UNIVERSITY OF CALIFORNIA

Santa Barbara

The Syntactic Bits of Nouns: How Prior Syntactic Distributions Affect Comprehension, Production, and Acquisition

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Linguistics

by

Nicholas Andrew Lester

Committee in charge:

Professor Fermín Moscoso del Prado Martín, Chair

Professor John W. Du Bois

Professor Stefan Th. Gries

Professor René Weber

June 2018

The dissertation of Nicholas Andrew Lester is approved.

John W. Du Bois

Stefan Th. Gries

René Weber

Fermín Moscoso del Prado Martín, Committee Chair

May 2018

The Syntactic Bits of Nouns: How Prior Syntactic Distributions Affect Comprehension, Production, and Acquisition

by

Nicholas Andrew Lester

ACKNOWLEDGEMENTS

I must first thank the Department of Linguistics and the University of California, Santa Barbara, for their generous financial support throughout my stay at the university. I further recognize the efforts of my committee members, who worked continuously on my behalf to secure funding, research space, equipment, participant pools, research assistants, and all of the other necessities that made this research possible.

Many people have contributed to the success of this project. First, I would like to acknowledge my early mentors from the Department of Linguistics and Technical Communication at the University of North Texas: Dr. Shobhana Chelliah, Dr. John (“Háj”) Ross, and most notably, Dr. Stefanie Wulff, who steered me towards quantitative, cognitive, and corpus-based linguistic approaches, a path which has shaped my entire career since. I would also thank Bárbara Michea, Chris Armijo Ayyoub, and Sayako Itakura. This excellent team of undergraduate research assisstants worked diligently and effectively at collecting and coding data for the picture naming experiments reported in this dissertation.

Thanks also to my committee members. Dr. René Weber allowed me to cut my experimental teeth on a large-scale project in his Media Neuroscience Lab. Dr. John W. Du Bois revealed to me the vast multiplicity and complexity of natural language data, particularly naturally occurring conversation, and the many competing motivations that shape our utterances. Dr. Stefan Th. Gries was my first statistics and programming teacher. He taught me how to think scientifically and quantitatively, and also how to answer meaningful linguistic questions using large corpora of naturally occurring text. Personally, he has been a source of rational, clear-sighted guidance in my times of doubt. My chair, Dr. Fermín Moscoso del Prado Martín, taught me everything I know about psycholinguistics, natural language processing, and information theory, the three cornerstones of this

dissertation. Moreover, the basic ideas behind this dissertation stem directly from his research. Any light that I might shed finds its source in the spark of his ingenuity. At a deeper level, his patience and steadfast commitment to my growth as a scientist are perhaps the only reasons that I would even venture to apply such a label to myself.

Finally, I would like to thank my family. My parents, Leland and Cathy Lester – both credentialed linguists, as luck would have it – taught me the value and joy of learning, which has carried me through my many, many years as a student. I also thank my wife Ava, for her strength, which bore the weight of this endeavor, for her tender love, which soothed the pain of setbacks, and for her adventurous spirit, which carried us 1,500 miles across the country to make all of this possible.

VITA OF NICHOLAS ANDREW LESTER
May 2018

EDUCATION

Bachelor of Arts in English (Linguistics) and Philosophy, University of North Texas, May 2009 (magna cum laude)
Graduate Certificate in Teaching English to Speakers of Other Languages, University of North Texas, May 2012
Master of Arts in Linguistics, University of North Texas, May 2012
Doctor of Philosophy in Linguistics, University of California, Santa Barbara, June 2018 (expected)

PROFESSIONAL EMPLOYMENT

2012-2018: Teaching Assistant, Department of Linguistics, University of California, Santa Barbara
2012: Instructor of Record, LING 3060: Principles of Language Study, Department of Linguistics and Technical Communication, University of North Texas
2012: Instructor of Record, TECM 2700: Technical Writing, Department of Linguistics and Technical Communication, University of North Texas
2009-2012: Teaching Assistant, Department of Linguistics and Technical Communication, University of North Texas

PUBLICATIONS

Lester, N. A. (accepted). *That*'s hard: Relativizer use in spontaneous L2 speech. *International Journal of Learner Corpus Research*.

Lester, N. A., Auderset, S., & Rogers, P. G. (accepted). Case inflection and the functional indeterminacy of nouns: A cross-linguistic analysis. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Lester, N. A., Baum, D., & Biron, T. (accepted). Phonetic duration of nouns depends on de-lexicalized syntactic distributions: Evidence from naturally occurring conversation. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Tsai, K., Lester, N. A., & Moscoso Del Prado Martín, F. (2018). Bi-dialectal homophone effects in Kansai Japanese: An auditory lexical decision experiment. *Acoustical Science and Technology, 39*, 158-159.

Wulff, S., Gries, S. Th. & Lester, N. A. (2018). Optional *that* in complementation by German and Spanish learners. In A.Tyler & C. Moder (Eds.). *What is applied cognitive linguistics? Answers from current SLA research*(pp. 99-120). New York: de Gruyter Mouton.

Lester, N. A. (2017). Lexical Priming in Spoken English Usage. *Corpus Linguistics and Linguistic Theory, 11*, 341-361.

Lester, N. A. (2017). The Louvain International Database of Spoken English Interlanguage. *Corpus Linguistics and Linguistic Theory, 11*, 363-371.

Lester, N. A., Du Bois. J. W., Gries, S. Th., & Moscoso del Prado Martín, F. (2017). Considering experimental and observational evidence of priming together, syntax doesn't look so autonomous. *Behavioral and Brain Sciences, 40*. doi:10.1017/S0140525X17000486

Lester, N. A., Feldman, L. B. & Moscoso del Prado Martín, F. (2017). You can take a noun out of syntax…: Syntactic similarity effects in lexical priming. *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2537-2542).

Lester, N. A. & Moscoso del Prado Martín, F. (2016). Syntactic flexibility in the noun: Evidence from picture naming. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 2585-2590)*. Austin, TX: Cognitive Science Society.

Chelliah, S. and Lester, N. A. (2016). Contact and convergence in the Northeast. H. H. Hock & E. Bashir, (Eds.), *The languages and linguistics of South Asia: A comprehensive guide* (pp. 300-309). Berlin: Mouton de Gruyter.

Lester, N. A. & Weber, R. (2016). Construal level affects intuitive moral responses to narrative content. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society (pp. 1847-1852)*. Austin, TX: Cognitive Science Society.

Lester, N. A. (2015). Linguistic input overrides conceptual biases: When goals don't matter. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society.

Lester, N. A. & Moscoso del Prado Martín, F. (2015). Word order in a grammarless language: A 'small-data,' information-theoretic approach. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society.

Lester, N. A., & Moscoso del Prado Martín, F. (2015). Constructional paradigms affect visual lexical decision latencies in English. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin: Cognitive Science Society.

Wulff, S., Lester, N., and Martinez Garcia, M. (2014). *That*-variation in German and Spanish L2 English. *Language and Cognition 6* (2): 271-299.

AWARDS

Regents Special Fellowship, University of California, Santa Barbara, 2012

FIELDS OF STUDY

Major Field: Usage-based Linguistics and the Syntax-Lexis Interface

Corpus Linguistics and Second Language Acquisition with Stefan Th. Gries

Computational Linguistics and Psycholinguistics with Fermín Moscoso del Prado Martín

Discourse and Grammar with John W. Du Bois

ABSTRACT

The Syntactic Bits of Nouns: How Prior Syntactic Distributions Affect Production, Comprehension, and Acquisition

by

Nicholas Andrew Lester

Usage-based linguistic theory argues that experience is the fundamental organizing principle of language. Linguistic representations are extracted from – and continuously tuned by – probabilistic features of language use. Much psycholinguistic evidence supports this argument, particularly in the domain of lexical processing. For example, how a word is distributed across its various lexical and morphological contexts influences how quickly it is recognized and produced in isolation. Fewer studies have explored how syntactic distributions affect lexical processing, and of these, even fewer have adopted comprehensive, abstract measurements of syntax. In this dissertation, I present several new information-theoretic tools for measuring the syntactic distributions of words based on the Dependency Grammar formalism. This formalism allows me to contrast two independent dimensions of syntactic structure: hierarchical status and word order. Further, I provide a new method for teasing apart information bound to syntactic and lexical contexts. I compute these measures for nouns based on two large corpora of English.

These measures are correlated with behavior in several contexts. First, I re-analyze the noun-based trials of two previously published databases of visual lexical decision response time data, one simple and the other primed. I then turn to production, reporting two picture-naming studies. In the first, participants produce nouns in isolation. This task consitutes a stong attack on the hypothesis that syntactic distributions affect noun production; at least on its face, it does not require participants to access syntactic information in order to successfully complete the task. In a follow up, participants were asked to name the images using a syntactic frame (*the* + NAME). This task should promote syntactic access, increasing the likelihood that prior syntactic distributions should play a role. Finally, I test whether children are senstive to these syntactic distributions (based on adult speech) as they begin to produce nouns in syntactic contexts for the first time using a large, densely sampled longitudinal corpus of child speech.

Results show that isolated noun processing is affected by prior syntactic distributions in both comprehension and production. However, the specific nature of these effects differs across modalities, and in production, as a function of whether the nouns were produced in isolation or within a syntactic frame. The measures also predict the age at which nouns first emerge in the speech of children.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. Words and syntactic structures

Traditionally, linguistic theory has drawn a strict divide between what must be memorized and what can be predicted on the basis of abstract rules. The motivating principle behind such theory is the pursuit of balance between the seemingly infinite generativity of language on the one hand, and the arbitrary conventions of how meaning is mapped onto form on the other. Clearly we must memorize some aspects of a language; otherwise, there could be no cross-linguistic variation in what combinations of sounds encode what meanings. Equally clear is the fact that these arbitrary pairings of form and meaning are combined according to systematic rules or at least very strong statistical regularities. With knowledge of these rules, one can create novel combinations of memorized chunks even if no such combination has ever been produced before and yet still be perfectly understood.

Most theories now agree that lexical items must maintain direct links to syntactic structures. For some, these links are represented within the lexical entries. Each word is annotated for the set of syntactic frames which it may head (e.g., which argument structure constructions fit a given verb), along with categorical information about how it may be integrated into the frames of other words (e.g., part of speech, mass/count distinctions, and so on; Bresnan, 2001; Chomsky, 1995; Pollard & Sag, 1994). These theories tend to emphasize linguistic competence over performance, grammatical potential over actual language use (Chomsky, 1965). Probabilistic aspects of language use are seen as ancillary and derived from language-external constraints on human cognition. Other theories argue that words and syntactic structures are represented independently, but connected via direct links in a network-like mental structure (e.g., Diessel, 2015; Goldberg, 2006; Langacker, 1987). These theories emphasize the fact that words and syntactic constructions share certain critical properties. For example, both convey meaning, and in the case of partially lexicalized constructions (e.g., idioms such as KICK *the bucket*), both may contain phonological content. These theories also tend to emphasize the role of performance in structuring competence. In Diessel (2015)'s usage-based construction grammar, for example, associative links develop between the words and syntactic structures (among other components) based on our experience with language during acquisition and use. The bonds themselves mirror the categorical specifications of the theories described above; however, they are enriched by probabilistic information at several scales (e.g., single items, classes of items, and so on). Diessel's model therefore construes the lexico-syntactic space as a distributed stochastic network – one in which words are situated within a rich, hyperdimensional syntactic space. In this dissertation, I build on prior research to refine our understanding of (i) how these syntactic spaces are structured, (ii) how to measure the information carried by these networks, and (iii) how this information impacts lexical comprehension and production in adults, as well as lexical acquisition in young children.

A. From distributional learning to language processing

Infants learning a language are confronted with a significant problem, what William James referred to as a “blooming, buzzing confusion” of raw experience (James, 1890; Goldstein, Waterfall, Lotem, Halpern, Schwade, Onnis, & Edelman, 2010). However, as Zellig Harris (1991) points out, the buzzing confusion of the raw data are actually highly intrinsically structured, even more so when one considers regularities in the extrinsic social context in which such data are presented (Goldstein et al., 2010). This structure reveals itself in the biased distributions of repeated content – linguistic, interactional, and contextual – which infants can exploit to parse, classify, and predict language use (Saffran, Newport, & Aslin, 1996). Over time, the information carried by these distributions allows children to generalize more abstract patterns, such as those that constrain the co-occurrence of words, that is, syntactic constructions (Goldberg, 2006; Tomasello, 2003). These processes have also been observed for adults learning miniature artificial languages (Wonnacott, Newport, & Tanenhaus, 2008). Crucially, Wonnacott and colleagues show that even after relatively small amounts of exposure, adults show sensitivity to distributional information in production, comprehension, and even in abstract linguistic competence (in the form of grammaticality judgments). Thus, it is reasonable to expect that the distributional properties of words shape acquisition, as well as language processing, all the way into adulthood. Indeed, the natural distributions of words, as measured on the basis of very large corpora, have been shown to impact child and adult language use across multiple types of context: lexical contexts (words in sequence), morphological contexts (stems and affixes), and syntactic contexts (words and the structural frames they inhabit).

1. Lexical context

The frequency of words has long been known to play a strong role in lexical processing (Oldfield & Wingfield, 1965). However, recent work suggests that lexical frequency actually summarizes a number of different factors (e.g., Baayen, 2011). One such factor is the diversity of lexical contexts in which words occur. For example, more frequent words are more likely to surface near a greater variety of other words within small or immediate contextual windows. McDonald and Shillcock (2001a,b) introduce a measure which they call *contextual distinctiveness* to measure the diversity of these lexical contexts. They use an information-theoretic measure called relative entropy to capture the amount of information carried by the lexical context of a word relative to the prior expected frequencies of words within the language generally. They find that this measure outperforms word frequency as a predictor of adult performance in visual lexical decision. However, they did not find a correlation between contextual distinctiveness and age of acquisition. Instead, they found that it correlated strongly with lexical ambiguity, suggesting that the measure taps into semantic representations. Later research affirms the usefulness of contextual windows for modeling semantics (e.g., Bullinaria & Levy, 2012).

2. Morphological context

Studies of the morphological distributions of words have uncovered a number of effects in adult processing. In a pioneering study, Kostić, Marković, and Baucal (2003) propose a measure capable of accounting simultaneously for (a) the base frequency of inflectional variants of words and (b) the number of syntactic functions possible for that inflectional category, relative to the overall syntactic variability of cases within the paradigm.[1] For example, the Serbian paradigm for feminine nouns contains six morphological variants. For each stem, one takes the probability of its occurring in a particular case-inflected form, then divides that by the number of syntactic functions served by the case. This ratio is then divided by the sum of the complete set of ratios for each of the possible inflectional variants.

[1] Cases typically differ in the number of syntactic functions they may serve. For example, the Serbian accusative case expresses object status, as in *Uzeo je svoju knjig-u* 'He took his book-ACC', as well as temporal adverbials, as in *Došao je u sred-u* 'He came (on)

The negative (base-2) log is applied to transform the value to (positive) bits. The resulting value increases when a case-inflected variant is both syntactically heavy and takes up less of the overall probability of its stem. Crucially, case-inflected variants with higher values for a given stem are processed more slowly, with this measure accounting for up to 88% of the variability in visual lexical decision latencies for adults. These results show that inflectional distributions carry information about both the likelihood of a word occurring, as well as the uncertainty of the syntactic function that the word will serve.

Building on this research, Moscoso del Prado Martín, Kostić, and Baayen (2004) introduce a measure based on Shannon's definition of entropy for discrete distributions (Shannon, 1948), known as the *inflectional entropy*. This measure captures the productivity of a stem across its inflectional variants, given the overall probability of encountering a word from the same inflectional class, where the probabilities reflect maximum-likelihood estimates based on large corpora of naturally occurring language use. Inflectional entropy increases as stems are more evenly distributed across the available case-inflectional variants. Moscoso del Prado Martín and colleagues find that inflectional entropy correlates negatively with response times in a visual lexical decision task for adults. Therefore, words are processed best when they have been experienced in a more diverse array of inflectional environments. Similar results have been reported for English, which uses analytic means for expressing the same inflectional meanings (Lester & Moscoso del Prado Martín, 2015). Baayen, Feldman, and Schreuder (2006) explore the correlations between inflectional entropy and a number of other variables. Crucially, they find that inflectional entropy

Wednesday-ACC' (Kostić et al., 2003).

significantly predicts subjective age of acquisition ratings, such that words with higher entropies are learned earlier. Stoll et al. (2012) find that in Chintang, a polysynthetic Tibeto-Burman language with complex verbal morphology, children begin to produce greater shares of verbs as they approach adult-like verbal inflectional entropies for in their own speech.

Milin, Filipović-Đurđević, & Moscoso del Prado Martín (2009) elaborate these findings by introducing the notion of paradigms. They use the relative entropy (basically the same measure applied by McDonald and Shillcock, 2001a) to measure the typicality of the distribution of Serbian nouns across their inflectional exponents relative to the overall pattern for words of the same gender (masculine or feminine). They find that nouns with more typical distributions are processed faster.

3. Syntactic contexts

Earlier work from theoretical linguistics suggests a strong link between the processes that operate within words and those that operate between words (Marantz & Halle, 1993; Marantz, 1997). These similarities have also been noted on the typological scale: where some languages express a syntactic relationship morphologically, others may encode the same relationship analytically. Less research has examined the role of syntactic distributions in isolated lexical processing. Baayen, Milin, Filipović-Đurđević, Hendrix, & Marelli (2011) modify the approach of McDonald and Shillcock (2001a,b) by restricting the contextual variation to a specific syntactic construction, namely, the prepositional phrase. They measure the typicality of the distribution of nouns within prepositional phrases (in this case, just prepositional trigrams of the form PREP + DET + NOUN, as in *on the table*) against the baseline frequency of prepositions. They found similar results to those reported by

McDonald and Shillcock: words with more typical distributions within the syntactic frame were recognized faster in visual lexical decision. Effects of this measure have since been observed in the electrophysiological signature (Hendrix, Bolger, & Baayen, 2016). Note, however, that this approach still relies on small co-occurrence windows and purely lexical variability (i.e., variation occurs within a syntactic construction, not across syntactic constructions).

Linzen, Marantz & Pylkkänen (2013) take first steps towards defining a truly syntactic distributional space for verbs. Rather than relying on lexical variation within a single syntactic construction, they measure the distribution of verbs across the set of argument structures that they project. Unlike Baayen et al. (2011), they do not find a behavioral effect in visual lexical decision. However, they do find an effect in the electrophysiological signature. This could be due to many factors, including a change in the word class being investigated. Lester & Moscoso del Prado Martín (2015) create a hybrid syntactic space for English nouns based on a large corpus of English text. They model the space after the functions encoded by the rich case systems of the Finno-Ugric languages (e.g., Finnish, Estonian, and Hungarian). This space was defined using a combination of prepositional phrases (e.g., *of* for the genitive relation) and positions within abstract phrase-structure trees (e.g., N heads of NPs that are simultaneously leftward sister to VPs and daughter to S nodes for the nominative relation). Unlike Baayen et al. (2011), purely syntactic frames were included. Unlike Linzen et al. (2013), measurements were taken for nouns, and specific syntactic functions were prioritized over full argument structure frames. Unlike either of these studies, the (bias-corrected) entropy of the frequency distributions was taken directly, rather than taking the relative entropy. Similar to Baayen et al. (2011), but unlike Linzen et al. (2013), they find a strong negative correlation between diversity and response times in visual lexical decision. Lester & Moscoso del Prado Martín (2016) refined this strategy by creating three fine-grained, purely syntactic entropies of nouns based on dependency graphs. Dependency graphs have individual words as nodes. Arcs connect pairs of nodes, and each arc represents a typed syntactic relationship. For each pair, one word "depends" on the other (e.g., in the phrase *a* c*ake, a* depends on *cake* via the *determiner* relation). Lester and Moscoso del Prado Martín define the syntactic space as the frequency distribution of nouns across the different dependency relation types in a large corpus of American English. They compute separate entropies for nouns as heads and nouns as modifiers. They correlate these measures with production latencies in a bare-noun picture naming task. They find a positive correlation for as-head diversity, but a negative correlation for as-modifier diversity, suggesting that different aspects of syntactic distributions can impact word processing in different ways, at least in word production. However, the dependency formalism used in that study presents a potential issue: while the syntactic relationships themselves are abstract, they are instantiated by specific words. In many cases, the nature of the syntactic relationship may be fully reducible to the identity of the words. This predictivity between word and syntactic relationship suggests a potential confound between those measures and the contextual distinctiveness measures of McDonald and Shillcock (2001a,b).

