
	SDB	186	-0.16	*0.031**	0.15	*0.037**
	Daytime sleepiness	186	-0.16	*0.024**	0.02	0.772
	Total	186	-0.34	<0.001***	0.09	0.220

Having established sample-wide associations between CSHQ and CCC General, a weighted regression was run to predict CSHQ Total score from a binary measure of whether children were reported by their parent as being language disordered or not. Five predictors were controlled for in the model, as they might be expected to explain variance in the dependent measure independently of the main predictor of interest. The continuous predictor Age (in months), and the binary predictor Sex were included, along with binary predictors describing whether or not parents reported a difficulty with Attention, Literacy and/or Social interaction, each of which has been associated with sleep differences in children (Carotenuto et al., 2016; Díaz-Román et al., 2018; Mehta et al., 2019 respectively). Parents had an opportunity to report these issues either in response to whether their child had a developmental disorder such as dyslexia or ADHD (‘*Does your child have a diagnosis, or possible diagnosis, of any other developmental disorders?*’), or when they described a language disorder (‘*Please describe your child's language difficulties and what their diagnosis is, if they have one.* ‘). Of these five predictors, Age (*B* = -0.02, z = -2.83, p = 0.005), Literacy (*B* = 2.90, *z* = 3.99, *p* <0.001) and Social interaction (*B* = 3.02, *z* = 2.58, *p* = 0.010) were predictive of Language Disordered group membership. These five predictors were used to calculate propensity scores for membership of the Language Disordered group.

Table 3 shows the details of a linear regression model predicting CSHQ Total by LD group membership, Literacy, Social skills, Attention, Age and Sex weighted by propensity score. The weighted model controls for differences between the LD and TD groups with respect to those factors included in the propensity score. After weighting, the groups are matched with respect to these factors, allowing an analysis of the effect of LD group membership only.

The model significantly predicted CSHQ Total; *F* (11.2, 189) = 3.213, *p* = 0.005, with membership of the Language Disordered group being the sole independently significant predictor. An unweighted model was also run as is presented alongside the weighted model in Supplementary Materials (Table SM1). Notably, membership of the Language Disordered group remains a significant predictor in the unweighted model. The relationship between language skill and CSHQ Total score is illustrated in Figure 2, which shows CSHQ Total for the LD and TD groups. Despite some visual indication of bimodality in the LD scores, both groups show unimodal distributions according to Hartigan's Dip Test (for TD *D* = 0.037, *p* = 0.120 and for LD *D* = 0.519, *p* = 0.246).

Table 3. ***Weighted regression model predicting CSHQ Total scores. ***p<0.001.***

	B	Lower 95% CI	Upper 95% CI	*t*	*p*
Intercept	44.96	40.03	49.88	17.879	<0.001***
LD group	3.21	0.98	5.44	2.820	0.005***
Age	-0.04	-0.08	0.01	-1.530	0.128
Sex	2.05	-0.24	4.34	1.756	0.081
Attention	8.90	-0.21	18.01	1.915	0.057
Literacy	2.93	-1.30	7.15	1.359	0.176
Social	6.72	-1.22	14.66	1.659	0.099

The CSHQ asks parents for their child's 'bed time', 'waking time' and their 'usual amount of sleep each day'. Given that bed times, wake times and sleep duration change as children get older (see Acebo et al., 2005; Iglowstein et al., 2003), we ran partial correlations to assess relationships with CCC General, taking Age (in months) into account. Controlling for Age, significant partial correlations emerged between CCC General and 'bed time' ($r_{partial}$ (174) = 0.18, p = 0.020), as well as 'usual amount of sleep' ($r_{partial}$ (160) = -0.20, p = 0.013), but not with 'waking time' ($r_{partial}$ (162) = 0.13, p = 0.100). So as general language ability increased, bed time got later in this sample, and sleep amount was reduced, suggesting that children with poorer language got *more* sleep than their peers rather than less. This result was unexpected in the context of the rest of the CSHQ showing the opposite pattern of association with language skill.

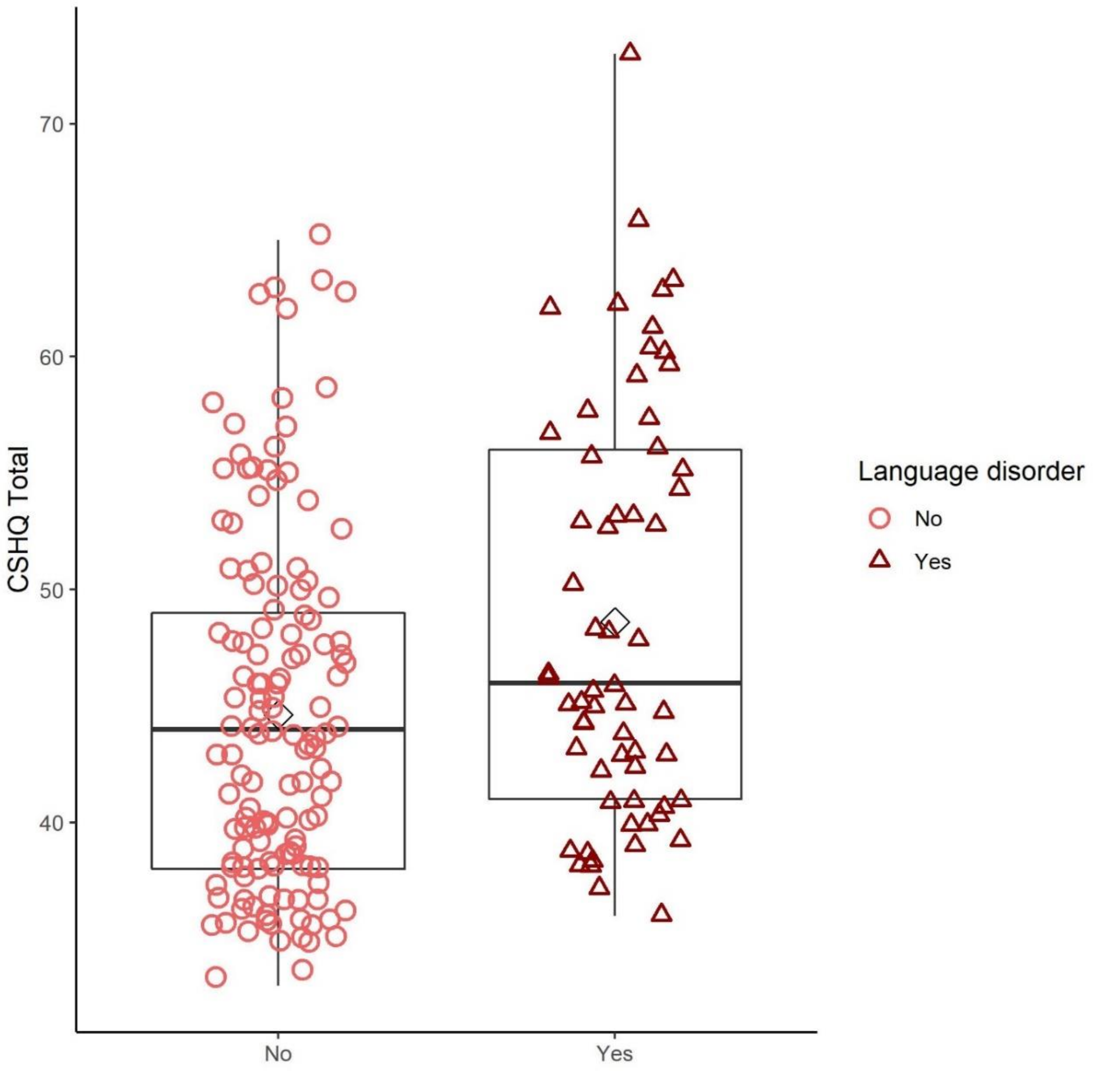


Figure 2. ***Box plot of CSHQ Total by membership of the Language disordered group.***

Discussion

Study 1 aimed to describe parent-reported sleep behaviour in primary-school aged children as a function of parent-reported language ability. We saw an association between better general language ability as described by the CCC-2 and better scores in the following domains of sleep behaviour as described by the CSHQ: Sleep Duration, Sleep Anxiety, Night Wakings, and Parasomnias (Sleep Disordered Breathing and Daytime Sleepiness were also associated with language ability but did not survive Bonferroni-Holme correction). CCC Social was shown to correlate with the Night Wakings and Disordered Breathing scales of the CSHQ (although the latter did not survive Bonferoni Holm correction), with more sleep problems seen in those with better social/pragmatic skills relative to general language skill. CCC Social was included here to consider whether sleep variables co-vary more closely with structural or social/pragmatic aspects of language. The relative lack of correlations between CSHQ subscales and CCC Social does not indicate an absence of any relationship between social competence and sleep behaviour, just an absence of clear relationships between social difficulty over and above language difficulty, and sleep behaviour. The positive correlation between CCC Social and two CSHQ scales may reflect the fact that children in the LD group have higher CCC Social scores and more sleep problems in this sample.

In support of an association between fewer sleep problems and better language ability, we went on to show that parent-reported language difficulty (a binary measure) predicted CSHQ Total score. Interestingly though, while parents reported poorer sleep behaviours in those children with poorer language, when asked for numerical estimates of bed time, wake time and usual sleep amount, children with poorer language were shown to get *more* sleep than their peers rather than less.

In order to further understand the links between sleep behaviour and language ability in young children, Study 2 objectively measured sleep duration and efficiency using actigraphy in a subgroup of children with or without clinically significant language disorder. We focused on 4-to-6 year old children, as this is the earliest age at which language disorder is routinely diagnosed in clinic, and an age at which vocabulary development is rapid as children start school. The findings from Study 1 were used to pre-register hypotheses and analyses for Study 2 (https://osf.io/yftqb).

The predominant pattern to emerge from Study 1 was more sleep problems in children with poorer language; and while we also saw evidence for *longer* parent-reported sleep in those same children, we suspected that this might be due to the approximate nature of parent-reported bed and wake times. Under-estimates of time spent awake after lights out are particularly prevalent in parent-reported estimates of sleep (Dayyat et al., 2011). In Study 2 we therefore expected to see support for the idea that children with poor language have worse (objectively measured) sleep than their peers. Hypothesis 1: c*hildren with Language Disorder will show shorter sleep duration, and lower sleep efficiency, or more variability in these measures, compared to typically developing age-matched peers; we also hypothesise more parent-reported bedtime anxiety;* Hypothesis 2: l*anguage composite score will show a positive relationship with mean sleep duration, and sleep efficiency over and above the predictive power of social/pragmatic ability; while the latter will better predict sleep onset latency, and bedtime anxiety.*

Study 2

Method

Participants

Participants in Study 2 were a sub-sample from Study 1. Parents from Study 1 were recontacted if their children fulfilled criteria for inclusion in Study 2, that is they were aged 4.0-to-6.11 and either showed no indication of neurodevelopmental disorder, or were reported to have language concerns and also had a CCC-2 profile indicative of language difficulties. More typically developing males than females were invited to take part in order to match gender ratio across groups. Children were excluded from participation if they did not have sufficient oral language to provide assent or if English was not their first language. Children were also excluded if they were being raised bilingual as at the age of this sample, bilingualism is likely to be an important factor explaining variability in English language skill and may overshadow any influence of sleep.

We recruited and tested 52 children in total for this study. However, having decided before pre-registration to only include children with structural language difficulties, we then excluded data from children who had a speech sound difficulty for which they were receiving speech and language therapy, but who did not show a profile of language disorder on standardised assessment and whose parents did not report a language difficulty on the CCC-2; on these grounds, 10 children were excluded. Two further children were excluded because they did not provide enough actigraphy data[2] (one from the LD group and one from the TD group). This left a sample of 40 children: 20 controls with no reported language issues (14 males, 6 female) with a mean age of 64.90 months (5 years, 5 months; *SD* = 9.84 months) and 20 children with language disorder (15 male, 5 female) with a mean age of 66.40 months (5 years, 6 months; *SD* = 11.53 months). One child in the LD group was reported to be taking melatonin to support sleep at the time of data collection. See Figure 1 for a description of CCC-2 scores for this sample in relation to the larger sample included in Study 1.

All those in the Language Disordered (LD) group were being seen by speech and language services for issues relating to vocabulary and/or syntax development at the time of recruitment. 14 children in the LD group were classified by the CCC-2 as having language profiles consistent with a diagnosis of DLD (CCC General ≤ 55 & CCC Social ≥ 9), while the remaining 5 either had a CCC General score slightly higher than 55 (range = 57-69), or a CCC Social score slightly lower than 9 (range = 4-6). The cognitive scores and questionnaire scores for the LD and TD groups can be seen in Table 4, with the one subscale of the CSHQ that shows a group difference illustrated in Figure 3. Note that although most of these children show considerable difficulties in more than one domain of language function, three children were unable to complete some assessments and one child did not score below 1*SD* on any task. Based on postcode data, mean national Indices of Multiple Deprivation (IMD) decile was 7.65 for the TD group and 6.65 for the LD group, a non-significant difference (t (38.0) = 1.24, p = 0.223).

The TD children who took part in Study 2 can be considered a representative subsample of Study 1 with respect to parental views on sleep. The representativeness of the Study 2 TD sub-sample is demonstrated by a non-significant two sample Kolmo-

[2] In the pre-registration 21 children are included in the TD group. One child was removed after pre-registration due to insufficient actigraphy data to provide reliable estimates of sleep. Their non-inclusion did not change the interpretation of the data.

gorov-Smirnov test of whether participants who were included in Study 2 can be considered to be drawn from the same population with respect to total CSHQ score as those not included (*D* = 0.189, *p* = 0.549).

Unfortunately, testing was interrupted by the COVID-19 pandemic; 25 children were tested before the UK lockdown on 23rd March, 2020 (17 TD and 8 LD), and 15 children were tested after. The children tested before lockdown were seen in person for their cognitive assessment, while the children tested after lockdown were tested via video call. This change meant that we were unable to use the Block Design subtest from the BAS-III for the children who were tested via video call; for these children we used the Matrices subtest from the BAS-III in order to measure visuospatial intelligence. We therefore report visuospatial intelligence for each group but do not use that measure as a co-variate. Regardless of how children were tested, parents provided written informed consent and each child gave verbal assent at the start of the first session. The study was granted ethical approval from the Department of Psychology's Departmental Ethics Committee at the University of York, as well as the Coventry and Warwickshire Research Ethics Committee on behalf of the UK National Health Service.

Measures

Data for this study consisted of: sleep measurements (up to ten nights of actigraphic recording, and up to ten nights of parent-reported sleep diary data – TD mean = 7.0 nights, *SD* = 0.00; LD mean = 7.25 nights, *SD* = 1.21); standardised language and cognition assessments; and parental questionnaires. For correlations between parent-report (CSHQ) and actigraphy measures of sleep and between parent-report and standardised measures of language, see Supplementary Materials (Tables SM2 & SM3).

Sleep Measures. Families were asked to complete seven consecutive nights of sleep measurement, using a Philips Respironics Actiwatch2 actigraphy watch and an online parental sleep diary (with nightly reminders provided via text or email). Children were asked to wear the watch on their non-dominant wrist during the night-time only. Children were not asked to wear the watch during the day as, after consultation with parents, it was felt that they may have removed and potentially misplaced the devices. Data from at least five nights was deemed sufficient to reliably establish objective measures of sleep duration and quality, including sleep onset latency and sleep efficiency (Acebo et al., 1999; Meltzer et al., 2012); participants who provided fewer than five nights were therefore excluded. Parents were asked to press a marker button on the watch to indicate when the child was left to sleep, and when they woke in the morning; the Actiwatch2 also has a luminance monitor. Parent diaries and luminance changes were used to mark the beginning and end of the rest period, from when children settled down to sleep to when they got out of bed in the morning. The actiwatch luminance monitor provides Lux-minutes (lux multiplied by sleep epoch length), to indicate the amount of light children were exposed to overnight. This measure did not differ between the TD (mean = 139.4 *SD* = 417.6) and LD (mean = 101.4 , *SD* = 245.9) groups: *t*(30.77) = 0.352, *p* = 0.728.

Actigraphy data were extracted via Respironics Actiware using the built-in algorithm. Data were collected in 30 second epochs. Sleep onset in the evening was calculated from the first epoch after which no activity was indicated for at least ten minutes (20 epochs). The two key estimates in this study were Sleep Duration and Sleep Efficiency; Sleep Duration refers to the total time that the child was asleep for (as distinct from the total time in bed), and Sleep Efficiency refers to the percentage of time the child was asleep for compared to total time in bed. Night waking was determined on an epoch-by-epoch basis, using an automated weighted calculation centred on the epoch of interest, and taking into account activity in the adjacent four epochs. The wake threshold was set to the default 'medium' (40 activity counts per minute). In 3-to-5 year old children the low (80 counts per second) and medium settings have both been shown to underestimate total sleep time relative to polysomnography (Meltzer et al., 2012), but as the high wake threshold (20 counts per second) can overestimate total sleep time in this age group, we kept the default. Sleep Onset Latency was also considered here and is defined as the time period between the start of the rest period and the first epoch marked as sleep. In pre-school children, sleep duration and efficiency metrics as measured by actigraphy correlate closely with concurrently measured polysomnography (intraclass correlations >.80), though number of awakenings show a weaker relationship (<.40) (Bélanger et al., 2013; Sitnick et al., 2008).