Each of these studies comes with certain shortcomings regarding how they operationalized syntax. Baayen and colleagues measure lexical variability while controlling for the syntactic form of the utterances. Linzen and colleagues measure syntactic variability, but do not control for differences across hierarchical levels (e.g., as-head vs. as-modifier contrasts for verbs), or for the contributions of variability across the component phrasal units that make up the argument-structure configurations. Lester and Moscoso del Prado Martín account for syntactic diversity, as well as hierarchical asymmetries, but do not control for lexical variability. None of those studies accounts for another crucial feature of syntax, namely the relative ordering of words bound by syntactic relations. Further, only Linzen and his colleagues compare syntactic diversity and typicality as competing predictors of behavior. A major goal of this dissertation is to build on the approach of Lester and Moscoso del Prado Martín (2016) to solve these issues. Doing so will help to clarify questions regarding how syntax guides the distributional learning of words, shaping patterns of use and ultimately the language processing architecture itself. A second important goal is to generalize the findings reported there to cover multiple aspects of word production and comprehension, as well as language acquisition during early childhood.

Each of the core chapters of this dissertation – Chapters II through IV – is designed to be a standalone paper. As such, the reader should expect to encounter some of the same information across these chapters, particularly regarding the computation of the measures. However, readers interested in specific topics will find everything they need within each chapter. Chapter II introduces several new information-theoretic measures of the prior syntactic distributions of nouns. Analyses of the effects of these measures on two previously published datasets of visual lexical decision, simple and primed, are reported. Chapter III explores similar effects in word production. Results of two picture-naming experiments are reported. The first experiment requires participants to name pictures with isolated nouns (e.g., “banjo!”), revealing effects of prior syntactic distributions in a non-syntactic task. The second explores whether these effects remain when the participants are required to produce the names within a syntactic frame (e.g., “the banjo!”). Chapter IV reports a corpus study of the first appearance of nouns in a dense longitudinal sample of two- to three-year-old child speech. A regression technique based on the logic of survival analysis is used to test whether prior syntactic distributions affect the emergence of nouns in the earliest stages of syntactic development. Chapter V synthesizes the results of the studies from comprehension, production, and acquisition to arrive at a general picture of the effects of prior syntactic distributions on how nouns are learned and processed. Limitations of the studies and directions for future research are also presented. Finally, the Appendix describes a large database – the Syntactic Diversity of English Nouns, or SynDI-EN – which contains measures of the syntactic diversity and prototypicality of thousands of English nouns. The methods for computing these measures are described in detail, and examples are provided for high- and low diversity/prototypicality nouns.

II. Effects of Prior Syntactic Distributions on Comprehension

A. Simple lexical decision

Much of what we know about lexical processing comes from studies that analyze behavioral or neurological responses to isolated words. The goal of such research is to probe the inner workings of the lexicon by limiting interference from syntax and other external factors. Accordingly, the focus has been on lexically endogenous variables, including those related to orthography, phonology, semantics, surface or lemma frequency, and so on (Balota et al., 2007). It has often been claimed that contextual variables do not impact lexical processing. For example, Balota, Paul, and Spieler (1999), summarizing the state-of-the-art in lexical processing research at the time, state that "discourse-based syntactic and semantic information do not contribute to isolated word recognition" (p. 15). However, converging evidence from comprehension and production suggests that isolated word processing is also sensitive to the semantic and syntactic contextual distributions of words (Adelman, Brown, & Quesada, 2009; Baayen, Milin, Filipović-Đurđević, Hendrix, & Marelli, 2011; Hendrix, Bolger, & Baayen, 2016; Kostić, Marković, & Baucal, 2003; Landauer & Dumais, 1997; Linzen, Marantz, & Pylkkänen, 2013; Milin, Filipović-Đurđević, & Moscoso del Prado Martín, 2009; Moscoso del Prado Martín, Kostić, & Baayen. 2004). The logic often invoked to explain these effects is that our natural experience of language involves simultaneous activation of words and syntactic structures (e.g., when we read a sentence, listen to a friend, etc.). If the system develops in response to our natural experience, then we should not expect words to dissociate from syntax simply because we have contrived to present the former absent the latter (Linzen et al., 2013).

Several issues remain open, of which I consider three. First, behavioral results have been mixed. Baayen and colleagues observe effects of syntactic distributions in both behavior (Baayen et al., 2011) and neurophysiological signals (Hendrix, et al., 2016). Linzen et al. (2013) replicated the neurological effect but did not find an effect on behavior. This discrepancy may come from several sources. First, the two studies considered different types of distributions. In the study of Baayen et al. (2011), syntactic distributions were measured within a single structure: the prepositional phrase. They compute the frequency with which nouns and prepositions co-occur. The label "syntactic" in this context thus refers to variability between word forms that share a syntactic bond. That is, the frequencies reflect word/word co-occurrence. Such lexical co-distributions are known to reflect semantics (Bullinaria & Levy, 2012), thus bringing the syntactic nature of the effect into question. Linzen and colleagues, on the other hand, looked at variability of words across syntactic structures. They compute the frequency with which verbs occur in each argument-structure construction (e.g., the ditransitive construction <NP_{AGT} VERB NP_{REC} NP_{PAT}>, as in *The boy sent his grandmother a letter*). These frequencies reflect word/structure co-occurrence. Perhaps a similar measure for nouns would likewise show no correlation with response times (RTs).

Second, the role of probability is not yet well understood. Linguistic models differ in whether the relationships between words and syntactic structures are categorical (licit vs. illicit; Chomsky, 1995) or probabilistic (usage-based; Diessel, 2015). Distributional measures will necessarily capture aspects of both: forms that license more structures naturally have more diverse frequency distributions than those that license fewer structures.

However, no study to my knowledge has directly pitted these two explanations against each other. Simply because probabilistic measures have been successful does not mean that analogous categorical predictors could not produce the same effect.

Third, in reality, syntactic contexts are far richer than simple mappings from word to structure. For example, they involve both hierarchy – functional asymmetries between related units, often termed *headedness* – and word order. For example, some theories argue that words only carry syntactic information about the structures that they head (e.g., Chomsky, 1995). Heads are the functional cores of syntactic structures, as exemplified by the noun *man* in the noun phrase *the tall man*. In this case, word/structure distributions are expected to impact processing only when sampled across structures that the word heads. The measures of Linzen et al. (2013) focus on head structures, while those of Baayen et al. (2011) focus on non-head structures (nouns are not heads of prepositional phrases). Word order, on the other hand, may relate directly to processing speed. For example, structures that require words to be produced earlier may have different aggregate effects on word processing than those that require words to be produced later. Specifically, the former structures may produce facilitatory effects, the latter inhibitory effects, in line with the positions they enforce upon the words.

In the present study, I address each of these points. First, I develop several probabilistic measures of the fully delexicalized syntactic distributions of nouns. These measures serve as analogues to the measures used by Linzen and colleagues. They differ from prior measures by accounting for both hierarchy and word order. Second, I develop an alternative set of measures based on the assumption that syntactic information in the lexicon is purely categorical. I carefully decorrelate the probabilistic and categorical predictors and allow them to compete as predictors of RTs in lexical decision. If syntactic information does not affect lexical decision, the improved noun measures offered here should not correlate with RTs. On the other hand, if syntactic information does impact isolated word recognition, we should derive a better estimate of the shape and magnitude of this effect from these measures compared to those proposed in earlier studies. Finally, if probabilistic syntactic information is relevant, then we should find that the probabilistic measures explain a unique portion of the variance in RTs, over-and-above what can be attributed to categorical distributions alone.

1. Effects of syntactic context on lexical decision

Different aspects of syntactic context have been found to affect how quickly we recognize words in visual lexical decision. Early work focused on inflectional morphology (the syntactic component of word building). Moscoso del Prado Martín, et al. (2004) found that in Serbian, a language with seven inflectional cases, nouns are recognized faster to the extent that they spread their probability evenly across the case inflections. Serbian case inflections reflect syntactic functions (e.g., status as subject or direct object). Therefore, this finding suggests that when nouns are used in a diverse array of syntactic constructions, they are easier to process. Henceforth, I refer to this type of effect as a *diversity effect.* This interpretation was later challenged by evidence that similar effects could be simulated using distributional semantics alone (Moscoso del Prado Martín, 2007). But later work pointed out that these distributions are actually hierarchically organized. Words are nested within inflectional classes. These classes define the formal properties of the morphosyntactic variants of word roots. Serbian nouns may belong to many such classes depending on how they inflect. Milin et al. (2009) measured the case distributions of nouns against the average, or *prototype*, distribution specific to their inflectional class. Words that matched the average distribution of nouns from their class were recognized faster. I refer to this type of effect as a *prototypicality effect.* Excusing the homoncular analogy, these lexical prototypes may be thought of as the “expectations” of the processor. Words that fit the system better are processed more efficiently. However, these prototype measures still suffer from the same issues as the general distributional measures proposed by Moscoso del Prado Martín et al. (2004). Based on Moscoso del Prado Martín (2007), these effects could be attributed to semantic prototypes, though this possibility has not to my knowledge been explored.

Baayen et al. (2011) scale the investigation up to analytical syntactic relations. They treated prepositional trigrams in English (e.g., *in the bucket*) as proxies for an English analytical case system. They computed prototype measures for each noun across the set of prepositions in these trigrams given the average distribution of prepositions. I refer to this measure as the *in-structure approach* to reflect the fact that the distribution is calculated over lexical co-variability within a single syntactic structure. They found that nouns that matched the prototype were recognized faster. In other words, nouns are processed most efficiently when they serve best the functions we need most. Importantly, the effect is the same for both morphological and analytical manifestations of syntax.

Linzen et al. (2013) attempted to replicate these effects for verbs. They looked at the distributions of verbs across 28 different argument-structure configurations (i.e., the highest order syntactic units that must accompany the verb to create an acceptable utterance). The distributions so defined differ crucially from those proposed by Baayen et al. (2011) in that they measure variability across instead of within abstract structures. I refer to this as the *across-structure approach*. They computed diversity and prototypicality measures based on these distributions. They found neural correlates in areas consistent with syntactic processing (Broca's area), suggesting that their measures had successfully tapped into syntax. I refer to this as the *across-structure diversity* effect. However, they did not replicate the behavioral effect for either measure in lexical decision. They tentatively conclude that syntactic information does not impact lexical decision. Importantly, however, they found that the diversity and prototypicality effects had different localizations and time-courses. These differences suggest that diversity and prototypicality constitute distinct aspects of word processing.

The in-structure and across-structure measures (as applied in these studies) differ in at least two important ways, both of which could help to explain the discrepancy in the behavioral findings. First, in-structure measures capture word/word co-occurrence distributions while across-structure measures capture word/structure co-occurrence. Any approach that measures the distribution of words relative to other words will be sure to capture a great deal of semantic information (Bullinaria & Levy, 2012). Therefore, the positive effect of in-structure diversity on behavior found by Baayen et al. (2011) could be due to semantics (see Moscoso del Prado Martín, 2007). Second, the across-structure approach as applied by Linzen and colleagues produces head distributions, while the in-structure approach as applied by Baayen and colleagues produces non-head distributions. Therefore, the discrepancy might be due to a contrast in the effects of head-based and non-head-based distributions. We should thus prefer a measure that captures word/structure instead of word/word relationships, but which also accounts for headedness.

2. How is context represented?

I have so far focused on how context affects responses to stimuli. But we still need an account of what types of representations or processes can account for these responses. Several possibilities have been considered in the linguistic and psycholinguistic literature. At one extreme, some linguistic theories argue that words (specifically, open-class roots) bear no syntactic information whatsoever (Borer, 2005; Marantz, 1997). To explain the behavioral effects, these theories could argue that reading usually involves syntactic processing. The artificiality of the task does not overcome the expectations of the system, and the independent syntactic system kicks on when exposed to the word. This syntactic activity could feed into the lexical decision. Such an account would be difficult if not impossible to distinguish from theories that include syntactic information in the lexicon.

Other accounts annotate words for syntactic features. In these theories, syntactic features are matched against labeled positions in syntactic trees to ensure that each word is slotted into its appropriate position. This is the general logic behind terminal productions in context-free grammars (CFGs). For example, in the production N → *chicken*, the nonterminal N category is equivalent to a syntactic annotation for *chicken* that constrains its distribution in the broader syntactic system. These annotations can be much more elaborate than simple part-of-speech labels. For example, in Lexical Functional Grammar (LFG; Bresnan, 2001; Neidle, 1994), lexical entries are marked for the complements (*arguments*) they take, both at the functional (e.g., *agent*) and structural level (e.g., *subject*). This information is stored as categorical feature labels that allow the word to trigger syntactic building processes. These theories can therefore account for the relationship between verbs and the diversity of structures studied in Linzen et al. (2013) without having to invoke spontaneous task-irrelevant syntactic activity. However, these theories are explicitly non-probabilistic. In order for these theories to be correct, the number of syntactic structures should explain the RTs better than the frequency distribution of a targets across those structures. Earlier probabilistic findings would be recast as noisy approximations of the number of syntactic types per word (type count and entropy are positively correlated).

Another class of theories emphasizes the functional and theoretical similarities between words and syntax. Many such theories even go so far as to define syntax as simply the most abstract end of the lexicon (e.g., Langacker, 1987; Goldberg, 1995; Diessel, 2015). Words relate directly to syntactic structures, similar to the representations in LFG. However, they can relate to *any* type of syntactic structure, irrespective of whether they are functional heads of that structure. This more inclusive position predicts that syntactic distributions beyond the subcategorization frames studied by Linzen and colleagues should also impact processing. Possible support for this comes from Baayen et al. (2011), who found distributional effects for a syntactic structure in which the noun is not head.

Another important feature of these theories is that they treat word–syntax relationships as fundamentally probabilistic (Diessel, 2015). Relationships between nodes in the network are tuned by experience: the more often and more distinctively that two linguistic units are experienced in close syntagmatic or syntactic conjunction, the stronger the connection between them (and the weaker the connections between these and other structures). This model allows for a straightforward interpretation of the distributional effects observed in lexical decision. These effects could arise through a pattern of feedback between lexical and syntactic nodes, where the local feedback potentials are proportional to frequency.

Psycholinguists have proposed their own set of models of the lexicon and lexical access specific to comprehension (Baayen et al., 2011; Coltheart et al., 2001; Davis, 2010; Grainger & Jacobs, 1996; Norris, 2006; Plaut, 1997; Morton, 1978; Seidenberg & McClelland, 1989). None of these models contains a syntactic component, but neither do any specifically preclude syntax. However, as pointed out by Norris (2013), some models are more flexible than others. For example, the interactive activation models of Seidenberg and colleagues (e.g., Harm & Seidenberg, 2004; Seidenberg & McClelland, 1989) require a new component and interfaces between that component and the others. The fundamental organization of the model would change, but the functional properties would remain the same. Specifically, a tier of syntactic nodes would be added, with connections at least to orthography/phonology, and likely to semantics as well. Prototype effects could be modeled by setting resting activation at the syntactic tier according to the prototypical distribution. Other models, such as the Bayesian Reader of Norris (2006), simply need to “plug” syntax into the existing machinery. For example, syntactic information could be fed into the prior probability of the Bayesian equation. Similarly, task-specific models such as the Drift-Diffusion Model of two-way choices (Ratcliff et al., 2004) could also easily accommodate new information streams when making predictions about behavior in lexical decision.

Other models might not need any change at all. Baayen et al. (2011) introduce a two-tier network of input orthographic nodes and output meaning nodes. They couple this network with an expectation-based, error-driven learning algorithm (Rescorla & Wagner, 1972). The network was able to model the syntactic paradigm effects from Milin et al. (2009). This means that a morphological paradigm exerted its effect without being represented in the model! They explain the success of the model in terms of *discriminative learning*. Syntax provides stable points of variability within the input. For example, English prepositional phrases define a position relatively close to nouns in which prepositions may vary. Over time, this variability helps to carve away the incidental aspects of the context to solidify the connection between the noun's form and its meaning. What remains is the most cross-contextually stable meaning that coincides with the presence of the noun. With more diverse distributions come stronger and more targeted inferences from noun form to meaning. Similarly, prototypical words, whose distributions match the expectations of the system the best, stand to benefit the most from the contextual variability that drives learning. This leads us to the third hypothesis:

The argument from discriminative learning runs into a problem with the findings of Linzen et al. (2013), who found no effect of diversity or prototype measures on lexical decision RTs. Why should the discriminative logic play out for nouns but not verbs? A possible answer presents itself if we consider the different ways that the two studies defined their syntactic distributions. Baayen and colleagues looked at *lexical* variation within a *single* syntactic construction. Their measure therefore amounts to a syntactically constrained version of the lexical co-occurrence measures discussed in Bullinaria and Levy (2012), which are typically interpreted as capturing semantics, not syntax. Furthermore, by looking at lexical variation, Baayen and colleagues bias the question in favor of the abilities of their two-tier model. By contrast, Linzen and colleagues looked at syntactic variation within a single lexical item. The distributions they considered were based on abstract syntactic templates. Therefore, they specifically ignore the lexical contribution of the syntactic context, where Baayen and colleagues rely on it completely. Without the overt lexical cues for the different syntactic constructions, discriminative learning may not apply. The question is swhether other measures of syntactic diversity and prototypicality will likewise produce null results for nouns once lexical cues have been filtered out.

The above literature review suggests three general hypotheses about possible syntactic effects on lexical recognition. These three hypotheses relate to syntactic diversity, measured categorically and probabilistically, and prototypicality. They are outlined below:

- *categorical hypothesis:* words that are attested in more syntactic constructions are recognized faster.
- *probabilistic hypothesis*: nouns that are distributed more uniformly across the syntactic structures in which they occur will be recognized faster.
- *prototypicality hypothesis*: nouns with syntactic distributions that resemble that of the prototypical noun will be recognized faster.

There are two points to note about these hypotheses. First, the two diversity measures, categorical and probabilistic, are treated separately. This is because the number of available syntactic structures is logically independent of the frequencies with which a word occurs in those structures, and so may independently affect RTs. Second, prototypicality is treated alongside the diversity measures. This is because Linzen et al. (2013) found neurophysiological evidence that diversity and prototypicality tap into separate mental processes. Therefore, predictions about effects from these two sources are not logically attached to the same null hypothesis.

Across the three hypotheses, there are eight possible outcomes. Of these, only four are seriously associated with (psycho)linguistic theory. These patterns along with the compatible theories are given in Table 1. Pluses indicate support for the hypothesis; minuses indicate no support.

Table 1: Possible outcomes across hypotheses and compatible theories.

Diversity			
Categorical	**Probabilistic**	**Prototypicality**	**Supported theory**
+	+	+	Baayen et al., 2011 Diessel, 2015
-	+	+	Goldberg, 2006
+	-	+	*not predicted*
+	+	-	*not predicted*
+	-	-	Bresnan, 2001 Chomsky, 1995
-	+	-	*not predicted*
-	-	+	*not predicted*
-	-	-	Linzen et al., 2013

The discriminative learning and usage-based models are compatible with positive results for probabilistic diversity and prototypicality. Two outcomes meet this requirement, shown in the first two rows of Table 1. The difference between these outcomes speaks to a secondary question regarding the independence of categorical and probabilistic diversity effects. Do they tap into distinct processes, or are the categorical measures merely worse approximations of the same phenomenon underlying both measures? If the latter is true, the effect of probabilistic measures should swallow that of the categorical measures. If not, we should see independent effects of each. Importantly, we shouls not see an effect of categorical diversity and prototypicality, but not probabilistic diversity. This outcome would require that probabilistic information is represented and exploited by the prototypicality system but ignored by the diversity system. No theory reviewed here could explain this outcome in a principled way.

In the next section, I introduce categorical, probabilistic, and prototypical measures of the syntactic distributions of nouns. I adopt a lower-level approach than Linzen et al. (2013) based on Dependency Grammar (Hudson, 2007; Mel'čuk, 1988; Nivre, 2005; Tesnière, 1959) that accounts for both word order and headedness in a straightforward way. With the help of these measures, we can evaluate whether (truly) syntactic distributions affect isolated noun processing and, if so, how.

3. Measuring syntactic distributions

Defining the syntactic space for any word is tricky. For one, syntactic constructions may be defined at multiple levels of abstraction, ranging from the binding of individual pairs of words to the ordering of entire phrasal units. Moreover, these constructions may or may not differ as a function of specific lexical content occupying one or more positions in the abstract syntactic frame (e.g., the *what's* X *doing* Y frame; Kay & Fillmore, 1999). To make matters worse, any actual syntactic token of moderate complexity is thought to simultaneously instantiate all hierarchically embedded component constructions (Langacker, 1987; Goldberg, 1995). Therefore, even if you could identify all constructions (you cannot, for several reasons beyond the scope of this study), you would have to decide at which level(s) to count. This situation creates serious practical problems. The more inclusive we become in terms of what counts as a distinctive syntactic entity, the less likely we are to observe sufficient frequencies of those constructional types to form a reliable picture of a given lexical item.