The study was presented to children as the PJ Heroes study, relating it to the children's TV programme 'PJ Masks', in which three children wear actigraphy-like watches to battle night-time villains. Children were given PJ Masks pyjamas and an Amazon voucher to thank them for their participation.

Questionnaires. CCC-2, CSHQ and descriptive Sleep History data from Study 1 were re-analysed for Study 2, and in addition, parents were asked to complete the Social Responsiveness Scale 2 (SRS; Constantino & Gruber, 2012); and Brown Attention Deficit Disorder scales (ADD; Brown, 2001). The SRS assesses difficulties in social behaviour associated with autism symptomology; parents were asked to complete this in order to establish whether sleep parameters were better explained by language or social/pragmatic factors. The ADD assesses attention behaviour in daily life, and was included here in order to describe the groups appropriately. The Sleep History questionnaire (see Supplementary Materials), included the question *Does your child get anxious about going to bed at night?* where parents were given the options 'no' (coded 1), 'somewhat' (coded 2) and 'yes' (coded 3).

Cognitive Battery. Children were assessed on cognitive and language ability using a series of standardised tasks in accordance with administration instructions: British Picture Vocabulary Scale, 3rd Edition (BPVS-III; Dunn et al., 2009); British Ability Scale 3rd Edition (BAS-3), Naming vocabulary subscale (Elliott & Smith, 2011); Nonword repetition subscale from the Comprehensive Test of Phonological Processing-2nd Edition (CTOPP-2; Wagner et al., 2013); Sentence Repetition subscale from the Clinical Evaluation of Language Fundamentals 5th Edition (CELF-5; Semel & Wiig, 2017); BAS-3, Pattern Construction subscale/ Matrices subscale. Cognitive assessment and questionnaire scores are given in Table 4.

Table 4. *Mean standard scores (and SD) for cognitive assessments carried out in Typically Developing (TD) and Language Disordered (LD) groups, and parent questionnaires.* * $p< 0.05$, ** $p< 0.01$, ***$p<0.001$. *SDB = Sleep Disordered Breathing; NWR = non-word repetition*

		TD Mean (*SD*)	LD Mean (*SD*)	*t*-test
Standardised	BAS-3 Naming	115.71 (12.0)	87.79 (14.5)	t (35.1) = 6.60, p < 0.001***
	BPVS-2	111.00 (11.4)	90.28 (13.2)	t (34.0) = 5.20, p <0.001***
	CELF-5 Recalling Sentences	118.10 (14.1)	82.22 (15.4)	t (34.9) = 7.55, p <0.001***
	CTOPP-2 NWR	108.33 (19.1)	70.28 (12.9)	t (34.2) = 7.37, p < 0.001***
	BAS-3 non-verbal measure	101.52 (14.0)	88.32 (21.7)	t (30.4)= 2.26, p = 0.031*
Parent questionnaires	SRS total (T-score)	45.86 (5.1)	58.15 (12.3)	t (25.1) = -4.14, p < 0.001***
	ADD total (T-score)	44.19 (6.4)	53.80 (7.7)	t (37.0) = -4.35, p < 0.001***
	CCC-2 General	85.90 (18.3)	44.65 (13.2)	t (36.4) = 8.32, p <0.001***
	CCC-2 Social	-0.14 (7.8)	15.50 (7.0)	t (38.9) = -6.76, p <0.001***
	CSHQ_Bedtime resistance	8.19 (2.7)	8.30 (2.7)	t (38.9) = -0.13, p = 0.898
	CSHQ_Sleep onset delay	1.62 (0.7)	1.60 (0.8)	t (38.8)= 0.08, p = 0.935
	CSHQ_Sleep duration	3.86 (1.2)	4.15 (1.5)	t (35.7) = -0.70, p = 0.489
	CSHQ_Sleep anxiety	5.67 (1.9)	6.40 (2.0)	t (38.4)= -1.20, p = 0.239
	CSHQ_Night wakings	3.95 (1.2)	4.30 (1.5)	t (36.7)= -0.80, p = 0.430
	CSHQ_Parasomnias	8.57 (1.5)	9.25 (1.8)	t (37.0)= -1.33, p = 0.191
	CSHQ_SDB	3.05 (0.2)	3.55 (0.7)	t (22.6)= -3.13, p = 0.005**
	CSHQ_Daytime sleepiness	10.52 (1.9)	10.30 (2.5)	t (36.2)= 0.32, p = 0.748
	CSHQ_Total	42.81 (5.8)	46.60 (8.5)	t (33.4) = -1.65, p = 0.108

Confirmatory analysis plan

To assess Hypothesis 1, mean and night-to-night variability (standard deviation) of objective Sleep Duration and Sleep Efficiency estimates were established for each participant. To analyse group differences between the typically developing (TD) and LD groups, t-tests were run on the mean and variability observed for each objective behaviour estimate.

To assess Hypothesis 2, linear regressions were run to test whether performance on a composite of standardised scores (Language Composite) from all four language measures (Receptive Vocabulary, Expressive Vocabulary, Sentence Repetition, and Non-word Repetition), SRS total score, or an interaction between the two would predict mean objective Sleep Duration, mean objective Sleep Efficiency, and mean objective Sleep Onset Latency. A logistic regression was run (N = 40) to assess whether Language Composite scores, total SRS score or an interaction between the two, could predict the presence of parent-reported bedtime anxiety from the Sleep History questionnaire.

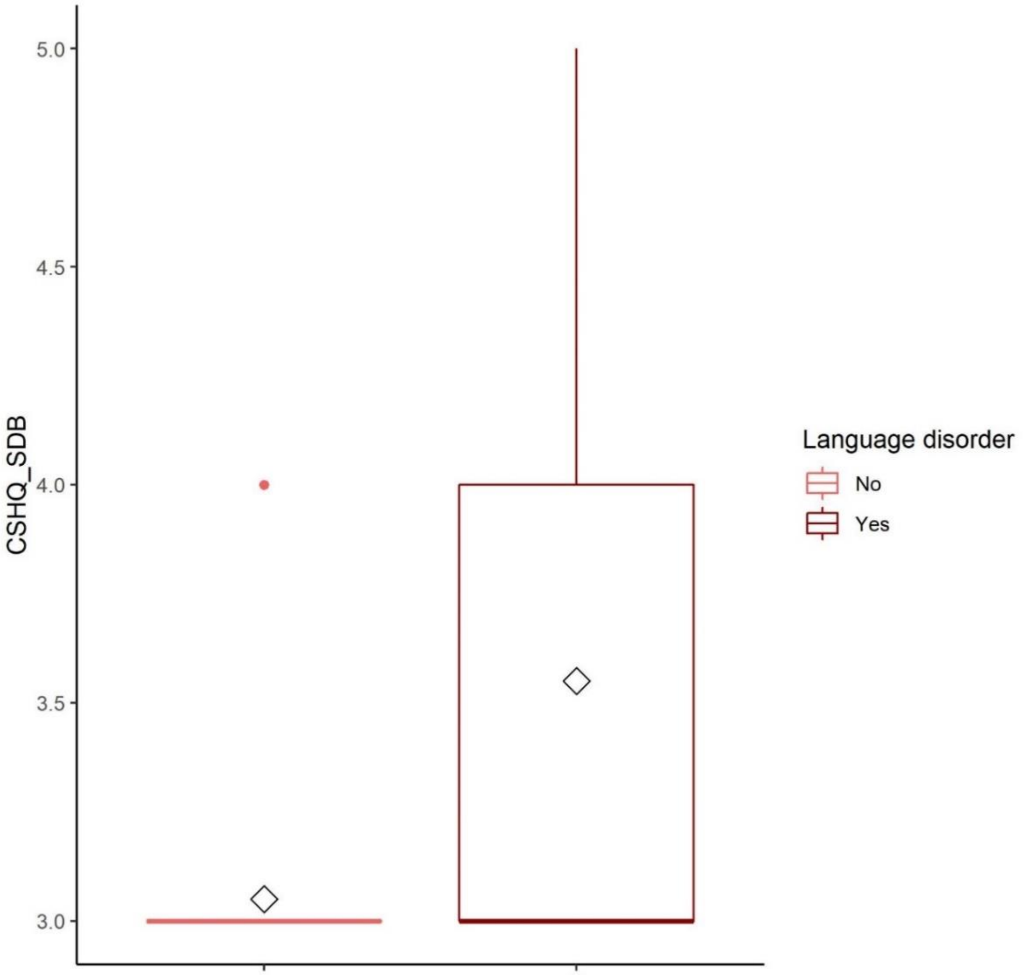


Figure 3. ***Box plot showing a group difference on the Sleep Disordered Breathing subscale of the CSHQ (sum of 3 items, each scored 1-3).***

Results

Confirmatory Analysis

Confirmatory Group Differences: Hypothesis 1. Mean Sleep Duration differed significantly between groups, but contrary to Hypothesis 1, the TD group showed shorter Sleep Duration (mean = 518.7 minutes, *SD* = 21.9) than the LD group (mean = 546.3, *SD* = 45.6), *t* (27.39) = -2.44, *p* = 0.022 (see Figure 4). No group differences emerged for night-to-night variability in Sleep Duration (TD mean = 45.4 minutes, *SD* = 14.7; LD mean = 41.1 minutes, *SD* = 15.9; *t* (37.76) = 0.88).

Mean Sleep Efficiency also differed between groups, and again contrary to Hypothesis 1, the TD group showed lower efficiency (mean = 78.6%, *SD* = 3.5) than the LD group (mean = 81.3%, *SD* = 4.1), *t* (37.00) = -2.28, *p* = 0.032. Night-to-night variability in Sleep Efficiency was equivalent across groups (TD mean = 5.5, *SD* = 1.8; LD mean = 5.3, *SD* = 2.5; *t* (34.73) = 0.34. Finally, a group difference in bedtime anxiety fell just short of significant in the anticipated direction, *W* = 148, *p* = 0.055. For the TD group, mean response on the three point scale was 1.10 (*SD* = 0.3), while for the LD group, mean response was 1.45 (*SD* = 0.69).

Previous actigraphy estimates (Acebo et al., 2005) for typically developing children aged 60 months (5;0 years) have shown a total sleep duration of 8.6 hours for girls (516 minutes) and 8.9 hours for boys (534 minutes) with a standard deviation of 48 minutes for both; and sleep efficiency estimates of 88.6% (*SD* = 4.5%) for girls, 87.9% (*SD* = 4.9%) for boys. The TD group in the current sample showed Sleep Duration in line with this previous estimate, though Sleep Efficiency fell below -*1SD* of the previous estimate.

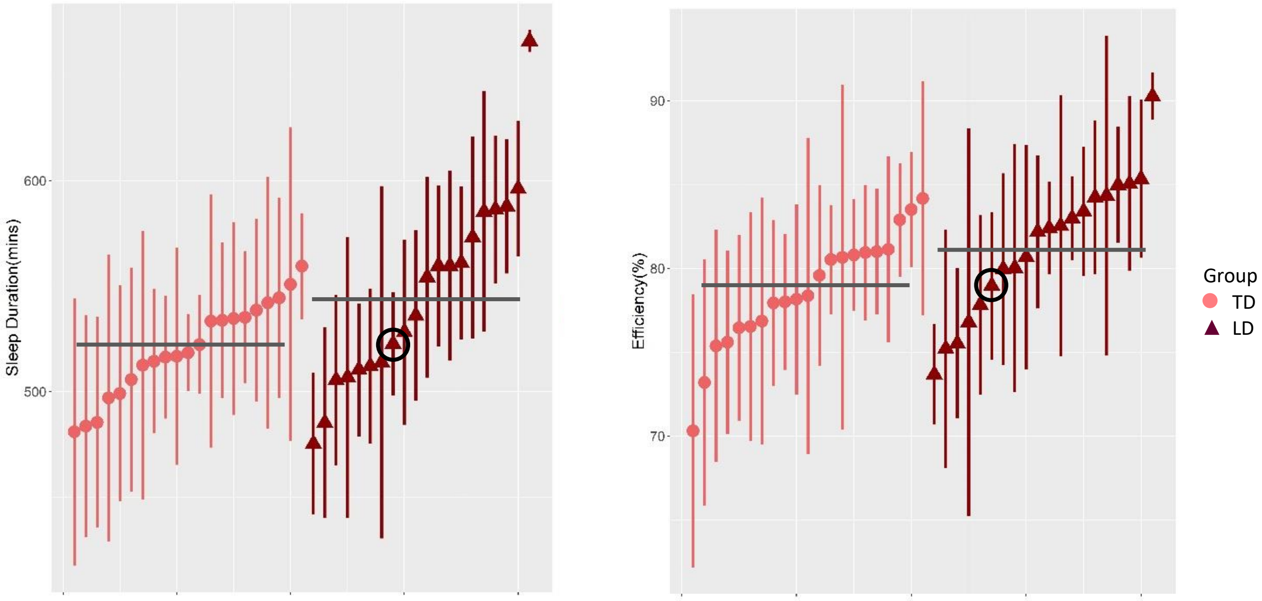


Figure 4. ***Individual mean and night-to-night variability (standard deviation) and group means for a) Sleep Duration and b) Sleep Efficiency for the typically developing group (TD) and the language disordered group (LD). Grey bars indicate the mean for each estimate and the individual circled in black was the only participant in the study to be taking melatonin at the time of testing. The error bars show standard deviation in each direction.***

Follow-on Exploratory Analyses. In our confirmatory analysis, mean night-to-night variability (intra-individual variability) in Sleep Duration did not differ between the groups. However, the summary statistics suggest that the LD group showed more inter-individual variability in mean values. We evaluated this possibility with an F-test for equality of variance, which supported the notion of more variability in mean Sleep Duration within the LD group than the TD group, F (19/19) = 0.23, p = 0.002. By contrast, variability in mean Efficiency did not differ between groups (F (19/19) = 0.72).

Confirmatory Linear Regressions: Hypothesis 2. Linear regressions were run to assess the predictive power of the Language Composite and SRS total score on objective Sleep Duration, Sleep Efficiency, and Sleep Onset Latency, as anticipated in Hypothesis 2. For each of these models, variance inflation factors were above 10 when the interaction term was included in the model (Language Composite VIF = 33.9, SRS VIF = 31.3, Language Composite*SRS VIF = 25.3), while variance inflation factors with the interaction term removed were acceptable (Language Composite VIF = 1.6, SRS VIF = 1.6). This possibility was anticipated in our pre-registration as the language composite significantly correlated with SRS total score (r(37)= -.55, p<0.001). Consequently, no models are presented with the interaction term included.

No models significantly predicted sleep parameters: for mean Sleep Duration, F(36,2,) = 2.67, p = 0.083; for mean Sleep Efficiency, F(36,2) = 1.47, p = 0.244; for mean Sleep Onset Latency, F(36,2) = 0.365, p = 0.697. Finally, the prediction of bedtime anxiety was assessed via ordinal logistic regression, and again no significant predictors emerged, although SRS approached significance: Language Composite odds ratio = 0.99 (97.5% CI: 0.93 – 1.05), and SRS odds ratio = 1.09 (97.5% CI: 1.00 – 1.22).

Further Exploratory Analyses

Confirmatory analyses for Study 2 broadly failed to support our hypotheses; we therefore ran a series of exploratory analyses in order to better understand these data and develop new hypotheses moving forward.

We were interested to explore differences between parent-reported sleep behaviour and objective sleep estimates of duration and efficiency in children with language disorder. The relationship between good parent-reported language (CCC General) and good parent-reported sleep behaviour (CSHQ Total) that we saw in Study 1 held in the sub-sample of children who completed Study 2: r(38) = -0.45, p = 0.003, so those with better parent-reported language also had better subjective sleep behaviour. We then considered the relationship between CCC General and objective estimates of Sleep Duration, which fell short of significance (r(38) = -0.28, p = 0.083), and Sleep Efficiency, which showed a negative correlation, r(38) = -0.32, p = 0.047. So children with poorer parent-reported general language scores slept more efficiently according to objective data.

The CSHQ seems to capture aspects of sleep behaviour that are unrelated to objective estimates of sleep quantity (Markovich et al., 2014). Our results suggest that parents of children with language difficulties show concern regarding the sleep behaviour of their children over and above what would be expected given estimates of sleep quantity (both subjectively and objectively estimated). We therefore considered group differences in parental anxiety about children's sleep. In the Sleep History questionnaire we asked parents *Are you currently worried about your child's sleep?* and *Were you worried about your child's sleep when they were younger?*. Running the same ordinal logistic models for these variables as we did to consider children's bedtime anxiety, it emerged that current parental concern about sleep was predicted by high SRS total score (Language Composite odds ratio = 0.99 (97.5% CI: 0.94 – 1.06, p = 0.844), and SRS odds ratio = 1.11 (97.5% CI: 1.01 – 1.24, p = 0.050)), while previous concern was predicted by low language composite (Language Composite odds ratio = 0.957(97.5% CI: 0.913 – 0.997, p = 0.043), and SRS odds ratio = 0.977 (97.5% CI: 0.909 – 1.046, p = 0.504)). Parental concern about children's sleep was more likely in the past if the child currently has language difficulties, while parental concern about current sleep is more likely if the children shows autism symptomology.