I attempt to side-step these issues by abandoning completeness in favor of compactness. I define the syntactic space based on the low-level, relatively economical syntactic categories found in the dependency-based grammatical formalisms (e.g., Hudson, 2007; Mel'čuk, 1988; Nivre, 2005; Tesnière, 1959). In the more typical phrase-structure model of syntax, syntactic structures are viewed in terms of constituency, or functionally bound groupings of words (e.g., the $\bar{X}$ formalism; Jackendoff, 1977). Such formalisms require both word nodes and abstract/phrasal nodes (e.g., NP standing for a noun and all of its dependents). Dependency approaches differ by focusing only on the immediate syntactic relations between pairs of words. Each syntactic relation has a tri-fold structure, which I refer to as a *bundle*. Each bundle consists of a *head*, a *modifier*, and a typed functional relation, or *dependency*. The head is roughly the semantic and syntactic core of the dependency. For example, the head usually (but not always) determines the behavior of the bundle, and is usually modified in some respect by the modifier. More precise definitions are difficult to pin down, and vary across dependency-based theories. The modifier is defined negatively as the word which is *not* the head. As mentioned previously, it typically modifies some aspect of the meaning of the head, though it may also serve other syntactic functions. For example, in some systems, conjoined nouns are directly related by the *conj* relation. In these relations, the non-initial coordinand(s) are modifiers of the initial coordinand. Thus, in the phrase *the dog and the cat*, *cat* would serve as modifier to *dog*, even though the relationship between the two is much different from that of *yellow* to *bumblebee* in *the yellow bumblebee*. Finally, the dependency label specifies the type of syntactic relationship between the words. As an example, consider the sentence *The rabbit hopped.* This sentence consists of three dependencies. First, the word *hopped* is bound to the abstract sentential *ROOT* via the *root* dependency. This is the dependency version of the starting symbol (for English, the S node) in context-free phrase-structure grammars. Next, *hopped* serves as head to modifier *rabbit* via the *nsubj* dependency. This dependency binds subject head nouns to verbs. Finally, *rabbit* calls *the* as a modifier via the *det* dependency, which binds determiners to nouns. The diagram of these relationships, known as a *dependency graph*, is presented as Figure 1. In the diagram, arrows point from heads to modifiers.

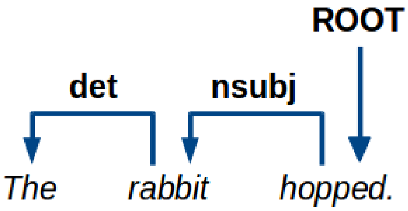


Figure 1: A dependency graph of the sentence The rabbit hopped

Dependency formalisms differ with respect to the types of syntactic relationships they recognize. These differences can arise for many reasons, for example, variability in the salient features of the languages to be modeled, the goals of the designer (e.g., fine-grained analysis of a single language vs. comparison across many languages), and so on. Despite this variability, the most commonly used dependency sets show a great deal of overlap in terms of general inventory and how the inventory is mapped into particular linguistic structures. For convenience, I use the CLEAR dependency labels (Choi & Palmer, 2012), which have been operationalized in the spaCy dependency parser for Python (Honnibal & Johnson, 2015). The CLEAR labels are somewhat more narrowly defined than other popular dependency sets (e.g., the Universal Dependency labels; Schuster & Manning, 2016), and so offer a slightly more fine-grained perspective.

These dependencies provide a compact, low-level representation of the English syntactic system. Importantly, they capture similarities between instances of the fully realized argument structure that were not captured by previous measures (Linzen et al., 2013). For example, intransitive and transitive uses of a verb both involve the *nsubj* subject dependency, and so with *dobj* object relations in transitive and ditransitive uses, etc. Given that language processing has been shown to be highly incremental (e.g., Ferreira, 1996; Novick, Kim & Trueswell, 2003), the lower-level perspective offered by the dependency parse may present a more realistic picture of the primary units of interaction in everyday language processing. And in many linguistic theories, lexical items interface with broader structures mostly through intermediary embeddings, from the nested structures of Cognitive Grammar (Langacker, 1987) to the maximal projection rules of generative grammar (Chomsky, 1970). Moreover, dependencies provide simple operationalizations of headedness (the head/modifier contrast) and word order (the direction in which the dependency faces). They therefore allow us to go well beyond prior measures in assessing how syntactic context impacts word processing.

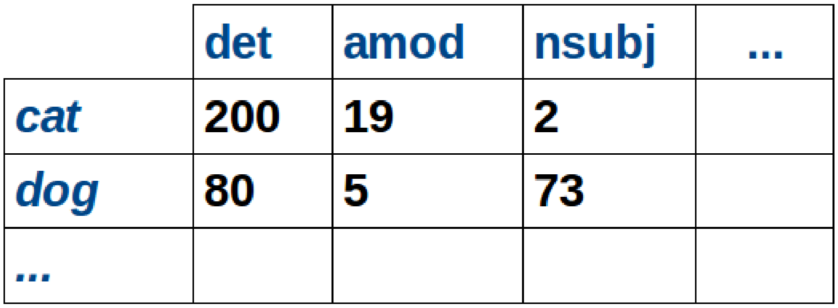
i. Syntactic diversity: Probabilistic vs. categorical approaches. I define the syntactic diversity of a word as its frequency distribution across the binary syntactic dependencies in which they occur. The standard tool for summarizing frequency distributions that has been employed by all previous studies in this vein comes from information theory, namely, the entropy (Shannon, 1948). Entropy is defined as the average negative log probability of any given outcome of a random variable, for example, the occurrence of a particular English word. The formal expression of the Shannon entropy is given in Equation 1**.**

$$H = -\sum p(w) \log p(w) \tag{1}$$

Higher entropies indicate more diverse distributions. For discrete entropies, the upper limit is defined as the negative log of the total number of possible outcomes (i.e., the uniform distribution, where all outcomes are equiprobable). This would be the case for words that occur equally often in each of the dependencies available to them. Such words could be said to carry the maximal amount of syntactic information. The lower bound is 0, which arises only under conditions of perfect certainty (i.e., when only one outcome is ever attested). This would be the case for words that only occur in a single dependency. Such words carry no information about the syntactic system.

Measuring syntactic diversity in the manner proposed above requires that we consider the joint occurrence of nouns and syntactic dependencies. Accordingly, *p*(w) in Eq. 1 will be replaced by the joint probability of the target word *t* and any given syntactic dependency *d*, expressed as $p(d\,, t)$ where $d \in$ D, and D is the set of all dependencies. In practice, D must be defined relative to a particular dependency annotation scheme. In this case, D is the set of unique CLEAR dependencies.

These probabilities are based on frequency counts, which can be organized into an *n x d* matrix, where *n* is the number of target nouns and *d* is the number of dependency types in D. Each row stands for a unique noun and each column for a unique syntactic dependency. Each cell contains the joint frequency of the noun and the dependency. Thus, for each instance of a target noun, I count how many times it occurs with each syntactic dependency, resulting in a frequency distribution of length *d*. Sample frequency distributions are provided in Figure 2.



	det	amod	nsubj	...
cat	200	19	2	
dog	80	5	73	
...				

Figure 2: Sample syntactic frequency distributions

These distributions still conflate the hierarchical status of the noun in the dependency (i.e., whether it is head or modifier) and the local word order (i.e., whether the dependency extends to the left or right). I therefore condition the frequency distributions based on these dimensions. For example, instead of tallying all instances of a noun in a given dependency, I can count only those instances for which the noun serves as head, or as rightward head, or modifier in any direction, and so on. Considering all possible combinations of hierarchy and word order yields nine distributions. These are schematized in Figure 3.

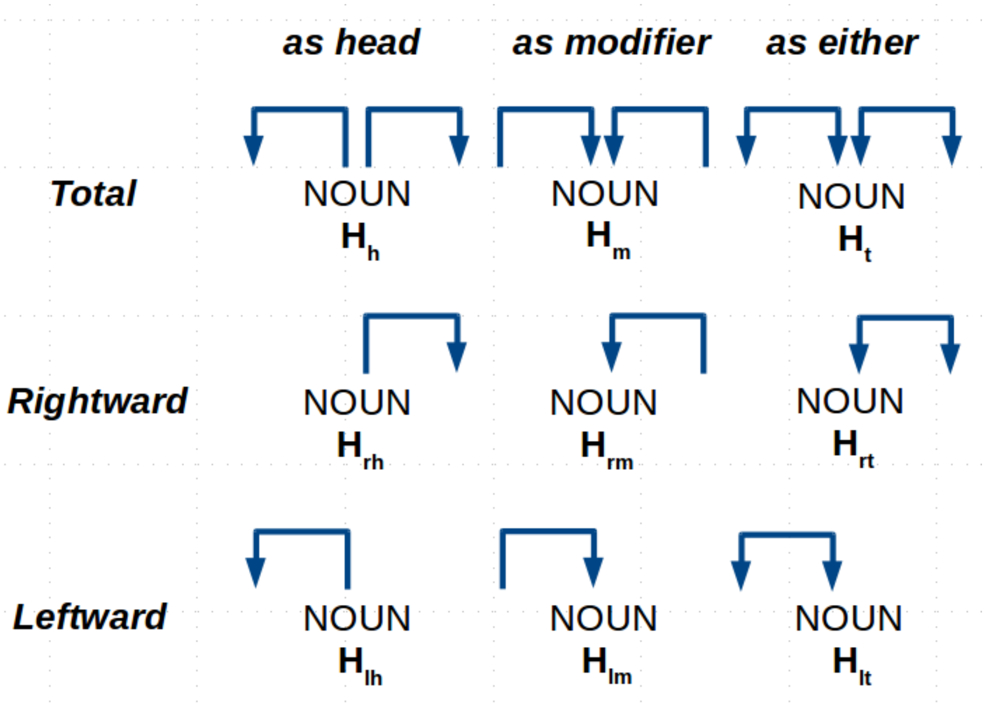


Figure 3: Schematized vector types.

In Figure 3, arrows point away from heads towards modifiers. Double-headed arrows indicate that hierarchy was not considered (i.e., head and modifier dependencies were not distinguished). Arrows extend to the left or right of the noun to indicate word order. By comparing the entropies of these distributions, I can explore which dimensions of syntax, at what granularity, are important for word processing.

These nine measures can be refined further. As I wish to measure the fully abstract syntactic information carried by nouns, I have ignored the words to which each noun was connected. Instead, I have counted only the abstract dependency types. These abstract dependency types have been assumed to capture syntactic relationships beyond what is available from the words alone (e.g., Lester & Moscoso del Prado Martín, 2016). In this respect, it resembles the entropies that Linzen et al. (2013) defined for verbs. However, dependencies never manifest apart from the words which instantiate them (setting aside the thorny issue of *null* or *zero* phenomena; see Fillmore, 1986), and some are restricted to only a few words. For example, the *det* relation, which binds determiners to nouns, allows very few words in modifier position (i.e., *the, a*, *this*, *these*, *that*, *those*, and so on). Furthermore, determiners are almost entirely restricted to the *det* relation in their own distributions. Thus, the information carried by the *det* category is largely bound up in the information carried by the determiners that appear in the context of a noun. No study to my knowledge has yet controlled for this relationship. However, as evidenced by the theoretical debate, we must tease apart the lexical and syntactic sources of contextual variability if we hope to draw sound inferences about the structure of the lexicon. In the extreme case, the information carried by words could be indistinguishable from that carried by syntactic structures. If so, we should expect naïve discriminative learning to apply just as it does for syntactically constrained lexical variation (Baayen et al., 2011). We would be then faced with two possible models: the model with word forms and meanings vs. the model with word forms, meanings, and syntactic categories. Given equal explanatory power, the former model should be preferred because it is simpler; it does not require the additional tier of syntactic generalizations. This conclusion – declared for morphology and hinted at for syntax by the proponents of discriminative learning (Baayen et al., 2011) – diverges markedly from prior linguistic and psycholinguistic models, and so must be considered carefully and rigorously.

To address this issue, we need some way to clean the measures of lexical information. Information theory provides a measure well-suited to this task: *conditional entropy*. The conditional entropy of distribution D given knowledge of distribution L is defined as $H(D \mid L) = H(L, D) – H(L)$, where $H(L, D)$ is the joint entropy of L and D. Let D be the frequency distribution across syntactic dependency types, and let L be the frequency distribution across lexical types of words bound to the targets. $H(L, D)$ is the entropy taken over the joint probabilities $p(l, d)$ for $l \in L$ and d for $d \in D$. This amounts to the sum of the entropies of the words and the dependencies independently minus the 'overlapping' mutual information between words and dependencies, or $H(L, D) = H(L) + H(D) - \mathrm{I}(L ; D)$. Based on this definition, conditional entropy can be rewritten as $H(D \mid L) = H(D) - \mathrm{I}(L ; D)$, or the information carried by the abstract dependencies minus the information shared between the dependencies and associated words. These relationships are schematized in Figure 4.

$H(D \mid L)$ has a lower bound of 0 when $D = L$, and an upper bound of $H(D)$ when D and L are completely independent. Conditional entropy applied in this way captures the information unique to abstract dependencies, completely divorced from the information carried by the surface forms.

Often, researchers estimate these measures using a maximum-likelihood estimators (e.g., Baayen et al., 2011; Kostić, Marković, & Baucal, 2003; Milin et al., 2009; Moscoso del Prado Martín et al., 2004), when based on samples, are known to be biased (Miller, 1955): they underestimate population-level (i.e., *true*) entropies. One way to combat this bias is to apply an entropy correction to account for the contribution of unobserved tokens.

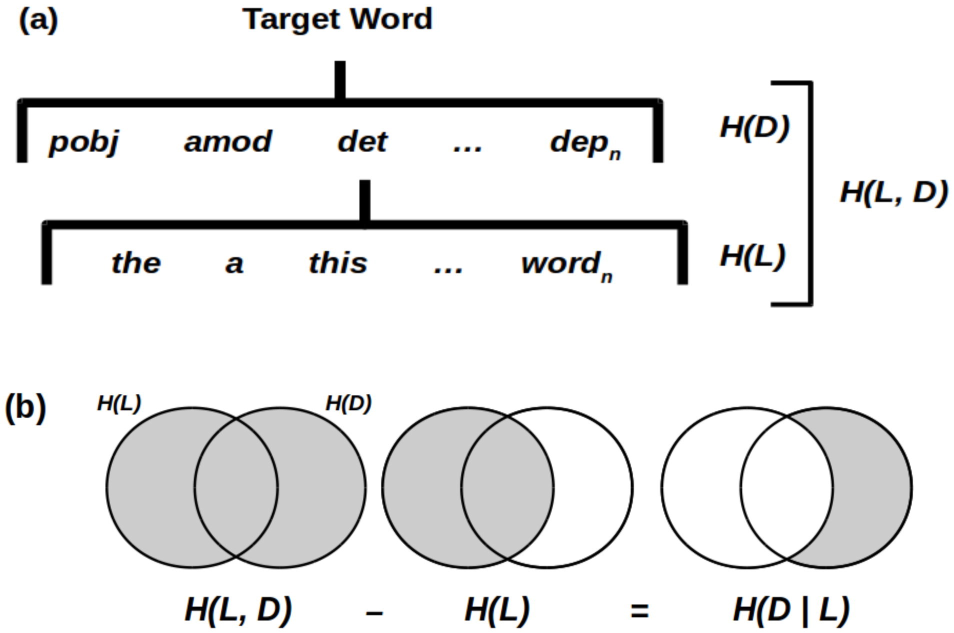


Figure 4: Schematization of the relationships involved in conditional entropy: (a) nesting of lexical items within dependency types for target words; (b) Venn diagrams depict the calculation of conditional entropy. Shaded areas reflect the portion of the entropies corresponding to the label beneath the Venn diagrams.

Moscoso del Prado Martín (2016) proposes using the method of Chao, Wang, & Jost (2013) for correcting the estimation bias.

To test the categorical hypothesis, we need a set of measures to account for the number of dependencies licensed for nouns irrespective of their probability distributions across those dependencies. This proves to be a much simpler enterprise. Assuming that activation of categorical representations is constant, processing should depend on the syntactic coverage of the noun, which can be expressed as the number of dependencies attested with non-zero frequency. Higher values reflect more robust interaction with the syntactic system. I assume that zero frequencies can be interpreted as “undefined” specifications and hence as functionally inert.

ii. Syntactic prototypicality. Prior studies have operationalized prototypicality as the *relative entropy*, sometimes known as the *Kullback-Leibler Divergence* (KLD; e.g., Milin et al., 2009; Baayen et al., 2011). Relative entropy measures the average number of bits required to recode a signal from one distribution as if it had come from an alternative distribution. Formally, it is expressed as Equation 2:

$$KLD(P||T) = \sum P\log\frac{P}{T} \tag{2}$$

This measure can be applied to the dependency distribution *f*(D), similar to the sample distribution in Figure 2. *T* is the frequency distribution of a given target word across syntactic dependencies $f\,(\mathrm{D}_{\mathrm{target}})$, and *P* is the 'prototype' distribution created by summing the distributions of all words $f(\mathrm{D}_{\mathrm{total}})$. The prototype is thus construed as the average distribution of words in the class.

So defined, relative entropy suffers from two problems. First, Equation 2 defines the maximum-likelihood estimate of KLD. It will therefore suffer from the same underestimation bias mentioned above for conditional entropy when applied to samples. For pointwise comparisons between distributions, this underestimation basis can be corrected by smoothing the frequency counts prior to taking the entropy. Several such methods are available. I select the James-Stein plug-in shrinkage estimator (Hausser & Strimmer, 2009).

This smooth is optimal for closed-class distributions (i.e., distributions for which the number of possible types is known). I assume that the set of dependencies encountered in a 15-million word corpus of English serves as a reasonable approximation of the total syntactic space (given the CLEAR dependency labels used by the spaCy parser).

The second problem is that relative entropy is asymmetric. The magnitude of prototypicality depends on whether the target is measured against the prototype or the prototype against the target. Ideally, one would not want to have to decide on a direction a priori (unless one's theory allows one to make such predictions). However, the relative entropy can be modified to produce symmetrical distance estimates using the Jensen-Shannon Divergence (JSD; Lin, 1991). The JSD between two distributions *P* and *T* is defined as the average relative entropy taken from each distribution to the midpoint between them *M*. JSD is defined formally in Equation 3:

$$JSD(T||P) = \frac{1}{2}KLD(T||M) + \frac{1}{2}KLD(P||M) \tag{3}$$

where

$$M = \frac{1}{2}(T + P) \tag{4}$$

With these refinements in mind, I define the syntactic prototypicality of nouns thus. Prototypicality is operationalized as the sum of all noun distributions in the sample. However, in this case, I take the sum over estimates corrected via the James-Stein plug-in method. Let *T* be the syntactic distribution of a given target noun $f_{\text{James-Stein}}(\text{D}_{\text{target}})$. The prototypicality of *T* relative to *P* can thus be given as JSD(*T*||*P*). As with the diversity measures, prototypicality can be measured in each of the nine syntactic distributions defined above.

4. Materials and Methods

I test the hypotheses by reanalyzing visual lexical decision RTs for the nouns of the English Lexicon Project (ELP; Balota et al., 2007). Only monomorphemic nouns are considered. Further, to avoid interference from out-of-class homographs (e.g., *the **hound** sniffed the stump* [noun] vs. *The protesters **hound** the representatives* [verb]), only unambiguous nouns are included. Word-class annotations were taken from the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012).

i. Critical predictors. I calculate the conditional entropy $H(D \mid L)$ for each noun from the sample. The component entropies – the entropy of non-target words H(L) and the joint entropy of words and dependencies $H(L, D)$ – were estimated using the Open American National Corpus (OANC)[2], a freely available, approximately 15-million word collection of American English writing and transcribed speech from many different genres and registers. First, I parsed the OANC using the spaCy dependency parser (Honnibal & Johnson, 2015). Then, for each of the target words, I generated 18 frequency distributions, two for each of the syntactic spaces in Figure 3. One of the distributions in each pair reflects the frequencies of the non-target forms that are bundled with the target $f\,(L)$. The other reflects the joint frequencies of non-targets and the dependencies that bind them to the target $f(L, D)$. Next, I compute the entropies of the distributions. I correct the entropies for underestimation bias using the Chao-Wang-Jost method (Chao, Wang, & Jost, 2013) before subtracting H(L) from H(L, D). Because I correct the entropies prior to taking the difference, some distributions

[2] http://www.anc.org/OANC

may produce negative conditional entropies (which is impossible for true populations). These obviously incorrect values reflect uncertainty given the limits of the sample. However, they are not useless if we shift our focus to the relative magnitudes that distinguish these from the other observations. With a reasonable sampling rate (e.g., at least 50 tokens), we should not expect that negative conditional entropy estimates would be generated for distributions that do not actually fall in the lower end of the entropy range. Next, I compute the categorical measures. I begin with the raw frequency vectors for each of the nine dependency-only distributions *f*(D). I count the number of dependencies with frequency > 0 in *f*(D) for each noun. Finally, I compute the prototypicality measures, likewise on the basis of *f*(D).