Finally, we needed to establish whether the relatively good objective measures of sleep duration and efficiency shown in the LD group were due to more of that group being tested during the COVID-19 pandemic lockdown. To check this, we split the LD group into those who had been tested before lockdown (n = 8) and those who were tested during lockdown (n = 12). Neither objective mean Sleep Duration (t(17.9) = -0.89), nor objective mean Sleep Efficiency (t(16.8) = -1.63, p = 0.123) differed between groups.

General Discussion

The aim of this project was to test the hypothesis that sleep may be atypical in children who have developmental difficulties in the language domain. In Study 1, an exploratory analysis was conducted of subjective, parent-reported data concerning the sleep and language abilities of 4-10 year old children using the CSHQ and CCC-2 questionnaires. In agreement with Botting and Baraka (2017), our analysis indicated that poor sleep behaviour was associated with poor language development, but we extended the previous work to show that this relationship exists when no children with a diagnosis or suspected diagnosis of autism are included in the analysis. The better children's general language ability was reported to be, the better also their reported sleep behaviour with respect to Sleep Duration, Sleep Anxiety, Night Wakings, and Parasomnias (Sleep Disordered Breathing and Daytime Sleepiness were also associated with language ability but did not survive Bonferroni-Holme correction). Furthermore, overall CSHQ score was predicted by whether or not children were described by their parents as having a difficulty with language development. The only measures from the CSHQ to indicate anything other than a positive relationship between sleep behaviour and language skill were parents' numerical estimates of bed time and sleep duration, where, unexpectedly, *better* language skill was associated with *later* bed time and *less* overall sleep.

We took these data forward to pre-register Study 2, in which a sub-sample of 4-to-6 year old children from Study 1 with clinical language deficits, along with age matched peers, wore an actigraphy watch for 5-10 nights. Here, contrary to our hypotheses (but consistent with the subjective estimates of bed time and total overall sleep from Study 1), objective sleep estimates were negatively related to objectively measured language ability, again suggesting that those with language deficits actually slept for longer and more efficiently than their language-typical peers. So while weaker language skills are associated with parent-reported negative sleep behaviours (such as anxiety and night wakings), at the same time, both subjective and objective estimates of actual sleep episodes suggest that weaker language skills are associated with longer sleep duration and higher sleep efficiency.

Parents of children with more language difficulties reported a high degree of concern about their child's sleep, beyond what would be anticipated given objective estimates of sleep. This pattern of seeing more severe or broad difficulties with sleep in subjective parent-report compared to objective measures, has also been seen in the case of ADHD (Chin et al., 2018), ASD (see Díaz-Román et al., 2018), and visual impairment (Hayton et al., 2021). This pattern suggests that measures like the CSHQ are recording something quite different, and complementary, to actigraphy-derived objective sleep patterns.

One reason for heightened parental concern might be children's sleep history. In Study 2, the likelihood of parents reporting current concern about their child's sleep was positively predicted by autism symptomology, but the likelihood of parents reporting that they were concerned about their child's sleep in the past was predicted by language ability. Four parents of typically developing children expressed some degree of concern about their child's sleep in the past compared to ten parents from the language disordered group – all but one (who reported apnoea) said their child struggled to initiate and maintain sleep as infants and did not sleep through the night until at least 18 months. For example, one parent of a child with language disorder reported *'From about 5 months up until 18 months, would wake between midnight and 3am and would not return to sleep until about 6/ 7 am.'* The finding that language scores only predicted the extent of past parental concern about their child's sleep may speak to the complex and temporally extended nature of parental perceptions of sleep.

Parents highlighted some areas of difficulty that were not possible to assess with actigraphy. Sleep disordered breathing was more likely to be reported in children with poor language in Study 1, and in Study 2 this was the only area of the CSHQ were the language disordered group differed significantly from their typically developing peers. Sleep disordered breathing has been associated with deficits in both phonology and vocabulary skill (see de Castro Corrêa et al., 2017; Mohammed et al., 2021), and may be a contributing factor to the aetiology of language difficulty as experienced by a sub-group of children. Sleep disordered breathing may affect language development either by reducing the quantity of sleep children get and thereby resulting in daytime sleepiness, and/or by disrupting night-time sleep architecture resulting in a poverty of consolidation opportunities. Although we did not see disrupted (less efficient) sleep in our smaller clinical sample in Study 2, sleep disordered breathing can result in changes to sleep architecture without necessarily interrupting sleep efficiency (Shahveisi et al., 2018), and actigraphy is not thought to be a good indicator of the sleep fragmentation seen in sleep disordered breathing (O'Driscoll et al., 2010).

Unexpectedly, we saw with both subjective and objective data that children with language disorders actually slept for longer and more efficiently than their typically developing peers. A possible interpretation of this finding is that the maturation of the sleep cycle might be generally delayed in the language disordered group relative to age matched peers. Sleep duration, efficiency, and global sleep patterns (Iglowstein et al., 2003) change gradually through infancy and childhood, with the amount of sleep needed over a 24 hour period decreasing, and with night-time sleep getting more efficient (Acebo et al., 2005). Infants who go on to demonstrate lower language ability show immature sleep relative to peers with better language (Dionne et al., 2011; Knowland et al., 2021; Smithson et al., 2018), that is, more of their overall sleep occurs as naps during the day. If we see a continuation of delayed sleep maturation by the early school years, then what looks like better sleep could be construed as less mature sleep. In our data, longer night-time sleep duration could indicate a higher need for sleep in the context of less opportunity for day time napping (given the age of the children). The group effect of increased efficiency in the language disordered sample is more difficult to explain as sleep typically gets more efficient over developmental time (Acebo et al., 2005). This group effect may have emerged because the typically developing children included in Study 2 showed unusually low sleep efficiency. Alternatively, it could be reflective of the language disordered children needing more sleep over 24 hours, given that when habitually napping pre-school children miss a nap their subsequent night-time sleep is both longer and more efficient compared to a typical night (Lassone et al., 2016).

Longitudinal work with young children showing early language delay would allow an analysis of trajectories of change in sleep behaviour. Such trajectories should consider changes in parental evaluation of, and feelings about, their child's sleep, alongside objective measures. Both subjective and objective measures are highly informative but are not equivalent. This seems to be especially true in groups of children with neurodevelopmental disorders. As this story unfolds it is likely to reveal a dynamic interaction between multiple factors including neural maturation, behavioural manifestation of disorder, parental sensitivity to child development and the perceptions of the child themselves around sleep.

It should also be noted that even if sleep behaviour is unremarkable in children with language disorders, that does not guarantee that sleep performs the same functions in these children that it does in children who are developing as expected in the language domain (Earle et al., 2017). Future studies in this area therefore need to consider both the nature of sleep and the role that sleep plays in supporting language development over time in different developmental populations.

Limitations

The success of this work should naturally be evaluated within the context of its limitations. The size of the sample, particularly in Study 2, was limited. There are myriad influences on language development, and children present with profiles of strength and weakness across multiple dimensions. This heterogeneity in symptomology and aetiology means it is challenging to draw conclusions that can be extended beyond the current sample. The non-equality of variance in sleep duration seen across our groups here may well reflect that aetiological heterogeneity.

It should be noted that the summary statistics for the typically developing and language disordered groups in Study 2 both demonstrated relatively poor subjective sleep according to the CSHQ. The clinical threshold for concern on the CSHQ is a sum score of 41. Here, 66% of the LD group exceeded this threshold, as did 50% of the TD group, compared to 23% of the control group in the original description of the measure (Owens et al., 2000). This suggests that the TD group in Study 2 may experience more sleep-related difficulties than are typically observed in the general population, as supported by the lower than expected sleep efficiency for the TD group based on actigraphy data. This possibly reflects a sampling bias where parents whose children experience sleep difficulties are more likely to volunteer for sleep studies.

Testing for this study was interrupted by the COVID-19 pandemic and resulting national UK lockdown in 2020. We did not find an effect of lockdown on either sleep duration or efficiency for children in the language disordered group, and we have demonstrated elsewhere that sleep duration was not interrupted in children over the UK lockdown (Knowland et al., in press). We are therefore confident in our results, but of course the circumstances must be taken into account. In summary, this project is a starting point; the work should be replicated with a larger sample in less interesting times.