The nine syntactic spaces that I consider are necessarily intercorrelated. For example, the total distribution of a word across syntactic dependencies are decomposable into the head and modifier distributions[3]. This is true for both probabilistic and categorical measures. If these correlations are strong, a situation known as *multicollinearity*, statistical models can struggle to apportion explained variance across the correlated predictors (e.g., Baayen, 2008). Multicollinearity leads to untrustworthy parameter estimates and significance tests; it violates the assumption of the independence of error across predictor terms that is required for most regression techniques (Chapter 2 of Zuur, Ieno, Walker, Saveliev, & Smith, 2009). Because these measures are collinear, they cannot be compared directly within the same model. Therefore, we need some way to extract the independent sources of information that

[3] Technically, $H_t = H_h + H_m + H_{c,}$ where H_c is the entropy of the choice between head and modifier.

are distributed across the predictors. I tease apart these latent sources of information using Independent Component Analysis (ICA). The first step of ICA is to rotate (*whiten* or *sphere*) the raw variables (i.e., the 'mixed signals') to remove correlations between them. This step creates maximally Gaussian relationships among the dimensions of the PCA space. Then, the whitened space is rotated to maximize non-Gaussianity. Following the logic of the Central Limit Theorem, the mixture of independent source signals will be more Gaussian than any of the sources individually. Therefore, the rotation that produces the least Gaussian space captures the most non-Gaussian (i.e., non-random) structure between the variables. Crucially, the resulting components are fully statistically independent. The positions of words within the new component space(s) can now be used to predict RTs. The meaning of the components can be interpreted relative to the so-called *mixing matrix,* which contains the co-efficients needed to project each word from the raw multi-dimensional space into the doubly rotated component space. The higher the absolute value of the coefficient between a raw predictor and independent component, the stronger the relationship between that predictor and component. Predictors may differ in the signs of their coefficients, allowing for contrasts to appear within the components themselves.

I performed three ICAs, one over each set of nine measures: categorical, probabilistic, and prototypicality. I used the FastICA algorithm (Hyvärinen & Oja, 2000) as implemented in the *R* package *fastICA* (Marchini, Heaton, & Ripley, 2013). The algorithm can be used to generate any number *n* of components. I estimate an appropriate minimum *n* using Principle Component Analysis (PCA). An important difference between PCA and ICA is that for the former, the extracted components must be orthogonal (based on a chain from the first or principal component – the most variance explained – to the second, third, and so on). I define the number of independent components n to be the number of PCA components needed to achieve 95% cumulative explained variance. The PCA performed on the probabilistic measures showed that six orthogonal components explain 95% of variance. I therefore extracted six independent components. I followed the same procedure for the categorical measures and found that a single component captures ~ 93% of the variance. The second component added only 3% explained variance, and its factor loadings were virtually indistinguishable from those of the first component. I therefore extract one independent component. Finally, I perform a PCA on the prototypicality measures. Based on the results, I extract three independent components via ICA.

Pairwise scatterplots of the ten components (6 probabilistic + 1 categorical + 3 prototypicality) suggested no substantial correlations. This suspicion was confirmed statistically: the measure of collinearity k fell well within the acceptable range ($\kappa = 4.18$; Baayen, 2008, suggests that $k < 30$ indicates no serious collinearity). Therefore, these variables can be entered as competitors within the same model. Importantly, this allows us to (a) compare the categorical and probabilistic hypotheses and (b) treat the prototype and diversity effects as independent functional components of lexical recognition (Linzen et al., 2013).

ii. Control variables. A number of variables are known to impact response latencies in visual lexical decision (VLD). Therefore, it is necessary to exclude these factors as possible alternative explanations for any relationship between the target variables and the ELP RTs.

These variables include

— *word frequency*

— *orthographic similarity*

— *orthographic word length*

— *age of acquisition*

Together, these variables are known to account for the bulk of unique variance (> 40%) compared to other relevant but weaker predictors (~2%; Brysbaert et al., 2011).

Prior research has shown that *word frequency* is the strongest predictor of VLD RTs (though see Baayen, 2010, for a discussion of the ultimate sources of this effect). In particular, word frequencies derived from movie subtitles (SUBTLEX-UK; van Heuven, Mandera, Keuleers, & Brysbaert, 2014) perform the best, explaining more of the RT variance than even the carefully balanced, 100-million-word British National Corpus. For that reason, I include the SUBTLEX-UK frequencies as a predictor. There has been some debate about whether to use surface or lemma frequencies. The former refers to the number of observations of a single string (e.g., *float*), while the latter refers to the sum of the surface frequencies for all inflectional variants of a word (e.g., *float, floats, floating, …*). However, recent work has shown that (a) lemma and surface frequencies are highly correlated (r > .9), (b) they have an almost identical effect on RTs (Brysbaert & New, 2009), and (c) surface frequencies are robust predictors of RT for low-frequency words while lemma frequencies are not (Baayen, Wurm, & Aycock, 2007). Because the frequency distribution of any set of words will carry a strong positive skew, I take the logarithm of the frequency.

Another relatively strong determinant of visual word recognition is the formal (orthographic) similarity of that word to other words in the lexicon. Similarity may be operationalized in a number of ways (e.g., mean letter bigram frequency, Coltheart's *N*). However, Markoni, Balota, & Yap (2008) show that their measure – the *orthographic Levenshtein distance* or OLD20 – accounts for the largest amount of variance in word recognition RTs. OLD20 reflects the average number of insertions, substitutions, or deletions that would need to be made to a word to produce its 20 closest orthographic neighbors (i.e., the average Levenshtein distance). A low value means that the spelling of the word overlaps a great deal with other words in the lexicon; a high value means that the form of the word is rather idiosyncratic. OLD20 correlates positively with word recognition latencies (Markoni et al., 2008), meaning that people are faster at recognizing words that are similar in form to many other words. In light of these facts, I include OLD20 as a control predictor (estimates taken from the BLP annotation).

Word length has proven to be a less reliable predictor of recognition latencies. In some cases, it has been shown not to exhibit any effect (O'Reagan & Jacob, 1992); elsewhere, it has been shown to be inhibitory (at least in the longer extremes; New, Ferrand, Pallier, & Brysbaert, 2006); and in some cases, it has been shown to depend on other variables, such as frequency (with inhibitory length effects surfacing only for low-frequency words) or age (with length effects surfacing only for older participants; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004). New et al. (2006) uncovered a U-shaped effect of orthographic length. Each additional character in shorter words (< seven characters) actually *facilitated* response latencies. By contrast, each additional character for longer words (> seven characters) was inhibitory. While the source (and shape) of the orthographic length effect remains controversial, the majority of studies suggests that it is an important determinant of visual lexical devision latencies. Therefore, I include length in characters as a co-predictor in the statistical analysis.

Another variable proposed to affect visual word recognition is the age at which words are typically acquired by native speakers, or *age of acquisition* (Morrison & Ellis, 1995; Cortese & Khanna, 2007; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; cf. Zevin & Seidenberg, 2002, for limitations). Generally speaking, the earlier a word is acquired, the faster the word will be recognized. I control for this effect by including the mean subjective age of acquisition ratings collected by Kuperman et al. (2012). Subjective ratings reflect how old people think they were when they first learned a word. These ratings have been found to be largely consistent across participants and to correlate strongly with lexical decision (Kuperman et al., 2012).

iii. Response time data. Experimental data were taken from a previously published database of visual lexical decision RTs (Balota et al., 2007). Participants in that study completed approximately 3,400 lexical decision trials in two sessions, each broken into blocks of 250 items. Words were drawn from a master list of over >89,000 mono- and polymorphemic forms. Sublists for each block were controlled so that no single lexical root was viewed too many times (e.g., *joy, enjoy, enjoyable*, etc. were split across blocks). Nonwords were constructed by changing one or two characters in the target words, as long as the resulting form was plausible given English spelling conventions. Feedback on accuracy and speed was provided. RTs and accuracies were recorded.

5. Results

I fitted a generalized additive mixed model (GAMM) with the *R* function *bam* from the *mgcv* package (Wood, 2016) predicting RTs from the ELP. Prior to fitting the model, the RTs were log transformed to remove a strong positive skew. Such transformations are not inherently necessary in additive models; however, model residuals were substantially improved by taking the logarithm. Only trials with RTs within 1.5 times the interquartile range of the mean were included in the analysis. A pilot model revealed persistent underestimation problems for faster RTs; this issue was solved by discarding trials with RTs < 350 ms. This trim successfully corrected the underestimation, yielding approximately normally distributed model residuals. Only accurate trials were considered (i.e., trials for which the participant correctly identified the target noun as a word). This left 17,113 observations of 584 noun types across 815 participants. Spline-based smooths were applied to *word frequency, orthographic similarity,* and *age of acquisition*, along with the ten ICA predictors, to account for possible non-linearity of the effects. *Orthographic length* was treated as linear because it offered too few distinct values to accommodate the smooth. To account for autocorrelative effects, I include two terms based on how the participant performed on the immediately prior trial: a parametric term for accuracy and smoothed term for RT. Further, I include factor smooths for overall trial number per participant. Random intercepts were allowed by item.

To balance explanatory power against parsimony, I conducted a backward model selection informed by the method of Zuur et al. (2009; Appendix A). Selection was only applied to the critical predictors; control predictors were left intact. I began with the maximal model and proceeded to remove each non-significant critical predictor whose removal reduced the Akaike Information Criterion (AIC) the most. I continued this process until only significant critical predictors remained. The resulting model is summarized in Table 2 below.

Table 2: Summary of GAMM predicting ELP response times.

Parametric terms	***β***	**Error**	***t***	***p***
intercept	6.34	.02	389.04	<.001
orthographic length	.01	.003	2.96	<.01
prior accuracy	.01	.004	1.73	.08

Smooth terms	***eDF***	***refDF***	***F***	***p***
SUBTLEX frequency (log)	4.06	4.61	20.12	<.001
age of acquisition	1.00	1.00	70.49	<.001
OLD20	1.00	1.00	0.06	.81
prior RT	7.99	8.71	372.98	<.001
trial number by participant	21.82	25.72	7.66	<.001
within-stimuli variance	275.01	576.00	.92	<.001
categorical component	**1.00**	**1.00**	**5.42**	**<.05**
probabilistic component 6	**1.00**	**1.00**	**7.37**	**<.01**
prototypical component 1	**2.11**	**2.43**	**5.14**	**<.01**

Table 2 shows the coefficient estimates, standard error, *t* values and *p* values for the parametric terms. Smooth terms are provided with the expected and residual degrees of freedom, *F* values, and *p* values. Critical predictors are shown in bold.

First, I consider the controls related to the experimental design and procedure. As evidenced by the significant effect of the within-stimuli smooth, the nouns differed in the extent to which they differed from the group mean, all else being equal. Hence, some variability among the words lies beyond the control measures taken here (a notion supported by the adjusted R^2 of .26). The significant sequence-by-participant factor smooth indicates strong autocorrelative effects that differed in shape across individual participants. In other words, subjects responded differently to cumulative experience with the task. Narrowing in on sequential effects, I find that RTs increased with RTs from the previous trial. When subjects struggle to make a lexical decision, that struggle bleeds over into subsequent trials.

Orthographic length surfaced as significant, while orthographic neighborhood density (OLD20) did not. These predictors are highly correlated ($r = .79$, $p_{Pearson\text{-}Product\text{-}Moment} < .001$). Therefore, the lack of an OLD20 effect could be due to interference from length. I refit the model with length but not neighborhood density and neighborhood density but not length. In both cases, whichever variable I left in surfaced as highly significant ($p < .001$). According to the AIC, which measures model fit against model complexity, the model with orthographic length alone (AIC = -6987.78) should be preferred over that with OLD20 alone (AIC = -6983.84). No other effects were substantially altered by omitting either of these variables.

The other two item-specific controls were highly significant, as well. As expected, word frequency was strongly negatively correlated with RTs: more frequent words were recognized faster. Also as expected, subjective age-of-acquisition estimates were positively correlated with RTs: words that people feel they learned later in life are recognized more slowly.

The model uncovered significant effects for three of the ten critical predictors: the categorical component, probabilistic component six, and prototypicality component 1. I consider each in turn. Figure 5 plots the loadings for the categorical component (left panel) and the effect of the component on RTs (right panel). This component loads in the same direction for all distributions, indicating a general diversity effect. Loadings are heaviest for headship and total diversity. The smallest contributor is rightward modifiership. Scores for this component correlate negatively with RTs: words that appear in more syntactic dependencies are recognized faster. From extreme to extreme, this benefit covers an approximately 30 ms window (though the specific magnitudes are not at issue here). I therefore find initial support for the categorical hypothesis.

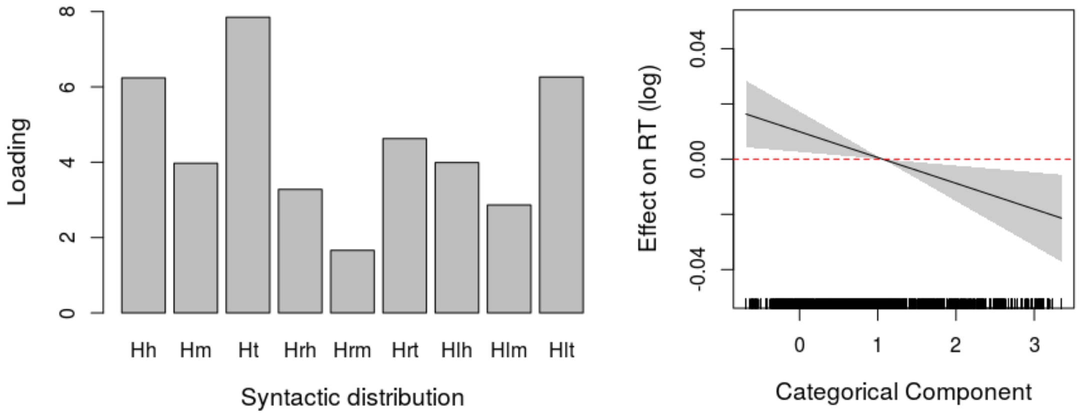


Figure 5: Significant effect of categorical diversity.Left panel: Component loadings of the single categorical component. Right panel: Effect of the categorical component on RTs. Y-axis shows the effect of the component on (log) RTs. The range of {-0.05, 0.05} is equivalent to a range of approximately {-40, 40}in milliseconds. Positive values indicate slower than average RTs while negative values indicate faster than average RTs. The dotted line indicates no effect. Shaded areas indicate 95% confidence intervals Density of observations are indicated by a rug along the x-axis.

Over and above the categorical effect, I found a significant effect of probabilistic component six. The component loadings and effect of this component are plotted in Figure 6. The loadings show that this component contrasts general modifier, general rightward, and rightward modifier diversities from the other measures. The co-loading of these variables suggests that words that score negatively on component six are distinctively associated with diverse rightward modifier distributions. Conversely, words that score positively on component six are those that distinctively eschew diversity as rightward modifier, but pursue it elsewhere in the system. Scores from component six correlate negatively with RTs, meaning that distinctively diverse rightward modifiers are recognized more slowly than words that avoid those structures. The relationship is nonlinear; the negative correlation is most pronounced for words falling in the negative range of the component scores. The relationship attenuates sharply around zero (until the density of observations drops off between scores of 2 and 3). These findings support the probabilistic hypothesis.

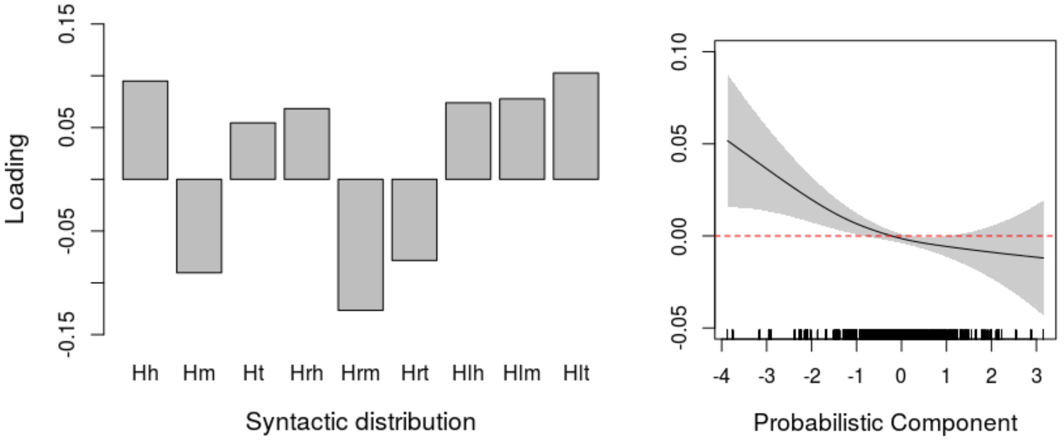


Figure 6: Significant effect of probabilistic component. Left panel: Component loadings of probabilistic component 6. This component contrasts distinctively diverse rightward modifiers (negative values) from everything else (positive values). Right panel: Effect of probabilistic component 6 on RTs (range in ms = {-40, 80}).

Finally, I found an independent effect of prototypicality. The left panel of Figure 7 shows the loadings for prototypicality component one. Similar to the categorical component, all measures load in the same direction, suggesting that distance from the prototype manifests itself in the same general way across all syntactic measures. Also similar to the categorical component, rightward modifiership stands out. Where it contributed the least to diversity, it contributes the most to distance from the noun prototype. The right panel plots the effect of this component on RTs. As expected by the prototypicality hypothesis, nouns that are more distant from the prototype are recognized more slowly.

6. Discussion

The results support all three of the hypotheses proposed above. Nouns that appear in more syntactic contexts are recognized faster (categorical hypothesis); nouns that are distributed more uniformly across these contexts are recognized faster (probabilistic hypothesis); and more prototypical nouns were recognized faster (prototypicality hypothesis). As predicted by the neurophysiological findings of Linzen et al. (2013), syntactic diversity and prototypicality showed independent, additive effects. However, unlike Linzen and colleagues, both types of effect were observed for RTs. This finding suggests that the dependency-based measures give more accurate estimates of the syntactic diversity of words.

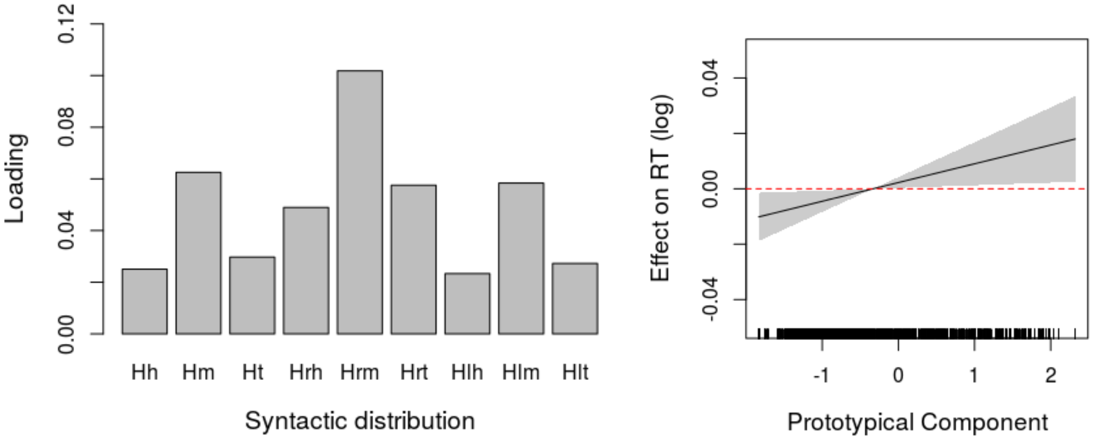


Figure 7: Significant effect of prototypical component. Left panel: Component loadings of prototypicality component 1. This component reflects general distance from the prototype. Distances in the modifier and particularly the rightward modifier distributions are prioritized. Right panel: Effect of probabilistic component 6 on response times.

Several novel effects were also observed. First, syntactic diversity breaks down into additive effects of categorical and probabilistic diversity. Second, the diversity and prototypicality effects depend most heavily on rightward modifier dependencies.

This pattern of findings is inconsistent with theories that posit only categorical syntactic representations in the lexicon. In these theories, words either are or are not licensed in a particular structure (e.g., Chomsky, 1995). These theories could account for the categorical diversity effect observed here. For example, words that activate more syntactic categories are processed faster, perhaps through a feedback mechanism. However, if this account were correct, we should not have seen effects from probability and prototypicality (both of which depend on frequency distributions). But we did see such effects, indicating that these theories are incomplete and underpredictive. Looking closer, we see that many of these theories are also unable to account for the modifier-driven diversity effect. Often, categorical theories only mark words for the structures that they may head (e.g., Bresnan, 2001; Chomsky, 1995). But the model revealed that the as-modifier diversities contributed the most to the categorical effect (Figure 5, left panel).

These findings also differ from those observed by Linzen et al. (2013) for verbs. In that study, they found no effect of either diversity or prototypicality on RTs. Other work on nouns has reported a syntactic effect; but the measures used there were actually based on lexical variation (Baayen et al., 2011). Lexical variation is known to reflect semantics (Bullinaria & Levy, 2012). Therefore, it was possible that the findings for nouns were tainted by semantics. Linzen and colleagues were the first to use fully abstract, cross-structural syntactic diversity. Therefore, it was possible that similarly abstract measures applied to nouns would likewise show no correlation with RTs. However, the opposite was true: syntactic distributions impact the processing of isolated words. This discrepancy could stem from at least three differences between the study of Linzen and colleagues and this one. First, they based their measures on phrase-structural subcategorization frames, whereas as I based mine on binary dependencies. From a construction-grammar perspective, both levels should be involved simultaneously: an argument-structure construction embodies the entire argument configuration, as well as the lower-level constructions that fill out the individual arguments (Goldberg, 1995; Langacker, 1987). The dependencies studied here correspond to those lower-level relationships. Perhaps these are more intimately tied to word processing given that they mark the point of entry for words into constructional frames. Second, I carefully corrected the entropy estimates to avoid underestimation biases. Noise attributable to biased estimates could have interfered with the model estimates for Linzen and colleagues. Third, regarding diversity, I made sure to remove all lexical information from the entropy estimates. Linzen and colleagues did not control for lexical information. This lexical information could interfere with the estimate, again obscuring the effect.

Several theories can explain all three effects. Usage-based construction grammar (UBCG; Diessel, 2015; see also Goldberg, 2006) proposes that language is best modeled as a complex network of interactions between linguistic units at all levels of abstraction. Categorical specifications on word forms are replaced by arcs between word-level and syntax-level nodes. Statistical information derived from the input tunes the strength of these connections, thus accounting for the probabilistic diversity effect. When these connections and connection strengths match up with the expectations of the system, words are recognized more quickly. System expectations can be modeled in several ways. For example, patterns of resting activation may develop over time based on the average behavior of the system. When a noun deviates from this pattern, it will not benefit as much from the activation that is fed back from the syntactic system. Alternatively, prototypical nouns, which better signal their membership to the noun category, may provide more compelling evidence to the system responsible for making the two-way lexical decision judgment. This could surface as input into the drift space between two choices (e.g., Ratcliff et al., 2004) or as an influence on prior probabilities of encountering a noun with such-and-such syntactic behavior (Norris, 2006; see Linzen et al., 2013, for a similar suggestion).

The discriminative-learning model of Baayen et al. (2011) could also account for these findings. According to this model, variability of contextual cues over time helps to solidify the bond between a target form and its meaning. Syntactic dependencies constitute one form of contextual cue. These cues are also thought to be paradigmatically bound, such that the information carried by a paradigm can influence discriminability of the connection between form and meaning. For example, the unconditioned distribution of prepositions (the paradigm) in English prepositional phrases represents the generalized potential for a preposition to be followed shortly by a noun. Nouns that function as objects to prepositions in proportion to the overall distribution of those prepositions will be more likely to surface when a preposition has been deployed (e.g., inversely proportional nouns would load too heavily on uncommon preposition types). Therefore, they stand to benefit the most from the information carried by the 'prepositional paradigm,' where that information comes in the form of structured contextual variability. The variability produces stronger, more stable connections between form and meaning, leading to more efficient reading. This, they argue, is the source of the typicality effect they observe.

The primary evidence for discriminative learning comes from measures based on overt cues – cues that are directly available in the input, for example, inflectional endings or prepositions. These types of measure bias the results in favor of the model. This is because the model developed by Baayen and colleagues contains only two layers: an input layer for letter *n*-grams (usually 2- or 3-grams) and an output layer for meaning. Therefore, they are naturally capable of modeling lexical variation, but ostensibly incapable of modeling variation for abstract categories. However, the measures considered here do not make any direct reference to lexical context. They are based solely on the syntactic dependencies that attach to the target noun. Information about the word to which the target is bound was explicitly removed. Therefore, these results constitute a challenge to the purely syntagmatic discriminative learning approach. Specifically, it appears that hierarchical, *non-overt* aspects of the contexts in which nouns appear also affect how well they are learned. We should therefore expect learning-through-discrimination to involve a multidimensional network of cues, including cues directly associated with the surface code (words built from graphs, phones, or signs) and higher-order, more abstract cues that emerge over time (e.g., Bybee, 2010; Diessel, 2015; Goldberg, 1995).

The contrast between rightward and leftward modifer distributions deserves further comment. Diverse and distinctively rightward modifiers were recognized more slowly, while diverse leftward modifiers were recognized faster. Why would diversity help in one context and hinder in another? Consider the nature of rightward modification. Nouns that modify words to their right participate in a head-final dependency. However, English has dominant head-initial word order, at least outside of the noun phrase (NP). Notice that NP-external relations of this sort are precisely where modifier relationships apply for nouns. Therefore, nouns with negative scores on this component fight against the typological orientation of English nouns as modifiers. Typological constraints like this should leave other traces. For example, they should affect word frequency. One would expect to find more words with low conditional entropies in the dispreferred dimension. Hence, the probability density function for rightward noun-as-modifier relations should be bunched up around 0. The density function for leftward noun-as-modifier relations (head-initial ordering) should be spread more evenly across the range of entropy values. Figure 8 plots the probability density functions for the rightward and leftward variants of the noun-as-head and noun-as-modifier conditional entropies.

Figure 8 supports this intuition. Words are clustered clustered below $H(D \mid L) = 0.5$ for the typologically dispreferred dimension, rightward diversity for nouns-as-modifiers (shown in orange). The greatest peak in density centered on 0. By contrast, the typologically preferred dimension –leftward diversity for nouns-as-modifiers – has a wide distribution, with higher rates of occurrence in the upper ranges of conditional entropy (up to ~1.5, a full bit higher than that observed for rightward modifier diversity).

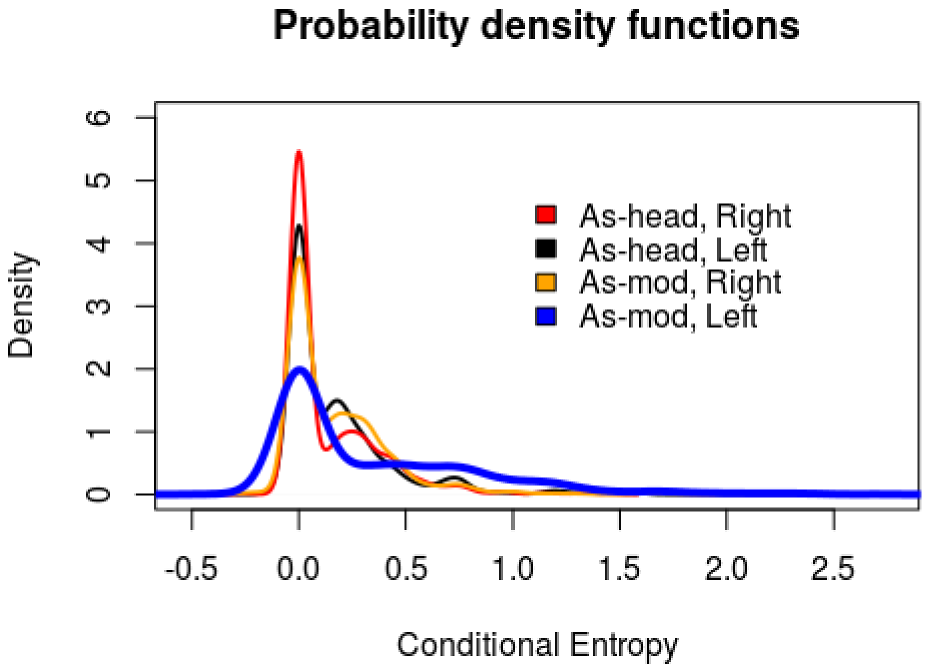


Figure 8: Probability density functions for the conditional entropies of nouns. Different colors reflect different syntactic dimensions.

Typology should also relate to prototypicality: nouns that are diverse rightward modifiers should be atypical, hence less like the other nouns of the language. By definition, the majority of other nouns would follow the dominant head-initial preference (a point supported by the curves in Figure 8). Therefore, the processing disadvantage associated with rightward modifiership might actually be due to a typological prototype favoring head-initial structures. This intuition is supported by the prototypicality effect observed here. The general prototypicality effect was most strongly driven by rightward modifiership. Distance from the noun prototype was associated with longer RTs. The common thread is that distributions in the rightward modifiership space exert the strongest effect. Following the logic of discriminative learning, nouns of this type would not have the same general opportunity to occur and hence would not receive the same discriminative benefit as words that fit the overall trend. This explanation directly links syntactic typology to the local word processing. Such a link clears the path for new predictions regarding the behavior of typologically distinct languages. For example, we should observe opposite effects in strongly head-final languages, such as Japanese: the processing disadvantage should emerge for leftward-facing noun-as-modifier diversity.

The primary take-away from this study is that reading a noun in isolation invokes the syntactic history of that word. While similar results have been observed before, this study is the first to demonstrate that purely syntactic distributions impact lexical decision RTs. These data falsify any theory that limits syntactic representation in the lexicon (e.g., Borer, 2005; Chomsky, 1995; Marantz, 1997; Ramchand, 2008). These data also provide converging support for the notion that syntactic information obligatorily impacts lexical access, regardless of task or whether the word is processed in isolation (e.g., Cubelli et al., 2005; de Simone & Collina, 2015; Lester & Moscoso del Prado Martín, 2016; Linzen et al., 2013). Even when the task requires no syntax, syntactic information impacts RTs. Finally, they underscore the need to decompose lexical frequency well beyond the typical type/token counts (Baayen, 2010; cf. Bybee, 2010). Each instance of a word is embedded in a multidimensional network of cues. Which cues are important to which tasks, and how frequency relates to these cues, are questions that may reveal important information about the processing mechanism.

One unexpected finding of the present study was the typological contrast in diversity effects. Words that match the typological properties of language are recognized more quickly than those that go against the grain. This question deserves further study. One possible extension would be to compare the size of these typological congruence effects across languages. I expect languages that have cross-linguistically less preferred orders in a given domain to benefit less from congruence than languages with more preferred word orders, irrespective of whether the system is consistent within those languages. Such studies would help to solidify the links between linguistic representation, processing, and typology.

B. Primed lexical decision

Lexical priming in visual lexical decision has a long history. Many models have been proposed, but the scope of the effects they have sought to explain has been surprisingly limited from the global linguistic perspective. Perhaps the two most important strands of research have concerned manipulations of the orthographic and semantic relationships between prime and target. Less studied but equally important are relationships between orthographic or semantic similarity and the morphological locus of these effects. However, because of the nature of the visual lexical decision task, syntax has been largely ignored. Researchers have generally assumed that determining whether a letter string is a legitimate word should not involve the syntactic system at all (what good would it do, if no clausal or phrasal parsing is required?). However, this assumption may not be warranted. For example, usage-based linguistic theory argues for direct connectivity between words and syntactic structures. These connections are in principle no different from those which bind word forms to conceptual-semantic representations. For example, Diessel (2015) argues that grammar can be captured by a network of relations in which any two nodes – including word forms and syntactic structures – may in principle become associated given appropriate statistical properties of live language use (production or comprehension in any modality).

The notion that syntactic information may be present in the lexicon is one of the rare points of agreement among grammar formalisms in linguistics. For example, even generativist approaches to syntax (e.g., Chomsky, 1995), which otherwise propose a strict divide between grammar and lexicon, acknowledge that lexical items must be specified for syntactic categories. For example, the syntactic operation responsible for constructing noun phrases of the form NP → DET N (as in *the waffle*) must select all and only words of category N to fill the second slot. In order to preclude ungrammatical sequences (e.g., **the into*), words must 'display' their membership in the appropriate category to the syntactic generator. Hence, all words are expected to contain information about syntactic categories relevant for the application of formal combinatorial rules. However, usage-based models go further, suggesting that words and syntactic structures are *directly* and *statistically* related based on one's experience with language (e.g., Diessel, 2015).

Converging evidence for these relationships comes from several sources, including language acquisition (Tomasello, 2003), lexical contributions to constructional meaning (e.g., Stefanowitsch & Gries, 2003), and semi-productivity of syntactic constructions (Goldberg, 2006; Zeldes, 2013). Importantly, such relationships have been demonstrated in online processing, even in a putatively non-syntactic task. Lester and Moscoso del Prado Martín (2016) showed that production latencies in a bare-noun picture-naming task were sensitive to the diversity of the probability distributions of the target names across the syntactic relations (as estimated on the basis of a large, syntactically annotated corpus of English writing). In the present study, I extend this research by asking two questions. First, do these syntactic diversity effects likewise surface in comprehension? And second, are these syntactic relations shared across lexical items, such that (dis-)similarity of the distributions between words will influence response times in a priming task?

1. Methods

To answer these questions, I first define a common syntactic space based on low-level syntactic relationships. Next, I estimate the frequency distributions of nouns within the syntactic space. I define a measure of distance in that syntactic space and compute this measure for prime—target pairs in a previously published database of primed visual lexical decision latencies. Finally, I correlate these distance measures with the target response latencies.

i. Data. Behavioral data come from the Semantic Priming Project (SPP; Hutchison, et al., 2013). The SPP contains response times and accuracies, along with a host of norming data, that were collected using a visual lexical decision task with overt orthographic priming. On each trial, participants were shown a centered fixation cross for 500 ms, followed by a prime word (all caps) for 150 ms. The prime was followed by a blank screen lasting either 50 or 1050 ms (the *interstimulus interval*, or *ISI*). Finally, the target word was displayed (all lowercase) until a decision was made or 3,000 ms elapsed, at which point the experiment would advance to the next trial.

I take only the trials containing primes and targets that also appear both in the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and the age of acquisition norming database of Kuperman, Stadthagen-Gonzalez, & Brysbaert (2012). I limit the data in this way to take advantage of the additional lexical controls afforded by these databases. I further limited the trials to include only those for which string-identical tokens of both prime and target received majority noun tags in the British National Corpus (BNC). I do so to minimize non-noun interpretations of the (potentially ambiguous) strings. This procedure leaves us with 1,305 unique primes and 821 unique targets (a total of 1,670 unique nouns).

ii. Defining a syntactic space. Now that we have a set of nouns, we can measure the relationship between these nouns and the syntactic system. To do so, we must first define the scope of that syntactic system. At least a century of research have failed to produce an exhaustive list of the syntactic constructions of English (much less any other language), and I do not presume to offer such a list here. Instead, I rely on the set of low-level relations as defined within Dependency Grammar formalisms (e.g., Hudson, 2006; Mel'čuk, 1988; Nivre, 2005; Tesnière, 1959). Dependency Grammars differ from the more commonly employed phrase-structure grammars in that they model relations (*dependencies*) between pairs of words only. These relations are asymmetric: each extends from a *head* (the syntactic and conceptual core word) to a *modifier* (whose appearance is contingent on the head). Each dependency is labeled to reflect its syntactic function. For example, *the* and *waffle* in the noun phrase *the waffle* would be bound by the *det* relation, which attaches a determiner (*the,* the modifier) to a noun (*waffle,* the head). Other examples include the *nsubj* relation, which binds a noun (modifier) to a verb (head) as its subject, and the *pobj* relation, which binds a noun (modifier) to a preposition (head) as its object. Much more can be said about these relations and constraints on their implementation in broader phrasal and clausal contexts. However, such questions are beyond the scope of the present study. I adopt the dependency notation as implemented in the *spaCy* parser (Honnibal & Johnson, 2015). I do so primarily for practical reasons: spaCy provides one of the fastest and most accurate dependency parsers on the market (compare e.g., the Stanford CoreNLP toolkit; Manning et al., 2014).

I define the syntactic space for nouns as the set of dependencies for which at least one noun from the sample of SPP primes and targets has been attested either as head or as modifier. I accomplish this in several steps. First, for each noun that appears both in the SPP and the BLP, I extract all sentences containing that noun from the BNC. Using the simplified CLAWS5 annotation (via the XML corpus reader provided in the Natural Language Toolkit; Bird, Klein, & Loper, 2009), I condition the search to include only sentences in which the word form was indeed tagged as a noun. Next, I parse those sentences in CoNLL format using the *spaCy* dependency parser (Honnibal & Johnson, 2015). I then compute the frequency distribution of each noun across the dependencies for which it serves as head or modifier. To increase the reliability of the frequency estimates, I discard vectors for all nouns that occurred in fewer than 100 sentences in the BNC (~1 per million words). The total syntactic space is defined as a vector in which each column reflects one of the set of unique dependencies occurring across all nouns. Finally, I merge the individual frequency distribution of each noun into the total syntactic space, creating a matrix of *n* rows by *m* columns, where *n* equals the number of total unique dependency types (46) and *m* equals the number of unique SPP/BLP nouns (1,241). The result is therefore a uniform syntactic space for all nouns, where individual nouns may or may not be attested in each possible dependency. In theoretical terms, I treat these vectors as reflecting the statistical connectivity between each noun and the syntactic structures it inhabits, as proposed in the usage-based literature. Psycholinguistic support for this treatment comes from an earlier study showing that these and similar dependency vectors affect processing latencies in noun production independently of other factors, such as token frequency (Lester & Moscoso del Prado Martín, 2016).

iii. Measuring syntactic similarity. We are interested in the possibility that pre-activation of shared syntactic representations will influence the speed of word recognition. Therefore, we need some measure of the similarity between the syntactic distributions of primes and targets in the behavioral data. Note that similarity in syntactic space outlined above does not reduce solely to shared *types* of dependencies. For example, consider two words, *w1* and *w2,* that occupy the same set of 20 dependency types. Suppose that *w1* and *w2* have roughly equivalent overall frequencies and that those frequencies are distributed equally across the dependency types for both words. In this case, we would call them syntactically similar, and consider the number of overlapping types as an appropriate measure of the strength of their similarity. Now suppose that the two words have similar overall frequencies, but that these frequencies are distributed over complementary sets of the dependencies that they share, such that *w1* has a frequency of 1 wherever *w2* has a frequency >100 and vice versa. In this case, we would call them dissimilar; crucially, however, the type-based metric could not tell us this. Thus, we need some way of accounting simultaneously for shared types, as well as similar apportioning of the probability mass across those shared types. One measure well suited to this task is the Jensen-Shannon Divergence (JSD; Lin, 1991). JSD is a symmetric variant of the Kullback-Leibler Divergence (KLD; sometimes referred to as the *relative entropy*). The KLD between two probability distributions *P* and *Q* is defined as follows (Eq. 5):

$$\mathrm{KLD}(P||Q) = \sum_i \quad P(i)\log\frac{P(i)}{Q(i)} \tag{5}$$

This measure captures the average amount of additional information that one would need in order to recode an event from distribution *P* as if it belonged to distribution *Q*. Importantly, KLD($P\|Q$) ≠ KLD($Q\|P$), meaning that one must decide *a priori* in which direction to take the distance. JSD provides a solution to the asymmetry problem by taking the midpoint between the two distributions, then taking the mean distance of the distributions to the midpoint. Formally, JSD is expressed as follows (Eqs. 6 and 7).

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{KLD}(P||M) + \frac{1}{2}\mathrm{KLD}(Q||M) \tag{6}$$

where

$$\mathrm{M}=\frac{1}{2}(\mathrm{P}+\mathrm{Q}) \tag{7}$$

Using this technique, JSD(P‖Q) = JSD(Q‖P), with values bounded such that $0 \leq \mathrm{JSD} \leq 1$.

In the most typical case (and in the present study), JSD measurements depend on *estimates* of the probability distributions of events within a distribution, not the *actual* probabilities. Any given frequency estimate necessarily grows with sample size, which means that the so-called *maximum-likelihood estimates* for any given sample size are sure to underestimate the true probabilities. Furthermore, this effect will impact lower frequency items more heavily than higher frequency items. To guard against this bias, and the attendant frequency confound, I smooth the frequency vectors. Because the comparison of any two distributions P and Q via JSD requires that they share the same number of cells, I select the James-Stein shrinkage estimator (Hausser & Strimmer, 2009). This smoother is optimal for vectors for which the number of cells is known (here, I assume that the observed set of dependencies is exhaustive, but we know that this is probably not the case; however, note that the dependency vectors will grow uniformly in number of cells across noun types as new dependencies are uncovered).