Summary & Conclusions

The aim of this paper was to investigate whether children with poor structural language development exhibit poor sleep and sleep behaviour. Over two studies we saw that children who had worse parent-reported language abilities also showed worse parent-reported sleep behaviours, such as more sleep anxiety and more night waking. Conversely, in both subjective and objective estimates of sleep duration, children with language disorder slept for longer and also more efficiently than their language-typical peers. Given that a weak relationship between objective estimates of sleep and the CSHQ has been shown before (Hayton et al., 2021; Markovich et al., 2014), we suggest that subjectively reported sleep behaviour and objective sleep estimates be thought of as complementary, together building a complete picture of the behavioural, cognitive and emotional components of sleep in young children. It is clear that the dynamic relationships between sleep and language are relevant not only to children's development but also the wider picture of family functioning and parental concern, and as such this is a topic that deserves further careful consideration.

References

Acebo, C., Sadeh, A., Seifer, R., Tzischinsky, O., Hafer, A., & Carskadon, M. A. (2005). Sleep/wake patterns derived from activity monitoring and maternal report for healthy 1- to 5- year old children. *Sleep, 28* (12): 1568-1577. doi: 10.1093/sleep/28.12.1568

Acebo, C., Sadeh, A., Seifer, R., Tzischinsky, O., Wolfson, A. R., Hafer, A., & Carskadon, M. A. (1999). Estimating sleep patterns with actigraphy monitoring in children and adolescents: how many nights are necessary for reliable measures? *Sleep, 22* (1), 95-103.

Anders, T. F., Iosif, A-M., Schwichtenberg, A. J., Tang, K., & Goodlin-Jones, B. L. (2011). Six-month sleep-wake organisation and stability in pre-school age children with autism, developmental delay, and typical development. *Behavioural Sleep Medicine, 9*(2): 92-106. doi: 10.1080/15402002.2011.557991

Axelsson, E. L., Williams, S. E., & Horst, J. (2016). The effect of sleep on children's word retention and generalization. *Frontiers in Psychology, 7,* 1192. doi: 10.3389/fpsyg.2016.01192

Backhaus, J., Hoeckesfeld, R., Born, J., Hohagen, F., Junghanns, K., (2008). Immediate as well as delayed post learning sleep but not wakefulness enhances declarative memory consolidation in children. *Neurobiology of Learning and Memory, 89,* 76–80. doi: 10.1016/j.nlm.2007.08.010

Bélanger, M-E., Bernier, A., Paquet, J., Simard, V., & Carrier, J. (2013). Validating actigraphy as a measure of sleep for preschool children. *Journal of Clinical Sleep Medicine, 9* (7): 701-706. doi: 10.5664/jcsm.2844

Bishop, D.V.M. (1998) Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry, 39*(6), 879-893.

Bishop, D. V. M. (2003). *The Children's Communication Checklist, 2nd Edition*. London: Pearson.

Bishop, D., Snowling, M., Thompson, P., Greenhalgh, T., & the CATALISE consortium. (2016). Phase 2 0f CATALISE: a multinational and multidisciplinary delphi consensus study: Terminology. *Journal of Child Psychology and Psychiatry, 58* (10): 1068-1080. doi: 10.1111/jcpp.12721

Botting, N., & Baraka, N. (2017). Sleep behaviour relates to language skills in children with and without communication disorders. *International Journal of Developmental Disabilities,* doi:10.1080/20473869.2017.1283766

Brown, T.E. (2001). *Brown Attention-Deficit Disorder Scales (Brown ADD Scales).* Pearson, London, UK.

Buckhalt, J. A., El-Sheikh, M., Keller, P. S., & Kelly, R. J. (2009). Concurrent and longitudinal relations between children's sleep and cognitive functioning: The moderating role of parent education. *Child Development, 80* (3), 875-892. doi: 10.1111/j.1467-8624.2009.01303.x

Chin, W-C., Huang, Y-S., Chou, Y-H., Wang, C-H., Chen, K-T., Hsu, J.F., & Hsu, S-C. (2018). Subjective and objective assessments of sleep problems in children with attention deficit/hyperactivity disorder and the effects of methylphenidate treatment. *Biomedical Journal, 41* (6): 356-363. doi: 10.1016/j.bj.2018.10.004

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale–Second Edition (SRS-2).* Torrance, CA: Western Psychological Services.

Corrêa, C., Cavalheiro, M.B., Weber, S.A.T., & Maximino, L.P. (2017). Obstructive sleep apnea and oral language disorders. *Brazilian Journal of Otorhinolaryngology, 83*(1): 98-104. doi:10.1016/j.bjorl.2016.01.017

Carotenuto, M., Esposito, M., Cortese, S., Laino, D., & Verrotti, A. (2016). Children with developmental dyslexia showed greater sleep disturbances than controls, including problems initiating and maintaining sleep. *Acta Pediatrica, 105*: 1079-1082. doi:10.1111/apa.13472

Díaz-Román, A., Zhang, J., Delorme, R., Beggiato, A., & Cortese, S. (2018). Sleep in youth with autism spectrum disorders: systematic review and meta-analysis of subjective and objective studies. *Evidence Based Mental Health, 21* (4): 146-154. doi: 10.1136/ebmental-2018-300037

Dionne, G., Touchette, E., Forget-Dubois, N., Petit, D., Tremblay, R. E., Montplaisir, J.Y., & Boivin, M. (2011). Associations between sleep-wake consolidation and language development in early childhood: A longitudinal twin study. SLEEP, 34(8), 987-995. doi: 10.5665/SLEEP.1148

Dayyat, E.A., Spruyt, K., Molfese, D.L., & Gozal, D. (2011). Sleep estimates in children: parental versus actigraphic assessments. *Nature and Science of Sleep, 3*: 115-123. doi: 10.2147/NSS.S25676

Dlouha, O., Prihodova, I., Skibova, J., & Nevsimalova, S. (2020). Developmental Language Disorder: Wake and Sleep Epileptiform Discharges and Co-morbid Neurodevelopmental Disorders. *Brain sciences, 10*(12), 910. doi: 10.3390/brainsci10120910

Dunn, L. M., Dunn, D. M., Styles, B., & Sewell, J. (2009). *The British Picture Vocabulary Scale III – 3rd Edition*. London: GL Assessment

Earle, F. S., Landi, N., & Myers, E. B. (2017). Sleep duration predicts behavioural and neural differences in adult speech sound learning. *Neuroscience Letters, 636:* 77–82. doi: 10.1016/j.neulet.2016.10.044

Echenne, B., Cheminai, R., Rivier, F., Negre, C., Touchon, J., & Billiard, M. (1992). Epileptic electroencephalographic abnormalities and developmental dysphasias: a study of 32 patients. *Brain and Development, 14*(4) : 216-225. doi:10.1016/SO387-7604(12)80233-6

Elliott, C. D., & Smith, P. (2011). *The British Ability Scales, 3rd Edition.* London: GL Assessment

Elrod, M.G., & Hood, B.S. (2015). Sleep differences among children with autism spectrum disorders and typically developing peers: a meta-analysis. *Journal of Developmental and Behavioural Pediatrics, 36* (3): 166-177. doi: 10.1097/DBP.0000000000000140

Fabbro, F., Zucca, C., Molteni, M., & Borgatti, R. (2000). EEG abnormalities during slow sleep in children with developmental language disorders. *SAGGI-Neuropsicologia Infantile Psicopedagogia Riabilitazione, 26*(1), 41-48. doi: 10.7860/JCDR/2015/13920.6168

Fletcher, F., Knowland, V. C. P., Walker, S., Gaskell, M. G., Norbury, C., & Henderson, L-M (2019). Atypicalities in sleep and semantic consolidation in autism. *Developmental Science,* doi: 10.1111/desc.12906.

Hayton, J., Marshall, J., & Dimitriou, D. (2021). Lights out: examining sleep in children with vision impairment. *Brain Science, 11*(4): 421. doi: 10.3390/brainsci11040421

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, G. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science, 15* (5), 674-687. doi: 10.1111/desc.12639

Horváth, K, Myers, K, Foster, R., & Plunkett, K. (2015). Napping facilitates word learning in early lexical development. *Journal of Sleep Research, 24* (5): 503-509. doi: 10.1111/jsr.12306.

Humphreys, J.S., Gringras, P., Blair, P.S., Scott, N., Henderson, J., Fleming, P.J., & Emond, A.M. (2014). Sleep patterns in children with autistic spectrum disorders: a prospective cohort study. *Archives of Disease in Childhood, 99*:114-118. doi: 10.1136/archdischild-2013-304083

Hupbach, A., Gómez, R. L., Bootzin, R. R., & Nadel, L. (2009). Nap-dependent learning in infants. *Developmental Science, 12* (6): 1007-1012. doi: 10.1111/j.1467-7687.2009.00837.x.

Iglowstein, I., Jenni, O. G., Molinari, L., & Largo, R. H. (2003). Sleep duration from infancy to adolescence: reference values and generational trends. *Pediatrics, 111* (2): 302–307. doi: 10.1542/peds.111.2.302.

Knowland, V.C.P, Berens, S., Gaskell, M, G., Walker, S. A., & Henderson, L-M. (2021). Does the maturation of early sleep patterns predict language ability at school entry? A Born in Bradford study. *Journal of Child Language,* 1–23. doi:10.1017/S0305000920000677

Knowland, V.C.P., Fletcher, F., Henderson, L-M., Walker, S., Norbury, C., & Gaskell, M.G. (2019). Sleep promotes phonological learning in children across language and autism spectra. *Journal of Speech, Language and Hearing Research,* doi.org/10.1044/2019_JSLHR-S-19-0098

Knowland, V.C.P., van Rijn, E., Gaskell, M.G., & Henderson, L-M. (*in press*). UK children's sleep and anxiety during the COVID-19 pandemic. In press, *BMC Psychology.*

Kurdziel, l., Duclos, K., & Spencer, R. M. (2013). Sleep spindles in midday naps enhance learning in preschool children. *Proceedings of the National Academy of Sciences of the United States of America, 110* (43): 17267-17272. doi:10.1073/pnas.1306418110.

Lam, J. C., Mahone, E. M., Mason, T., & Scharf, S. M. (2011). The effects of napping on cognitive function in pre-schoolers. *Journal of Developmental Behavioral Pediatrics, 32,* 9-97. doi: 10.1097/DBP.0b013e318207ecc7

Lassone, J. M., Rusterholz, T., Kurth, S., Schumacher, A. M., Achermann, P., & LeBourgeois, M. K. (2016). Sleep physiology in toddlers: Effects of missing a nap on subsequent night sleep. *Neurobiology of Sleep and Circadian Rhythms.* 1: 19–26. doi: 10.1016/j.nbscr.2016.08.001

Marini, A., Ozbič, M., Magni, R., & Valeri, G. (2020). Toward a definition of the linguistic profile of children with Autism Spectrum Disorder. *Frontiers in Psychology, 11*: 808. doi: 10.3389/fpsyg.2020.00808. eCollection 2020.

Markovich, A.N., Gendron, M.A., & Corkum, P.V. (2014). Validating the Children's Sleep Habits Questionnaire against polysomnography and actigraphy in school aged children. *Frontiers in Psychiatry, 6* (5): 188. doi: 10.3389/fpsyt.2014.00188

Mehta, T.R., Gurung, P., Nene, Y., Fayyaz, M., & Bollu, P.C. (2019). Sleep and ADHD: A review article. *Current Developmental Disorders Reports, 6*: 228–234. doi:10.1007/s40474-019-00178-6

Meltzer, L. J., Montgomery-Downs, H. E., Insana, S. P., Walsh, C. M., (2012). Use of actigraphy for assessment in pediatric sleep research. *Sleep Medicine Review, 16*(5): 463–475. doi:10.1016/j.smrv.2011.10.002.

Meltzer, L.J., Walsh, C.M., Traylor, J., & Westin, A.M.L. (2012). Direct comparison of two new actigraphs and polysomnography in children and adolescents. *Sleep, 35* (1): 159-166. doi: 10.5665/sleep.1608

Mohammed, D., Park, V., Bogaardt, H., & Docking, K. (2021). The impact of childhood obstructive sleep apnea on speech and oral language development: a systematic review. *Sleep Medicine, 81*: 144-153. doi: 10.1016/j.sleep.2021.02.015

Norbury, C.F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & 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 & Psychiatry, 57* (11): 1247-1257. doi: 10.1111/jcpp.12573

O'Driscoll, D.M., Foster, A.M., Davey, M.J., Nixon, G.M., & Horne, R.S.C. (2010) Can actigraphy measure sleep fragmentation in children? *Archives of Disease in Childhood, 95* (12):1031-1033. doi: 10.1136/adc.2009.166561

Overvliet, G.M., Besseling, R.M.H., Vles, J.S.H., Hofman, P.A.M., van Hall, M.H.J.A., Backes, W.H., & Aldenkamp, A.P. (2011). Association between frequency of nocturnal epilepsy and language disturbance in children. Pediatric Neurology, 44(5): 333-339. doi: 10.1016/j.pediatrneurol.2010.10.014

Owens, J. A., Spirito, A., & McGuinn, M. (2000). The Children's Sleep Habits Questionnaire (CSHQ): psychometric properties of a survey instrument for school-aged children. *Sleep, 23* (8), 1043-1051.

Paller, K. A., Creery, J.D., & Schechtman, E. (2021). Memory and sleep: how sleep cognition can change the waking mind for the better. *Annual Review of Psychology, 72*: 123-150. doi: 10.1146/annurev-psych-010419-050815.

Picard, A., Heraut, F. C., Bouskraoui, M., Lemoine, M., Lacert, P., & Delattre, J. (1998). Sleep EEG and developmental dysphasia. *Developmental Medicine & Child Neurology, 40*(9), 595-599. doi: 10.1111/j.1469-8749.1998.tb15424.x

Quach, J., Hiscock, H., Canterford, L., & Wake, M. (2009). Outcomes of child sleep problems over the school-transition period: Australian population longitudinal study. *Pediatrics, 123* (5): 1287-1292. doi: 10.1542/peds.2008-1860

Reynolds, A.M., Soke, G.N., Sabourin, K.R., Hepburn, S., Katz, T., Wiggins, L.D., Schieve, L.A., & Levy, S.E. (2019). Sleep problems in 2- to 5-year-olds with Autism Spectrum Disorder and other developmental delays. *Pediatrics, 143* (3):e20180492. doi: 10.1542/peds.2018-0492

Semel, E., & Wiig, E. H. (2017). *Clinical Evaluation of Language Fundamentals 5th UK Edition* (CELF-5UK). Pearson Education Ltd, Pearson, London: UK.

Shahveisi, K., Jalali, A., Moloudi, M.R., Moradi, S., Maroufi, A., & Khazaie, H. (2018). Sleep architecture in patients with primary snoring and obstructive sleep apnea. Basic Clinical Neuroscience, 9(2): 147-156. doi: 10.29252/NIRP.BCN.9.2.147

Sitnick, S.L., Goodlin-Jones, B.L., & Anders, T.F. (2008). The use of actigraphy to study sleep disorders in preschoolers: some concerns about detection of nighttime awakenings. *Sleep, 31*: 395e401. doi: 10.1093/sleep/31.3.395

Smithson, L., Baird, T., Tamana, S. K., Lau, A., Mariasine, J., Chikum, J., Lefebvre, D. L., Subbarao, P., Becker, A. B., Turvey, S. E., & Sears, M. R. (2018). Shorter sleep duration is associated with reduced cognitive development at two years of age. *Sleep Medicine, 48*: 131-139. doi : 10.1016/j.sleep.2018.04.005

Touchette, E., Petit, D., Séquin, J. R., Boivin, M., Tremblay, R. E., & Montplaisir, J. Y. (2007). Associations between sleep duration patterns and behavioural/cognitive functioning at school entry. *Sleep, 30* (9), 1213-1219. doi: 10.1093/sleep/30.9.1213

Verhoeff, M.E., Blanken, L.M.E., Kocevska, D., Mileva-Seitz, V.R., Jaddoe, V.W.V., White, T., Verhulst, F., Luijk, M.P.C.M., & Tiemeier, H. (2018). The bidirectional association between sleep problems and autism spectrum disorder: a population-based cohort study. *Molecular Autism 9*:8. doi:10.1186/s13229-018-0194-8

Wagner, R., Torgesen, J., Rashotte, C., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing 2nd Edition* (CTOPP2). Pearson Education Ltd, Pearson, London: UK.