We now have a means of formalizing the syntactic similarity between primes and targets for the SPP data. For each prime—target pair in the sample, I compute the JSD between them. A value of 0 indicates identity; a value of 1 indicates complete independence. However, we now face a broader issue. According to usage-based theory, (at least the bulk of) syntactic structure is meaningful – that is, directly linked to semantic representations in the same way as words (e.g., Diessel, 2015). This means that any effect we uncover for this measure may actually reflect *semantic* similarity, which is well known to impact response times in lexical priming (e.g., Neely, 1991). Fortunately, the SPP contains annotation of the degree of semantic similarity between prime and target; cosine similarity in the Latent Semantic Analysis space (LSA). LSA measures the extent to which words occur in similar stretches of text, with higher cosine values indicating greater similarity (for a detailed discussion of this approach, see Landauer & Dumais, 1997). To keep things simple, I transform the cosine measures in SPP by subtracting them from 1. This way, both the transformed semantic measure and the syntactic measure reflect distance, such that increasing values correspond to decreasing similarity. These measures have the added advantage of both scaling from 0 to 1, allowing us to compare the relative strength of their effects on response times.

Figure 8 shows the relationship between the JSD (y-axis) and LSA (x-axis) values for the present sample. As expected there is a slight but significant positive (linear) correlation, meaning that words that are similar in meaning tend to surface in the same syntactic contexts. While not central to the present study, an important feature of Figure 8 is the triangular shape of the variance: words that are very close in meaning vary only slightly in syntactic similarity, while words that are distant in meaning vary more widely. This relationship supports the account of Jackendoff (2013), who argues for the existence of syntactic generalizations (i.e., constructions) that allow structural inheritance among sets of semantically heterogeneous sub-constructions. At the very least, it suggests that syntax and semantics are not as tightly coupled as some would argue (e.g., Goldberg, 1995).

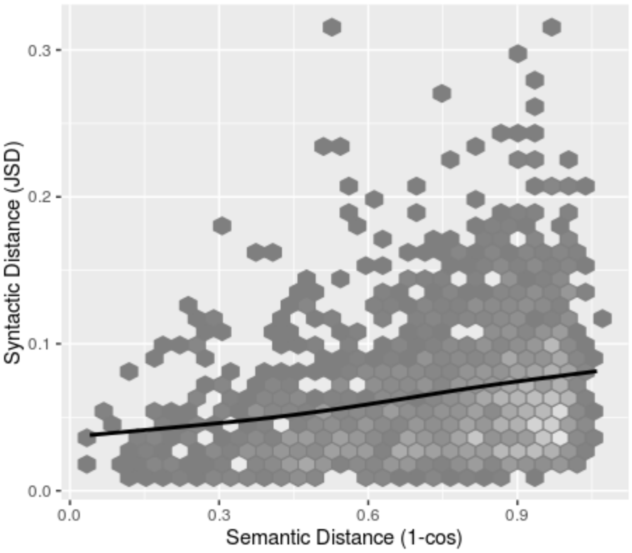


Figure 8: Relationship between syntactic and semantic distance measures

To avoid the semantic confound, I 'clean' the syntactic measure of its semantic content. I *residualize* the semantic measure out of the syntactic measure by performing a linear regression over the unique prime—target pairs in the SPP database. I predict JSD as a function of semantic similarity, then replace the original JSD estimates with the residuals of the model. In this way, I capture the information in JSD that is not attributable to semantics (for a recent defense of this method, see Hendrix, Bolger, & Baayen, 2017; cf. Wurm & Fisicaro, 2014).

iv. Further controls. A number of other factors are known to impact recognition latencies in the primed lexical decision paradigm. These fall into three categories: effects related to recognizing individual words, (other) effects based on the relationship between prime and target, and effects related to the nature of the task itself. From the first set, the most important predictor is the surface frequency of the target: more frequent words are recognized faster. I use the SUBTLEX-UK frequencies, which are based on movie subtitles and known to outperform estimates drawn from other corpora, including the BNC (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). I also include a measure of the density of the orthographic neighborhood of the target known as OLD20 (Yarkoni, Balota, & Yap, 2008). The more similar the spelling of the word to its closest neighbors, the faster it is recognized. Another important (if controversial) predictor is age of acquisition: the earlier a word is acquired in the lifespan, the faster it is recognized (e.g., Kuperman et al., 2012). Less important, but nevertheless known to exert an effect, is the orthographic length of the word: longer words take longer to recognize (New, Ferrand, Pallier, & Brysbaert, 2006). This effect is thought to be physical in nature. Lower acuity in the parafoveal region makes it more difficult to extract information from longer words (though see Veldre & Andrews, 2018, for evidence that semantic and syntactic information is recovered in sentence reading).

I include two predictors relating the prime and target besides the residualized syntactic measure. First, I include the LSA distance. As mentioned above, semantically similar primes are known to facilitate access to targets. In addition, I include the Levenshtein distance (LD;

Levenshtein, 1966; see van der Loo, 2014, for implementation in the *R* package *stringdist*) between prime and target. LD reflects the minimal number of changes (reversals, deletions, insertions) needed to transform one word into another. Orthographically similar prime—target pairs should result in slower recognition latencies on the assumption that orthographic overlap between prime and target increases competition among candidate word forms (Adelman, et al., 2014). In addition to these main effects, I include two-way interactions between the (factorized) interstimulus interval on the one hand, and LSA, LD, and residualized JSD on the other. In this way, I account for the possibility that priming effects will be reliably stronger at shorter offsets between prime and target.

Finally, I include the (log) sequential position of each trial in the overall experimental order of presentation. As participants move through the trials, some degree of fatigue should set in (each participant performed over 800 trials), producing generally longer RTs.

2. Results

I performed a linear mixed-effect regression predicting response latencies from the SPP primed lexical decision database as a function of the variables outlined above. In addition to fixed effects, I include random intercept adjustments for participants and prime—target pairs. I discard all latencies falling below 400 ms or 2 standard deviations above the mean (~1212 ms) as outliers (6.7% of all trials). In addition, I correct for a strong positive skew in the response times by taking the logarithm (as suggested by a Box-Cox power analysis; Box & Cox, 1964). Visual inspection of the model residuals with and without the corrections confirms the necessity of these steps.

All main effects for the control predictors besides OLD20 surfaced as significant at the α=.05 level, and in the expected direction. The model also uncovered a significant (*p*<.001) effect of the two-way interaction between LD and ISI: at 50 ms ISI, LD had a negative impact on response times (-2.5 ms per unit increase in LD), with no effect at 1050 ms. This result suggests that orthographic similarity between primes and targets indeed involves a (short-lived) competitive process, but I leave this question to future research. More importantly, the model revealed a significant interaction (*p*<.01) between ISI and LSA distance, this time in the expected direction: RTs increased by ~5 ms per .1 increase in cosine distance at short ISI. At long ISI, this effect was reduced to ~3 ms per .1 increase. The more semantically distant the prime—target pairs, the slower the target was recognized, with a sharper effect at short offsets.

Over and above the effects of the controls, the model returned a significant main effect (*p*<.001) of residualized JSD, as well as a marginal interaction with ISI (*p*=.07). I focus on the former: for every .1 increase in residualized syntactic distance, RTs increased by ~4 ± ~3 ms. Thus, the less related the prime and target in syntactic space, the longer it takes to recognize the target.

3. Discussion

The present study demonstrates a relatively strong effect of syntactic similarity on response times in a previously published database of primed visual lexical decision data. In fact, the effect was similar in strength to that of semantic similarity. To my knowledge, this study is the first to demonstrate that pre-activating a word's syntactic space affects access to that word in a *prima facie* non-syntactic comprehension task. The effect is perhaps all the more surprising, given that it was revealed for *nouns* – a word class largely thought to harbor the *least* amount of syntactic information (usually restricted to a few category specifications; e.g., Durán and Pillon, 2011). These findings have important implications for theories of language processing and representation.

Current models of “single-word” lexical priming (Neely, 1991) based on lexical decision evidence have not been designed to account for syntactic effects. Instead, they have focused on the role of semantic and orthographic relations between primes and targets. However, they may be instructive for interpreting the results. While semantic effects have been associated with facilitation, orthographic effects have been associated with inhibition. That is, priming a target with a semantic associate helps to reinforce the orthographic evidence for the target (Neely, 1991), while priming with an orthographic associate interferes with target identification (Adelman et al., 2014). The data I rely on here do not provide a non-primed baseline, meaning that we cannot be sure whether syntactic similarity is facilitative or dissimilarity is inhibitory. I leave this question to future research. However, the similar shapes of the syntactic and semantic effects suggest that syntax, like semantics, feeds back into lexical candidates prior to the lexicality judgment. Furthermore, it suggests that syntax, like semantics, is *obligatorily* accessed as soon as lexical forms become active. Crucially, the relationships between words and syntax become active even when (overt) syntactic structure is not built into the stimuli and not specifically required to complete the task. Recent psycholinguistic work on single-word production has echoed this point. For example, Lester and Moscoso del Prado Martín (2016) report chronometric findings suggestive of large-scale feedback from syntax to lexicon in a bare-noun picture-naming task. Other studies have found that syntactic category information is likewise obligatorily activated in non-syntactic production tasks (e.g., Durán and Pillon, 2011). The present study extends these findings from production to comprehension, from spoken language to written language, and from a simple to a primed paradigm. Hence, the converging evidence suggests that obligatory syntactic access, along with bi-directional feedback between syntax and lexicon, is a general, modality-independent property of language processing.

These data also speak to linguistic representation (Branigan & Pickering, 2017). In order for lexical priming to take place, some common connection must exist between the words and the representations underlying the measurement of their similarity. This notion is uncontroversially applied to the relationship between words and conceptual content in the semantic priming literature (e.g., Lam, Dijkstra, & Rueschemeyer, 2015). By extension, these results can be interpreted as reflecting a common set of syntactic structures to which each noun is individually connected. Moreover, the probabilistic nature of the measure suggests that connection weights – not just the set of shared syntactic types – are represented in the lexico-syntactic network, exactly as predicted by usage-based models of linguistic representation (Diessel, 2015). Importantly, these findings are *not* consistent with modular-syntactic models (e.g., Chomsky, 1995), which posit a strict divide between the generative syntactic mechanism and the memory store of lexical items. Adapting the old jingle, “you can take a noun out of syntax, but you can't take the syntax out of a noun.”

4. Future Directions

I used Latent Semantic Analysis as a proxy for semantic relation when 'cleaning' the syntactic measure of its semantic component. However, LSA has its limitations. It is based on similar distributions across the paragraphs of a large body of texts, regardless of the relative proximity of the words *within* those paragraphs. Hence, it may better capture broad, discursive-semantic similarity as opposed to the type of fine-grained, feature-driven semantic similarity which has also been demonstrated to impact lexical priming (Hutchison, 2003). For example, two words may tend to occur in the same paragraphs, but never in the same sentence, or in the same positions relative to other words within sentences. To reduce granularity, one could consider distributions of words relative to the other words that fall within a small window. This technique has been shown to produce quite reliable semantic representations (even considering only a one-word window to the left and right of the target can be quite effective; Bullinaria & Levy, 2007). Measures of this sort should provide a more stringent test of the syntactic (or, at least, the non-semantic) contribution of this measure.

Recall that the model revealed a marginal interaction between the measure and the temporal offset of the prime and target. The SPP contains only two such offsets: extremely fast and extremely slow. Therefore, one may find a more robust interaction at offsets intermediate to these extremes. Furthermore, by incrementally increasing the offset between 50 and 1050 ms, one could treat this variable not as a factor, but as a proper numerical variable (true to its nature; Feldman et al., 2015).   he

III. Effects of Prior Syntactic Distributions on Production

A. Introduction

The relationship between words and syntax – the set of configurations into which words may be organized – has been a fraught topic in linguistics. Early theoretical research assumed a strict divide between the two, both in terms of function and representation (Chomsky, 1957, 1995). A hallmark of these theories is that words are objects in memory while syntax is a combinatorial system designed to operate over these memory objects: Words present a set of affordances in the form of categories (e.g., part of speech, gender, and so on), and syntax uses matching algorithms to map the appropriate word to the appropriate position in the syntactic structure (usually some form of hierarchical tree). These theories predict that lexical processing can take place in the absence of syntax so long as the combinatorial system is not directly engaged (e.g., by the requirement to produce a syntactic frame). This prediction has received some, albeit limited, empricial support. For example, in a picture naming task, La Heij, Mark, Sander, & Willeboorsde (1998) find that distractors of the same grammatical gender as the target response impact performance, but only when the picture is named using a noun phrase (DETERMINER+ NOUN), for which the form of the response depends on that gender information. When the name is produced by itself, no effect was observed. However, others have argued that this gender-congruence effect is a reflex of lexical selection of the determiner, a sort of priming effect whereby preactivation of the syntactic information necessary to select the correct phonological form of the determiner facilitates production of the NP (e.g., Costa, Kovacic, Fedorenko, & Caramazza, 2003).

An opposed set of theories have posited that words and syntax are inextricably bound up in one another (e.g., Goldberg, 1995). These theories predict among other things, that lexical processing should never proceed in isolation, but should always engage the syntactic space attached to a given word. More recent work goes further to suggest that words and syntactic structures are (potentially) directly related within a single memory network, sometimes referred to as the *constructicon* (e.g., Fillmore, Lee-Goldman, & Rhodes, 2012). These relationships go far beyond simple category labels. For example, the word *cat* is not simply annotated for its status as *noun or countable*; it participates in a broad network of relations with morpho-syntactic constructions at multiple levels of abstraction (e.g., stem + -s plural inflection, subject of transitive construction, and so on). Importantly, these relationships are fundamentally probabilistic. The strength of any single association depends on a complex set of cues derived from one's experience with the related elements within the network (e.g., Bates & MacWhinney, 1989; Diessel, 2015). I shall henceforth use the language of probabilistic distributions and interactive activation networks to refer to the same underlying phenomenon. A growing number of psycholinguistic studies support these types of models. For example, contra La Heij and colleagues, several studies using different experimental paradigms (blocking, picture-word interference) have reported significant effects of category (non)congruence in non-syntactic tasks (Cubelli, Lotto, Paolieri, Girelli, & Job,, 2005; de Simone & Collina, 2015; Gregory, Varley, & Herbert, 2012). In comprehension, the probabilistic associations between words and syntactic structures have been shown to affect recognition of isolated words both in behavior (Baayen, Milin, Filipović-Đurđević, Hendrix, & Marelli, 2011; Linzen, Marantz, & Pylkkänen, 2013) and electrophysiology (Linzen et al., 2013). This work has so far looked at distributions within single constructions (nouns in the prepositional phrase; Baayen et al., 2011; Hendrix, Bolger, & Baayen, 2017) or if across constructions, only those for which the word is the syntactic head (verbs in argument constructions; ;Linzen et al., 2013). Only one study to my knowledge has used a production task (Hendrix et al., 2017). While they did find an effect, their task involved a partially syntactic component (the pictures to be named were preceded by an orthographic phrasal frame, e.g., *in the*). Therefore, it is still not clear whether production of nouns is sensitive to prior syntactic distributions.

The present study extends this body of work by answering several questions. First, I employ a cross-constructional measure of syntactic distributions for nouns (similar to the measures of Linzen et al., 2013, for verbs). Next, I distinguish two types of syntactic relationships, each with two levels: hierarchy, with a contrast between head and modifier functions, and word order, with leftward and rightward directions (measured from the noun). Third, I follow Linzen and colleagues by contrasting the information carried by the syntactic distribution to the prototypicality of that distribution. In so doing, I improve on the measures that Linzen and colleagues use for both of these kinds of information. Fourth, I compare production of the noun in isolation with its production in a predictable syntactic frame (*the* + NOUN). In this way, I test whether the syntactic connectivity – its resonance within the system – is powerful enough to impact production of the noun in a particular context. If syntactic information impacts the production of (truly) isolated words, then we have evidence that lexico-syntactic relationships are obligatorily accessed during lexical access. Such findings would fit with the probabilistic network account of Diessel (2015) and other usage-based accounts (e.g., Goldberg, 2006; Bybee, 2010). They would present a challenge for models of syntax that do not allow direct probabilistic relationships between words and syntactic structures *of all kinds* (including structures for which the word is not a functional head; e.g., Bresnan, 2001; Chomsky, 1995; Kay, 2013).

In what follows, I outline the evidence which links syntax to word processing and production. I then introduce two new approaches to measuring syntactic diversity and prototypicality, along with a set of 18 possible implementations of these measures. These measures are then correlated with response times in two picture naming experiments. Results are discussed in the context of the constructicon and how experience interacts with task demands to influence naming latencies.

B. Syntax and word production

The strongest evidence for obligatory syntactic activation in production comes from bare-noun object naming. In this paradigm, participants are presented with images of objects and asked to say their names aloud. The measures of interest are what name is produced and how long it takes to produce. Ostensibly, the task is non-syntactic in that the participants are not required to produce any syntactic structure. Empirically, La Heij et al. (1998) report that syntactic effects do not surface in bare-noun naming (e.g., *banjo!*), though they do surface in noun-phrase naming (e.g., *the banjo!*). La Heij and colleagues demonstrate this difference for Dutch nouns and noun phrases. When images are presented with distractor words from the same gender as the image name, participants produce the name of the image faster than when the distractors come from a different gender. Crucially, this effect only holds when the participants must also produce the gender-marked determiner. They interpret this finding as evidence for the selective activation of syntactic information under circumstances when that information is needed to complete the task. However, the syntactic interpretation of this result has been challenged. Schiller and Caramazza (2003) provide evidence for an alternative account which explains the gender effect through lexical selection of the determiner – that is, as a lexical effect. In addition, more recent research has reported effects from a number of syntactic categories on production latencies in bare-noun production. For example, Duràn and Pillon (2011), using a blocked priming task, found faster response times for trial blocks containing only nouns or only verbs than for blocks containing nouns and verbs (a *word class congruence* effect). Unlike gender congruence, word class congruence cannot be attributed to additional lexical search functions. Thus, it appears (1) that lexical access may involve obligatory activation of syntactic information, even in the absence of syntactic encoding, and (2) bare-noun naming can tap into this relationship (see also Gregory et al., 2012; de Simone & Collina, 2015).

Much of the prior research has focused on categorial information. In connectionist terms, these studies attempt to preactivate an abstract syntactic category node which shares links among words belonging to that class. Preactivation can facilitate (priming; Gregory et al., 2012) or inhibit (picture-word interference; de Simone & Collina, 2015) access to the target, depending on the task and design. Categorial constraints of this kind figure prominently in models of sentence production (Dell, Oppenheim, & Kittredge, 2008). However, these category labels are actually quite complex generalizations over both morpho-syntactic and morpho-phonological distributions. For example, the syntactic and inflectional potential of English nouns is not purely predictable by their belonging to the category NOUN. This is not a simple matter of exceptions to the rule. Instead, the category label belies a much richer network of semi-productive subgroups that stand at the intersection of a number of other features: the mass/count contrast, (historically derived) phonological patterns (i.e., patterns of suppletion), semantics (i.e., semantically constrained syntactic distributions; **the pitcher gave the water to the cup*), lexical prosodic effects (i.e., patterns of stress shift; *PERmit*, noun, vs. *perMIT*, verb; cf. *emBRACE,* noun, vs. *emBRACE*, verb), and so on). As such, it is not clear that they actually hold as independent symbols in the lexical network (Schiller & Caramazza, 2003; Milin, Filipović-Đurđević, & Moscoso del Prado Martín, 2009; Baayen et al., 2011). For example, noun lemmas could connect directly to the set of inflectional and syntactic representations with which they combine. When words share similar distributions within this space, they should be treated similarly with respect to those distributions – a 'ghost-category' effect. From this perspective, the different class congruence effects reduce to similarity of use.

A complementary line of research from comprehension has looked at morpho-syntactic distributions and their effect on word recognition. This work ultimately descends from studies of inflectional morphology (i.e., the syntactic branch of morphological paradigms). Moscoso del Prado Martín et al. (2004) show for Serbian that the more uncertain the inflection of a given stem, the faster it is recognized. Such a finding supports the probabilistic network account of Diessel (2015) and others. Activation spreads between a central lemma node and its set of inflectional variants. Where this connectivity is evenly distributed across the available forms, the overall lemma receives more efficient support from its inflected variants. This support rapidly boosts lexical activation across the network, leading to faster recognition in visual lexical decision. Morphological inflection presents an interesting case because it translates syntactic relations into formal variants. That is, inflected word forms wear their syntactic functions on their sleeves. For example, Russian собак-а [sobak-a] 'dog-NOM' applies primarily to sentential subjects whereas собáк-и [sobak-i] 'dog-GEN' would be used to indicate possession**.** Simply knowing the surface form allows us to infer something about its *in situ* syntactic role. By extension, nouns with diverse inflectional distributions are bound to have more diverse syntactic distributions. Thus, for inflectionally rich languages, inflectional diversity partially reflects syntactic diversity. However, for languages with limited morphology such as English, functions that are performed in other languages via inflection are performed almost exclusively through syntax. For example, the morphological *cases* of highly inflecting languages correspond largely to phrasal structures in English. Despite the difference in locus, the 'syntactic inflections' of English behave similarly to the morphological inflections of other languages: nouns that have more diverse distributions across preposition types are recognized faster (Baayen et al., 2011; Lester & Moscoso del Prado Martín, 2015) and produced faster (Hendrix, et al., 2017). In the terminology introduced above, these studies measure the lexicalized modifier diversity for nouns in a single dependency relation (*pobj*, for objects of prepositions). The diversity is lexicalized in that the frequency distribution is defined over prepositions as opposed to abstract syntactic categories. Therefore, these studies effectively measure lexical diversity within syntactic constructions.

An open question concerns the role of prior syntactic distributions in bare-noun picture naming. The only study to my knowledge to use a production paradigm did not use an isolated-production task. Hendrix et al. (2017) employ comprehension-to-production priming via the presentation of a syntactic context (preposition + determiner, e.g., *in the*) prior to the picture. For this reason, their results are ambiguous: should they be attributed to lexical features of the picture name or to aspects of the syntactic processes linking the name to the primed context? In this respect, Hendrix and colleagues find that the ERP signals reflect the distributional measures in a manner resembling the effects of word frequency. Word frequency is typically interpreted as a lexical effect, which suggests that the diversity effect is also anchored – at least in part – to the lexical representation. Moreover, the presence of the prime could only interfere with this effect by conditioning the likelihood of the noun. That the effect remains despite the prime, and that it shares its electrophysiological profile with the frequency measure suggests that the latter is most likely lexical in nature. What remains to be seen is (a) whether the syntactic information still affects naming in the absence of a syntactic context, (b) whether any such affect will surface in behavior (i.e. response times), and (c) whether cross-constructional measures of syntactic diversity and prototypicality will likewise impact naming. In the next section, I introduce a method for estimating cross-constructional syntactic information for nouns. Points (a) and (b) are addressed in the Experiment 1 and 2, reported below.

C. Measuring Syntactic Diversity

I operationalize syntactic relations using a dependency grammar formalism (Choi & Palmer, 2012). In this approach, syntactic relations apply to pairs of words – labelled the “head” and the “modifier” – with an additional label describing the precise nature of their relation. I refer to the triplet of dependency relation, head, and modifier as a *bundle* (also known as a *construction*). For the set of bundles involving a noun, I am interested in the *syntactic information* carried by each noun type. Prior work has investigated how the syntactic information carried by nouns *within* syntactic categories influences lexical production (Hendrix et al., 2017). This work has shown that each category, for example, the prepositional phrase construction, presents its own “paradigm” whose “cells” represent the possible head lemmas, e.g., prepositions. Here I apply the same reasoning, but at a higher level of abstraction. Instead of lemmas, the cells reflect syntactic dependencies. That is, I look at the distribution of nouns *across* rather than *within* syntactic structures. This approach is therefore similar to that employed by Roland, Dick, and Elman (2007) and Linzen et al. (2013), who measured the distributions of verbs across the different argument structures with which they combine. However, the dependency formalism allows us to take a finer-grained perspective on syntactic relations. Whereas the structures studied by Linzen and colleagues are headed by the verb, dependencies allow easy modeling of the hierarchical status of words as either head or modifier. Moreover, we can compare the direction of the relation (leftward or rightward), as well as any combination of hierarchical status and direction. Counting every possible combination, we arrive at nine syntactic distributions:

— Overall relations (ignoring hierarchy and direction)

— As-head relations (ignoring direction)

— As-modifier relations (ignoring direction)

— Rightward relations (ignoring hierarchy)

— Rightward as-head relations

— Rightward as-modifier relations

— Leftward relations (ignoring hierarchy)

— Leftward as-head relations

— Leftward as-modifier relations

A second goal of this study is to explore the relative importance of all nine dimensions of syntactic information described above. More specifically, we want to determine whether word order, hierarchical status, or some combination of the two are important for understanding the effects of syntactic diversity on noun production in picture naming.

Syntactic relations are partially, or in some cases wholly, identifiable based on the lexical forms involved. In other words, dependency relations and the lexical identity of heads and modifiers should be redundant to some extent. For example, if we see that the noun *ship* is paired with the word *the,* we immediately know that the relationship is *det*, for determiner modification (as well as that *ship* is the head and that *the* precedes *ship*). To ensure that we are dealing with truly syntactic information (and not information derivable from lexical context), we must "clean" the syntactic measures of information carried by words alone. To accomplish this, I take advantage of an information-theoretic measure called conditional entropy. The conditional entropy of a distribution S given a distribution L is defined as follows: $H(S \mid L) = H(S , L) - H(L)$. That is, the conditional entropy is equal to the joint entropy of S and L minus the entropy of L. $H(S , L)$ reduces to the negative sum of the weighted joint probability $p(s_i , l_j)$ where $s \in S$ and $l \in L$, i ranges over S and j ranges over L. If S is defined as the set of syntactic dependencies, and L is defined as the set of lexical forms that accompany these dependencies, then $H(S , L)$ reflects the information carried by the lexical and syntactic tiers together. $H(L)$ reflects the information carried by the lexical tier independently. By subtracting the lexical entropy from the joint entropy of syntax and lexicon, we arrive at the information carried by the abstract syntactic dependencies independent of the lexical information. If the information carried by syntactic dependencies is entirely redundant given the lexical content of the bundle, then $H(S , L) = H(L)$ and $H(S \mid L) = 0$. In this case, the syntactic information of the word could be read entirely off of the lexical context in which it is embedded. On the other hand, if $H(S \mid L)$ is greater than zero, then some other layer of stochastic generalization must be at work, namely, a syntactic layer. Henceforth, I refer to $H(S \mid L)$ as *syntactic diversity.*

Beyond the syntactic information carried by individual words, there is further information carried by word class. Word classes can be defined at many levels of granularity. Here, I examine part-of-speech. The underlying intuition is that words are not isolated within the linguistic system. Instead, they tend to be nested within groups of functionally related words. Earlier work on morphology shows that, in comprehension, these paradigms mediate processing efficiency. For example, Milin et al. (2009) find that Serbian nouns are recognized faster to the extent that they are distributed across their inflectional exponents in ways similar to other words of their class. Similar effects have been observed at the level of syntax for phrasal classes. Hendrix, et al. (2017) show that English nouns are produced faster when they are distributed across prepositional phrases in ways consistent with the overall pattern for prepositional phrases. In other words, any given *exemplar* is processed more efficiently when it resembles the *prototype* of its phrasal class, even in isolation. The precise cognitive mechanisms underlying this effect remain unclear. By one account, the effect arises as a function of discrimination learning: linking noun forms to meaning is more challenging when nouns differ from the rest of the linguistic system in how they are coupled with prepositions, or any other flexible aspects of their lexical context for that matter (i.e., the learner will be led by system-wide preferences to make worse guesses about the meaning representations for such nouns; Baayen et al., 2011). Another possibility is that lexical representations are highly distributed within a feature space that includes syntactic features, as well as semantic features, orthographic/phonological features, and so on. Prototypes could emerge as stable patterns of activation across words within this space. Such prototypes would have a greater impact on baseline activation within the system (as they represent larger-scale aggregates of experience than any of the individual words). In this case, the prototype effects might signal a difficulty in transitioning from global activation states to the target state (e.g., Plaut & Booth, 2000). Some have challenged the latter argument, saying that it would require computationally intractable storage of exemplars (e.g., separate representations for every prepositional phrase that one has ever encountered; Baayen et al., 2011; Baayen, Hendrix, & Ramscar, 2013). However, this “combinatorial explosion” could possibly be avoided by re-framing complex exemplars as shared (i.e., simultaneous or contingent) activation within the distributed language network. Such lexico-syntactic networks stand at the core of functional-linguistic theory, in particular the constructionist approaches (Bybee, 2010; Diessel, 2015; Goldberg, 1995; Langacker, 1987). In either case, these prototype effects are expected to surface for linguistic relationships only insofar as they are functionally relevant, whether to learning or to on-line transitions within a state space.

The studies above operationalize class-wise similarity using the Kullback-Leibler divergence (KLD; also known as the *relative entropy*) of the probability distribution of a given word from the summed distributions of all words. This is an information theoretical measure of the degree to which two distributions differ. The distance from prototype P to exemplar E is formalized as $KLD(P \parallel E) = \Sigma\, E \log E/P$. The measure is asymmetric, meaning that the divergence from P to E is not necessarily equal to the divergence from E to P. As such, one must decide *a priori* in which direction to take the divergence. This property is undesirable in the present context given that the cognitive mechanism underlying the effect is still not well understood. Lin (1991) proposes a symmetric alternative, commonly referred to as the Jensen-Shannon divergence (JSD). JSD is calculated similarly to KLD, but with one addition. Instead of taking the distance between two distributions, JSD first requires that we compute the midpoint M between E and P. Then, the KLD is computed from E to M and from P to M. JSD is defined as the mean of these two divergences. As with the entropies discussed above, the probability estimates entered into the JSD generally underestimate the true probabilities. I correct the probabilities in E and P using the James-Stein shrinkage estimator, which is optimal for cases when the number of columns is known *a priori* (Hausser & Strimmer, 2009). In this case, the number of columns is equal to the length of the set of dependencies observed at least once for all nouns in a given condition (e.g., the nine dependency conditions I describe above). I refer to JSD($E \parallel P$) as *syntactic atypicality* (rather than *prototypicality*) to capture the fact that the measure refers to distance from the prototype.

1. Computing the estimates

I derive probability estimates from the OANC (Reppen, Ide, & Suderman, 2005). I first parsed the corpus with the *spaCy* dependency parser (Honnibal & Johnson, 2015). I then identified all tokens tagged as nouns, retrieved the lemmas for those tokens, and computed a frequency matrix for all the noun lemmas. This process leaves us with a total of 10,684 noun lemma types. To simplify, I consider only monomorphemic lemmas (based on the CELEX labels as they appear in British Lexicon Project database; Keuleers, Lacey, Rastle, & Brysbaert, 2012). I thus sidestep the complexities of compounds and derivational families (Schreuder & Baayen, 1997). Still, some of the remaining lemmas are ambiuous between word classes. For example, *orange* may be a noun, as in *Orange is my favorite color,* but may also be used as an adjective, as in *the orange pumpkin*. Such cross-category relationships may influence the syntactic distributions of the noun forms, or may interfere with online processing by engaging multiple subspaces of the syntactic network in parallel (e.g., the subspaces associated with the noun and adjective uses of *orange*). To control for possible cross-categorical contamination, I consider only those nouns which appear predominantly as nouns (based on the annotation in the database of concreteness norms published by Brysbaert, Warriner, and Kuperman, 2014). These cuts brought the sample down to 3,124 distinct noun lemmas.

I constructed three frequency matrices around these lemmas. In the first matrix, the columns reflected all pairs of dependencies and related words observed for nouns in the corpus (e.g., *det* + *the*, *det* + *this*, *dobj* + *eat*, and so on). In the second, the columns contained just the related words (e.g., *the*, *this*, *eat*, and so on). Finally, in the third matrix, the columns contained just the dependencies (*e.g., det*, *dobj*, and so on). I take the entropy of each noun in the first and second formulation to arrive at $H(D, L)$ and $H(L)$, respectively. Each of these entropies was corrected for underestimation (see Miller, 1955) following the technique introduced in Chao, Wang, & Jost (2013). This technique has been shown to perform well when applied to the distributional profiles commonly found in language (e.g., the Zipf-Mandelbrot distribution; see Moscoso del Prado Martín, 2016). I then subtract the latter from the former to produce conditional entropies $H(D \mid L)$ for all of the nouns. As the goal is to investigate the role (if any) of delexicalized syntax, we must deal with tokens for which we observe zero de-lexicalized syntactic information. I first inspected the overall distributions (i.e., ignoring hierarchy and direction). Not so surprisingly, nearly half of the nouns in the sample had syntactic distributions that were entirely predictable from their co-lexical distributions (n = 4908). Under most circumstances, the relationships between nouns and other words are clearly defined (e.g., *the* can only instantiate the *det* dependency). Moreover, the semantic properties of nouns may dictate their distribution (e.g., knowing that the verb is *eat* almost certainly precludes any inanimate noun from being attached as subject[\4]). An examination of the frequencies of these lemmas revealed that the vast majority (80%) occurred less than 20 times, with 60% occurring less than 10 times. The probability distributions of these tokens are highly likely to be unrepresentative of the true distributions (even despite the corrections). Erring on the conservative side, I restrict further analysis to lemmas with frequency greater than 100 in the OANC (~7 pMw; the maximum frequency observed for any lemma with $H(D \mid L) = 0$ was 80). These further cuts left us with 1,563 distinct noun lemmas.

The third matrix type is needed to estimate the degree to which a noun is close to the “prototypical noun”. Unlike in the case of standard entropies, removing lexical information

[4] We stress that the assumed stochastic nature of language renders absolute certainty – within the limits of grammatical convention– an impossibility.

from divergences is not straightforward. There are many possible ways to approach this issue. As a first approximation, I ignore information carried by the other word in the dependency by computing the frequency distributions over the dependency types alone. Simplifying in this way allows us to compute straightforward divergences. I define the syntactic prototype of nouns as the summed distribution of all nouns in the sample (i.e., the vector created by taking the sums of all columns in the frequency matrix). I then take the JSD between each noun and the prototype (i.e., between each row and the summed vector). I consider this an acceptable compromise given that only tokens with reliable and non-zero estimates of $H(D \mid L)$ for their overall distribution are considered.

2. Decorrelating the measures

The raw estimates of diversity highly intercorrelated (k>30). This problem, known as multicollinearity, can damage the reliability of model estimates. To remedy this issue, I subject the nine predictors to *independent component analysis* (ICA) using the *fastICA* algorithm (Hyvärinen & Oja, 2000; Marchini, Heaton, & Ripley, 2013). The fastICA algorithm reconstructs a set of maximally statistically independent components from the observed vales. First, the matrix of observed values per word across the nine dimensions is centered and decorrelated to produce a matrix of n non-correlated components (i.e., they are subjected to a principal component analysis, PCA, of n components). The number of components n is determined *a priori*, and must fall within the range $\{1,k\}$ where k is the number of original dimensions. I did not have any principled reason to select n on purely *a priori* grounds. I therefore apply an empirical heuristic. I define n as the number of PCA components needed to explain at least 95% of the variance in the sample. After this initial decorrelation, the source signals are estimated. To accomplish this, the fastICA algorithm rotates the results of PCA to find orientations for which the resulting distribution diverges most from a normal distribution. These rotations are constrained to be orthogonal (from the observed matrix and from each other) to ensure that the resulting components are maximally uncorrelated. The resulting rotation matrix contains the coefficients needed to project the original predictors into new space. These coefficients therefore reflect the strength of the association between each original variable and the derived component.

I run two separate ICAs, one among the nine syntactic diversity measures, the other among the nine syntactic atypicality measures. Initial PCAs revealed that the diversity space and the atypicality space could be reasonably well expressed in five and four components, respectively. I followed this heuristic and produced five components for diversity and four for atypicality. The projected values of the nouns in each of the source spaces were recorded for each noun. Henceforth, I refer to these source spaces as *diversity* and *atypicality components*, followed by a unique number (e.g., *diversity component 1*).

D. Experiment 1: Bare-noun picture naming

1. Stimuli and design

The object images were taken from the set of 520 black-and-white line drawings of common objects that were used in the International Picture Naming Project (IPNP) research (Bates et al., 2003). Each participant saw 200 of the original 520 images. These 200 images were randomly selected at the onset of each experimental session, meaning that each participant saw a unique set of images. The 200 images were randomly divided into four sets of 50. Order of presentation within these groups was also randomized. All images and text were presented in black on a white background. The images were normalized to 300 X 300 pixels.

2. Participants

46 undergraduate students from a public university on the west coast of the United States were recruited to participate (N(female) = 35; mean age = 20.91), all of whom were native speakers of English with normal or corrected-to-normal vision. All participants were treated in accordance with the American Psychological Association guidelines for ethical human research.

3. Procedure

I follow the same general procedure as described in Bates et al. (2003). The experiment was carried out in a dimly lit, sound-attenuated room. All experimental materials were presented via the experimental software OpenSesame v. 3.1.2 (Mathôt, Schreij, & Theeuwes, 2012) on a 17-in LCD display with 1366 X 768 screen resolution. Participants were seated approximately 50 cm from the display. They were provided with written instructions which stated that they would be shown a series of images, and that their task was to say the name of each image aloud. They were instructed to say the name in isolation (“Banjo!”) as quickly and accurately as possible, and to avoid producing hesitations or fillers prior to saying the word. Finally, they were informed that they had a maximum of three seconds to name the image before the next trial would begin, and that they should remain silent if they could not find an appropriate name for the image before timeout. In the next phase, participants were trained on a set of three images taken from the database published by Bonin and colleagues (Bonin, Peereman, Malardier, Méot, & Chalard, 2003). These images were the same for all participants and were selected so as not to overlap with images from the IPNP set. During the training and critical trials, participants saw a fixation cross for 250 ms, followed by a white screen for 500 ms, then the image until the participant had named it, or for a maximum of three seconds. Stimuli were presented in four 50-image blocks with opportunities to rest after each of the first three blocks (to minimize fatigue effects). Responses were recorded with a Sony ECM-909 stereo microphone set to 90-degree spread. Responses were transcribed and response times were coded by hand using the audio-editing software Audacity[5].

4. Control variables

Many factors have been observed to correlate with word production latencies in picture naming. To control for these effects, I annotate the picture stimuli with the following information:

- *word frequency*
- *length in syllables*
- *subjective age of acquisition*
- *inflectional entropy*
- *diversity of names offered for an image*
- *shared name*s *across images*

[5] http://audacityteam.org

—	*objective image complexity*

The effects of the control variables are well established in the picture-naming literature. *Frequency* is known to have a facilitatory effect on production latencies (Oldfield and Wingfield, 1967), though the ultimate source of this effect is contentious (e.g., Almeida, Knobel, Finkbeiner, & Caramazza, 2007; cf. Bates et al., 2003). Frequencies were estimated from the SUBTLEX-US corpus, which contains frequencies for approximately 74,000 English words taken from a 51-million word sample of American English subtitles (Brysbaert & New, 2009). Length in syllables has a generally inhibitory effect on response times: longer words take longer to initiate, presumably due to the increased load on phonological and articulatory planning (e.g., Bates et al., 2003). Here I take syllabic lengths from the CELEX database (Baayen, Piepenrock, & Gulikers, 1995). *Inflectional entropy* (Moscoso del Prado Martín, Kostić, & Baayen, 2004) refers to the distribution of a word across its phonologically distinct inflectional realizations. For English nouns, this amounts to the relative probabilities of its occurring in either the singular or plural form (genitive clitic *'s* is not bound to the noun stem, and so is not considered). The inflectional entropy has been shown to have a somewhat weak effect on picture naming that depends on a number of factors (nouns: Baayen, Levelt, Schreuder, & Ernestus, 2008; verbs: Tabak, Schreuder, & Baayen, 2010). Generally, an inhibitory effect has been observed: more even splits between singular and plural realizations lead to longer naming latencies. To estimate these distributions, I tagged the 15-million-word Open American National Corpus (OANC; Reppen, Ide, & Suderman, 2005) for part-of-speech using the *spaCy* tagger (https://spacy.io) as implemented in the *Python* programming language. For each word, I took the entropy of its relative distribution across singular and plural realizations. I further corrected the entropies using the method introduced by Chao, Wang, & Jost (2013). *Diversity of names* and *shared names for images* pertain to codeability. Pictures that elicit many different names are less codeable; producing any name for these images takes longer as one must decide between many alternatives. Pictures that receive labels general enough to match names offered for other images are more difficult to differentiate with a basic-level name, which may lead speakers to seek hypernymic alternatives after a failed search at the basic level. This two-stage process results in later speech onset times. Shared names tend to be formally simpler and of higher frequency (Bates et al., 2003), leading to shorter response times for images that share names with other images. Finally, *visual complexity* is expected to be inhibitory: pictures with more pixels require heavier visual/conceptual processing before an appropriate name can be identified, leading to longer naming latencies.

5. Results

I remove all responses that contained hesitations, coughs, multiple naming attempts, or for which no answer was provided (n = 572; 6% of overall observations). I only consider responses that also appear in my database of noun entropies (*n* = 272 unique lemmas). RTs that fell beyond two standard deviations from the mean in either direction were removed (*n* = 146; 4% of observations). Visual inspection of the pairwise correlations between the full array of independent variables revealed a high degree of correlation among the variables, with an unacceptable collinearity index (condition number $\kappa > 41$). Pairwise scatterplots suggested that this high degree of intercorrelation is due to correlations between (log) word frequency, age of acquisition, and the diversity and atypicality measures. I therefore residualize these variables out of the frequency measure for unique lemmas by performing a linear regression with the lexical variables as predictors and (log) frequency as dependent variable. I then take the residuals in place of the original frequency measure and annotate each of the observed responses with its residual from model. A second check for collinearity with residualized frequency in place of the raw log frequency showed an acceptable degree of intercorrelation ($\kappa < 20$).