Williams, D., Botting, N., & Boucher, J. (2008). Language in autism and specific language impairment: where are the links? *Psychological Bulletin, 134*(6): 944-963. doi: 10.1037/a0013743

Data, Code and Materials Availability Statement

The dataset supporting the conclusions of this article are available on the Open Science Forum (https://osf.io/y4evw/) along with analysis scripts. A number of questionnaires and assessments were used in the construction of these data: a Sleep History questionnaire developed by the research team is available in Supplementary Materials at the end of this manuscript; the Children's Sleep Habits Questionnaire (CSHQ; Owens, Spirito & McGuinn, 2000) is available online at https://depts.washington.edu/dbpeds/Screening%20Tools/ScreeningTools.html; the following materials are copyrighted, on which grounds the editor has granted exemption from sharing (19th April, 2021):

- Children's Communication Checklist-2 (CCC-2; Bishop, 2003)
- Social Responsiveness Scale 2 (SRS; Constantino & Gruber, 2012)
- Brown Attention Deficit Disorder scales (ADD; Brown, 2001)

- British Picture Vocabulary Scale, 3rd Edition (BPVS-III; Dunn, Dunn, Styles & Sewell, 2009)
- British Ability Scale 3rd Edition (BAS-3; Elliott & Smith, 2011)
- Comprehensive Test of Phonological Processing- 2nd Edition (CTOPP-2; Wagner, Torgesen, Rashotte & Pearson, 2013)
- Clinical Evaluation of Language Fundamentals 5th Edition (CELF-5; Semel & Wiig, 2017)

Ethics Approval and Consent

Approval for this work was granted by the Department of Psychology Research Ethics Committee at the University of York. All parents gave informed, written consent, and all children gave verbal assent before any data were collected.

Authorship and Contributorship Statement

All authors designed the study, edited the manuscript and approved the final version. MR collected the data. VK analysed the data and wrote the manuscript.

Acknowledgements

The authors would like to thank all families who so generously gave their time. This work was funded by ESRC grant ref ES/N009924/1 awarded to Lisa Henderson, Gareth Gaskell and Courtenay Norbury.

Supplementary Materials

Descriptive sleep history

Q1. How old is your child (years, months; e.g., 12 years 6 months)
Years ___________ Months _________

Q2. What was your child's gender at birth?
Male
Female

Q3. Do they identify with a different gender now?
Yes
No

Q4. What is your child's main language? _______________________________

Q5. Does your child speak any other languages as well as they speak their main language? ________________________________

Q6. In total, how many children (0-18) live in your household? ___________

Q7. Of these, how many are older than the child you are filling in this questionnaire about? __________

Q8. What is the highest educational qualification achieved by someone in your child's household? ____

Q9. Does your child have difficulties with language development?

Yes
No
I'm not sure

Q10. Does your child receive support for their language development at school, or have they received support in the past? *{Asked if Yes or I'm not sure in response to Q9}*

____

Q11. Does your child see a speech and language therapist to support their language development, or have they seen one in the past? *{Asked if Yes or I'm not sure in response to Q9}* ____

Q12. Please describe your child's language difficulties and what their diagnosis is, if they have one. (Your description here might include whether your child has difficulties with understanding spoken language and/or with speaking, and whether they have a diagnosis such as Developmental Language Disorder.) *{Asked if Yes or I'm not sure in response to Q9}* ____

Q13. Does your child have a diagnosis, or possible diagnosis, of any other developmental disorders?

ASD
ADHD
Developmental Co-ordination Disorder
Dyslexia
Other ___________
None

Q14. Are you currently worried about your child's sleep?
Yes
No
Somewhat

Q15. Please tell us what currently concerns you about your child's sleep: *{Asked if Yes or Somewhat in response to Q14}* ___________________________________

Q16. Have you ever sought support for your child's sleep from a GP or other health professional? *{Asked if Yes or Somewhat in response to Q14}*
Yes
No

Q17. Were you worried about your child's sleep when they were younger?
Yes
No
Somewhat

Q18. Please tell us why you were worried about your child's sleep when they were younger, and how old your child was when their sleep was a concern: *{Asked if Yes or Somewhat in response to Q17}* ____

Q19. Have you ever sought support for your child's sleep from a GP or other professional? *{Asked if Yes or Somewhat in response to Q17}*__________________________

Q20. What does a good night of sleep look like for your child?
____

Q21. How many times a week do you typically see this pattern of good sleep?
____

Q22. What does a bad night of sleep look like for your child?
____

Q23. How many times a week do you typically see this pattern of bad sleep?
____

Q24. Does your child currently take daytime naps?
Yes
No

Q25. How many times a day does your child usually nap? *{Asked if Yes in response to Q24}*_________

Q26. When your child naps in the day how long do they usually sleep for? *{Asked if Yes in response to Q24}*_________

Q27. How many days a week does your child nap? *{Asked if Yes in response to Q24}*

____

Q28. At what age did your child stop napping in the day? Please tell us in Years and Months if you can (it might help to remember if it was linked with an event like starting nursery) _________

Q29. Does your child get anxious about going to bed at night?

Yes

Somewhat

No

Q30. Can you describe your child's bedtime routine? This might start when they have a bath, watch a special TV programme or when you ask them to go to bed.

____

Q31. How long does it take from when you start this routine to when your child falls asleep?

____

Q32. Once you've left your child's bedroom, how many times do you typically have to go back to their room or put them back to bed before they fall asleep?

____

	UnWeighted					**Weighted**				
	B	**Lower 95% CI**	**Upper 95% CI**	***t***	***p***	**B**	**Lower 95% CI**	**Upper 95% CI**	***t***	***p***
Intercept	47.98	42.72	53.24	17.866	<0.001***	44.96	40.03	49.88	17.879	<0.001***
LD group	2.72	0.15	5.29	2.071	0.040*	3.21	0.98	5.44	2.820	0.005***
Age	-0.07	-0.11	-0.02	-2.664	0.008**	-0.04	-0.08	0.01	-1.530	0.128
Sex	1.49	-0.75	3.74	1.305	0.194	2.05	-0.24	4.34	1.756	0.081
Attention	9.63	0.68	18.59	2.108	0.036*	8.90	-0.21	18.01	1.915	0.057
Literacy	1.86	-2.76	6.48	0.790	0.431	2.93	-1.30	7.15	1.359	0.176
Social	7.65	0.48	14.83	2.091	0.038*	6.72	-1.22	14.66	1.659	0.099

Table SM1. ***Unweighted and weighted regression models predicting CSHQ Total scores. *p<0.05, ***p<0.001. Unweighted model: F (6,189) = 4.40, p < 0.001; Weighted model: F (11.2, 189) = 3.213, p = 0.005.***

		Actigraphy measures					
		Sleep Duration (mins)	**Sleep Efficiency (%)**	**Sleep on-set latency (mins)**	**Average activity/ min**	**Bed time**	**Get-up time**
Parent-report CSHQ	**Bedtime resistance**					*r* (38)=-.09, *p*=.585	
	Sleep onset delay			*r*(38)=.46, *p*=.003			
	Sleep duration	*r* (38) = .08, *p*=.607					
	Sleep anxiety			*r*(38)=-.12, *p*=.460			
	Night wakings		*r*(38)=-.05, *p*=.768				
	Parasomnias				*r*(38)=.012, *p*=.941		
	SDB		*r*(38)=.24, *p*=.129				
	Daytime sleepiness	*r*(38)=.20, *p*=217	*r* (38)=.181, *p*=.264				
	Total score	*r*(38)=.12, *p*=.478	*r*(38)=.20, *p*=.215	*r*(38)=-.11, *p*=.491	*r*(38)=.05, *p*=.784	*r*(38)=-.08, *p*=.627	*r*(38)=.24, *p*=.132
	Bed-time					*r*(36)=.35, *p*=.033	
	Get-up time						*r*(30)=60, *p*<0.001
	Sleep duration (mins)	r(30)=.49, p=.004					

Table SM2. ***Pearsons correlations between parent-reported Child Sleep Habits Questionnaire (CSHQ) responses and actigraphy-derived measures of total sleep time, efficiency, bed time and get up times. Correlations are reported for theoretically relevant associations.***

		Cognitive battery (standardised scores)				
		BAS-3 Naming	**BPVS-2**	**CELF-5 Recalling Sentences**	**CTOPP-2 Non-word Repetition**	**BAS-3 non-verbal measure**
Parent report CCC-2	**CCC General**	*r* (37) = 0.58, *p*<0.001	*r*(36) = 0.50, *p* = 0.001	*r*(36) = 0.69, *p*<0.001	*r*(36) = 0.72, *p* <0.001	*r*(37) =0.36, *p* = 0.025
	CCC Social	*r*(37)=-0.56, *p*<0.001	*r*(36)=-0.49, *p*= 0.002	*r*(36)=-0.64, *p*<0.001	*r*(36)=-0.69, *p*<0.001	*r*(37)=-0.27, *p*= 0.095

Table SM3. ***Pearsons correlations between parent-reported Children's Communication Checklist 2nd Edition responses and standardized cognitive assessments. British Ability Scales 3rd Edition (Naming); British Picture Vocabulary Scale 2nd Edition; Clinical Evaluation of Language Fundamentals 5th UK Edition (Recalling sentences); Comprehensive Test of Phonological Processing 2nd Edition (Non-word Repetition); British Ability Scale 3rd Edition (Matrices or Block Design).***

License