The remaining 3,492 RTs exhibited heavy rightward skew, which can affect consistency of model performance across the range of the response variable. A Box-Cox (Box & Cox, 1964) analysis suggested an inverse-square transform was most appropriate to approximate normality. Model comparison validated this transformation of the data: residuals from untransformed-RT models were non-normally distributed (as expected), with a strong tendency to overestimate RTs in the middle range. Residuals from the transformed-RT models were normally distributed.

I fitted a generalized additive mixed model (GAMM)[6] predicting transformed response times. Smoother terms were fit for (log) trial number, (log) previous RT, (log) objective visual complexity, naming diversity, residualized frequency, age of acquisition, inflectional entropy, the five diversity components, and the four atypicality components. Parametric terms were fit for number of syllables and shared name. The former was included as a

[6] All models were computed both as GAMMs and as linear mixed effect models (LMM). Results from the two analyses converged for all critical predictors. GAMMs are reported because several of the control predictors showed strongly non-linear effects.

parametric term because of the small number of possible values (1, 2, or 3 syllables). In addition, random intercepts were fit for participant, image, and response lemma.

Significant predictors from the model are given in Table 3, along with the effective degrees of freedom (*eDF*), the estimated residual degrees of freedom (*refDF*), *F* values and *p*-values.

Table 3: Significant predictors of RTs in bare-noun picture naming

Smooth terms	***eDF***	***refDF***	***F* value**	***p* value**
trial number (log)	3.24	4.02	10.54	<.001
previous RT (log ms)	1.00	1.00	9.67	.002
age of acquisition	1.00	1.00	20.41	<.001
name diversity	3.70	4.06	44.90	<.001
visual complexity (log file size)	1.02	1.03	6.87	.008
inflectional entropy	1.13	1.18	8.82	.002
prototypicality component 4	1.00	1.00	4.58	.03
Random effects				
image	146.62	296.00	1.63	<.001
subject	42.15	45.00	15.49	<.001

Control predictors. Neither of the parametric terms surfaced as significant. Surprisingly, neither did residualized frequency. As mentioned above, the word frequency effect is one of the strongest and most reliable predictors of behavior in lexical processing and production tasks. Importantly, however, word frequency is bound up with other variables, as indicated by the high degree of collinearity observed in the present sample.

Residualizing the other variables out of word frequency produces a measure closer to pure repetition. Baayen (2011) performed a more extensive residualization and found that reducing frequency to pure repeition substantially decreased its explanatory power for lexical decision. Therefore, the lack of an effect here could be due to the relatively weak influence of pure repetition being swallowed by the much stronger predictors.

Several controls related to the experimental procedure surfaced as significant. First, trial number was negatively correlated with RT. Over the course of the experiment, RTs gradually slowed, eventually leveling out in the later trials with a slight upturn for the last trials. Because the variable was log transformed, this pattern suggests a mostly linear effect for the raw trial numbers. This could be due to fatigue or to interference from a higher number of recently experienced exemplars from the prior trials. The RT of the previous trial also exerted a strong effect. This effect was log-linear, suggesting a diminishing influence on target RTs as previous RTs increased. The diversity of names offered for the image across participants also correlated negatively with response times, increasing steadily, but leveling out in the upper ranges. This effect suggests that some images activate more possible names than others, making it more difficult to settle on the target form. Visual complexity was likewise negatively correlated with RT. The higher the number of kb needed to encode the image, the longer participants took to produce the name. This effect presumably arises during the visual decomposition and conceptual-semantic mapping stages.   Significant random effects show that subjects and images varied significantly within their respective groups, but lemmas did not.

Of the lexical variables, only age of acquisition and inflectional entropy were significant.

Age of acquisition showed the expected negative correlation: words that are learned later in life take longer to initiate. Counter to expectations, inflectional entropy showed facilitation: the more balanced the split of occurrences of a lemma across its singular and plural forms, the faster the naming RT. There is some precedence for this effect. Tabak, Schreuder, and Baayen (2010) report facilitation for the inflectional entropy of verbs, but only for non-targeted responses to the image. In this study, there was no clear target for the images. Tabak and colleagues suggest that this facilitation could arise because words with complex inflectional paradigms (i.e., higher inflectional entropies) make for strong competitors (i.e., a "gang effect") during lexical selection, sometimes overcoming the target. To test for this effect, I determined the lemma of the dominant response for each image across participants and annotated each observation for whether the response lemma matched (*match*) or did not match (*mismatch*) the dominant response lemma. I refit the model with separate factor smooths on inflectional entropy for the matched and mismatched responses. Based on Tabak et al. (2010), I expect a stronger effect of inflectional entropy for the mismatched responses. This is precisely what I found (p <.01, adjusted for multiple comparisons using the False Discovery Rate; FDR; Benjamini and Yekutieli, 2001). The factor smooths are plotted in Figure 9.

Figure 9 reveals that the matched tokens showed virtually no effect of inflectional entropy, while the mismatched tokens showed facilitation. Mismatched responses were slower than matched responses for low-complexity paradigms but similar for high-complexity paradigms). These findings thus replicate those of Tabak and colleagues regarding mismatched responses. However, I fail to replicate the inhibitory effect for

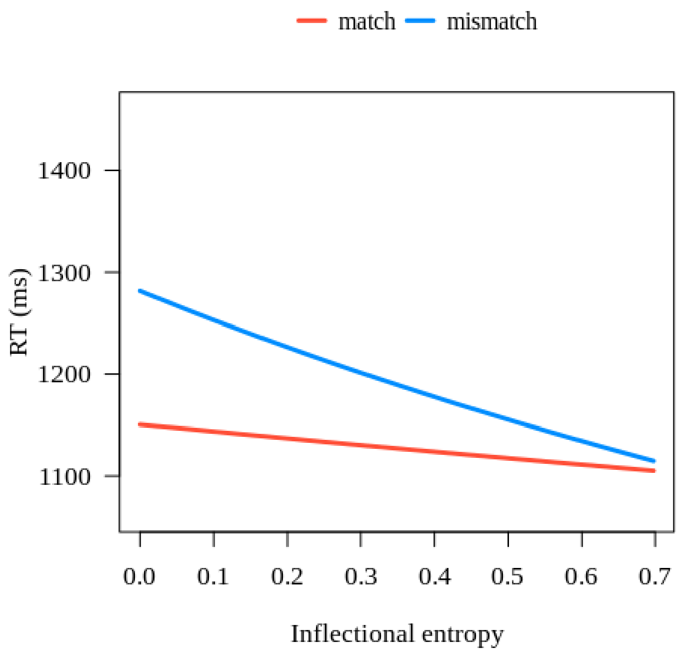


Figure 9: Different effects of inflectional entropy for responses that match (matched; red) and do not match (mismatched; blue) the dominant responses per image.

matched tokens. This is perhaps due to the fact that inflectional entropy is generally weak in production (Baayen, Feldman, & Schreuder, 2006).

Critical predictors. An initial model (not reported above) revealed significant effects of two of the atypicality components, 2 (*p* <.01) and 4 (*p* = .02). Closer inspection of these effects revealed curvature that was driven by a few observations at extreme values of the predictors. To guard against artifacts due to under-sampling at the extremes, I removed the outlying observations and refit the model (this is the model reported in Table 3). The cut proved necessary: after removing these values, the effect of atypicality component 2 disappeared entirely (*p* = .72). However, component 4 retained its significance (*p* = .03). This effect is plotted in the bottom panel of Figure 2. The component loadings are given in the top panel.

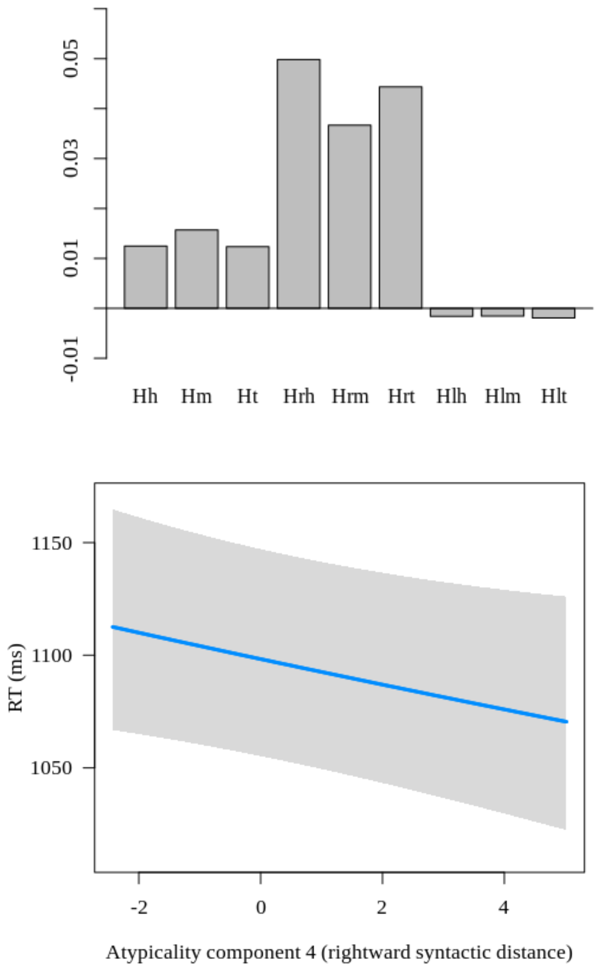


Figure 10: Effect of atypicality component 4. Top panel: Component loadings for atypicality component 4. Loadings indicate that the component reflects distance in the rightward syntactic space. Bottom panel: Effect of atypicality component 4 on RTs. The solid line plots the regression curve. The shaded area represents the 95% confidence interval.

The top panel of Figure 11 shows that this variable loads most heavily on the rightward JSD estimates to the exclusion of the leftward JSD estimates. The bottom panel shows that nouns with typical rightward syntactic distributions are processed more slowly than nouns with atypical distributions. As expected based on prior work, the effect is somewhat weak. Unexpectedly, the effect was facilitatory: words with more atypical rightward distributions are produced faster than words with idiosyncratic distributions. Based on the interaction I observed for inflectional entropy, I explore the possibility that the syntactic atypicality effect differs between matched and mismatched tokens. To this end, I refit the model, this time with separate factor smooths for component 4 based on the matched and mismatched responses. Indeed, a marginal difference emerges (*p* = 0.05). This relationship is plotted in Figure 11.

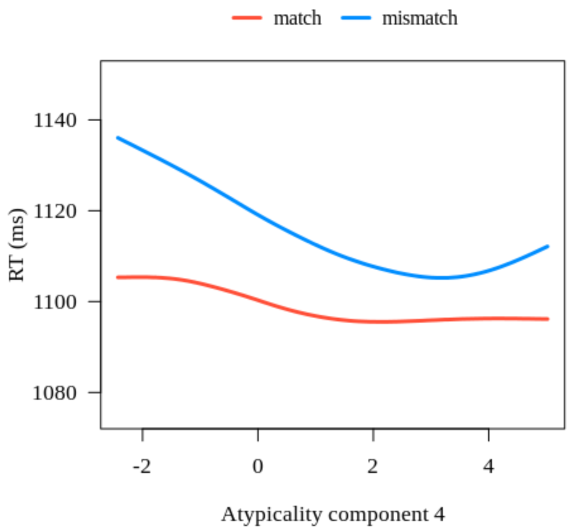


Figure 11: Different effects of atypicality component 4 for responses that match (matched; red) and did not match (mismatched; blue) the dominant response per image.

Figure 11 shows that the negative correlation holds for the mismatched responses but not for the matched responses. Moreover, the difference in predicted RT is greatest for the most prototypical nouns.

7. Discussion

I find evidence that de-lexicalized syntactic information affects the production of nouns in isolation. While I considered two types of information – syntactic diversity and syntactic atypicality – I only find support for the latter. This finding is consistent with the findings from morphology for comprehension (Milin et al., 2009). Moreover, the effect was limited to atypicality for rightward-facing dependencies. More atypical rightward distributions correlated with faster naming RTs. This seems to be at odds with the traditional interpretation of atypical words as being more difficult to process. An alternative explanation is that prototypicality leads to interference. From this perspective, prototypical words would spread activation to a greater number of competitors via their shared syntactic representations. This competition takes longer to settle as more competitors become more active. If so, this competition only arises when a highly conventionalized form is not immediately selected based on the visual input. In this way, the results align with those observed for inflectional entropy here and in Tabak et al. (2010).

This study differs from others that have applied prior syntactic distributions to noun production. For example, Hendrix et al. (2017) coupled their picture naming task with an overt syntactic context. Picture stimuli in that study were preceded by *preposition + determiner* contexts. Therefore, they observe syntactic effects in a (partially) syntactic task. Moreover, the relative entropy that they employ is based explicitly on the same syntactic construction that they test in the experiment (the prepositional-phrase construction). The data presented here did not require syntactic processing, and the syntactic measures were based on the entire dependency space. They are thus the first – to my knowledge – to demonstrate aggregate syntactic effects on RTs in isolated noun production. Crucially, the syntactic effect was observed over and above a number of controls known to impact production latencies. Moreover, the effect remained even after extreme values were removed. All of this suggests a robust, albeit weak, effect of syntax on bare-noun picture naming. I thus add to a growing number of studies that have uncovered syntactic effects in lexical production using tasks that do not, at least on their face, require any syntactic processing (e.g., Cubelli et al., 2005; de Simone & Collina, 2015; Gregory et al., 2012; cf. La Heij, et al., 1998). That syntactic representations should participate directly in the processing of words in isolation is also compatible with the constructionist theories of the structure of language. Most constructionist theories argue for a unified memory-based system, the *constructicon*, which encompasses everything from words to abstract phrasal and clausal templates (Bybee, 2010; Diessel, 2015; Goldberg, 1995; Langacker, 1987).

The atypicality effect was specific to dependencies in the rightward-facing direction. This could reflect a system tuned to the future. Much research has documented the highly incremental nature of speech production (V. Ferreira, 1996; Allum & Wheeldon, 2007), particularly under time constraints (F. Ferreira & Swets, 2002). The participants in this study were instructed to produce the picture names as quickly as possible. On analogy to the sentence production literature, perhaps the pressure to produce words quickly engaged a strategy that privileges forward-facing rather than backward-facing syntactic relations. Activation then circulates between lexical and syntactic relationships to produce the interference effect described above.

Finally, results of the analysis showed an interesting symmetry with earlier findings for inflectional entropy. Tabak et al. (2010) report a facilitatory effect for inflectional entropy, but only for non-target responses. I replicate that effect here. Interestingly, the atypicality measure also depended on whether the response matched the consensus responses across participants. Specifically, the effect surfaced only for non-dominant responses. No effect was observed for the dominant responses. This suggests that although the effect is robust, it is easily swallowed by other factors, including the efficiency of mapping in the pathway from visual to conceptual-semantic to lexical processing.

I have so far demonstrated that prior syntactic distributions affect isolated noun production. However, the effect was weak and only surfaced for atypicality, not diversity. We know from earlier research that syntactic production tasks are also sensitive to prior distributions (Hendrix et al., 2017) and sometimes even bring about effects that are not observed in bare-noun naming (La Heij et al., 1998). In a second experiment, I test whether requiring participants to produce nouns in a minimal syntactic context alters the relationship between prior syntactic distributions and response time in picture naming.

E. Experiment 2: Noun-phrase picture naming

I performed another picture naming experiment with the same general structure as Experiment 1. In this case, I asked participants to produce full noun phrases. Explicitly engaging the syntactic system in this way could have one of two effects. On the one hand, it could result in much stronger activation of the syntactic representations presumed to underly the effect observed for bare-noun naming in Experiment 1. Stronger activation cycling between lexical and syntactic representations could increase the effect of prior syntactic distributions, which filter the strength and pathways of these relationships. On the other hand, repeated production of a single syntactic construction could overwhelm any effects of prior syntactic distributions by 'clamping' the spreading activation within a single lexico-syntactic pathway (noun ↔ *det*). I test this possibility by pitting the diversity and atypicality measures against measures targeting the specific syntactic and syntagmatic properties of the response frame. I perform two RT analyses: one locked to the onset of *the*, one locked to the onset of the noun.

1. Stimuli and Design

The stimuli and design were identical to those described for Experiment 1.

2. Participants

31 undergraduate students were recruited to participate (N(female) = 24; mean age = 19.29), all of whom were native speakers of English with normal or corrected-to-normal vision. All participants were treated in accordance with the American Psychological Association guidelines for ethical human research.

3. Procedure

The procedure was almost identical to that of Experiment 1, with two main differences. First, participants were instructed to name pictures using the frame *the* + NAME. For example, they should have responded “the banjo!” upon being presented a picture of the banjo. As in Experiment 1, the responses were recorded into individual WAV format audio files. Second, RTs were automatically derived from the WAV files using forced alignment rather than hand coding. Forced alignment involves mapping phonological representations of orthographic words onto audio signals using an algorithm trained on detecting the most likely segment given various acoustic properties. I use the Prosodylab-Aligner (Gorman, Howell, & Wagner, 2011) to create force-aligned versions of the response files at the whole-word and segmental level. Performance of the aligner was checked by manual inspection using Praat (Boersma, 2001) for 5 randomly selected files per participant and found to be satisfactory. Response times were extracted from the time-aligned file for the determiner and the noun. Additionally, the quality of the vowel in the determiner was extracted: unstressed *uh* (ə), stressed *uh* (ʌ), or stressed *ee* (ɪ͡j).

4. Control predictors

All controls from Experiment 1 were applied here, as well. Due to the phrasal nature of the task, I introduce three additional controls related to the syntactic and syntagmatic structure of the responses. First, I include a measure of the association of the noun to the determiner relation *det.* I operationalize lexico-syntactic association as the log-odds ratio (LOR) of the noun as it appears in the det relation vs. out of the relation given the behavior of all other verbs and relations. This measure is best visualized as a cross-tabulation, illustrated in Figure 12.

	det relation	all other relations
target noun	*a*	*b*
all other nouns	*c*	*d*

Figure 12: Table for calculating log-odds ratio of target words in the *det* relation

In Figure 12, cells *a* through *d* represent frequencies. Using these frequencies, the LOR of the target noun in the *det relation* equal to the log of the odds that the target noun occurs in the det relation divided by the odds that any other noun occurs in the *det* relation, or LOR = (a / b) / (c / d). LOR takes increasingly positive values to the extent that the target noun's frequency in *det* is higher than it is in other relations, and/or to the extent that other nouns tend to load in the opposite direction. This situation indicates positive association between the noun and the syntactic relation. LOR takes increasingly negative values in the opposite situation, in which case the noun is negatively associated with the relation. I expect nouns that are strongly associated with the *det* relation to be prepared and articulated more rapidly in this relation than words that disprefer this relation. This accounts for the statistical expectation at the syntactic level.

At the syntagmatic level, I control for how likely the noun is to follow *the* in sequence. I computed the conditional bigram surprisal of each unique *the* + NOUN sequence produced by the participants. Conditional surprisal is defined as the negative log of the bigram probability $p(w_1, w_2)$ divided by the unigram probability of the first word $p(w_1)$, or $S(w_2 \mid w_1)$ = - log $p(w_1, w_2)$ / $p(w_1)$. This measure captures the unexpectedness of the noun given an immediately prior *the*. Higher values reflect less expected transitions. Estimates of $p(w_1, w_2)$ and $p(w_1)$ were drawn from *n*-gram lists based on the 560-million word *Corpus of Contemporary American English* (COCA; Davies, 2008–).[7]I expect conditional surprisal to correlate negatively with RTs, with more surprising transitions taking longer to produce.

[7] These lists are freely available at https://www.ngrams.info/download_coca.asp.

Finally, I include a three-level factor for the quality of the vowel of *the*: stressed and unstressed mid central vowels (*thuh*) and front vowel with palatal off glide (*thee*). The pronunciation of *the* is known to correlate other signs of production difficutly (e.g., pauses, filler such as *uh*, and so on)/ Specifically, speakers produce *thee* more frequently when they experience problems during production (Fox Tree & Clark, 1997). Furthermore, listeners respond to these cues in ways that suggest that they are subconsciously tuned to the difficulty faced by the speaker (Arnold, Tanenhaus, Altmann, & Fagnano, 2004). Production problems in this experimental context are expected to surface as increase RTs. Based on this reasoning, I expect longer RTs for responses introduced by *thee.*

5. Results: *the* RT analysis

I first fit a GAMM predicting RTs at the onset of the determiner. First, I removed all RTs greater than two times the standard deviation from the mean (4% of responses containing nouns for which I have diversity and atypicality measurements. The remaining RTs showed a strong positive skew. A Box-Cox power analysis suggested the inverse transform to normalize the RT distribution. I substitute the negative inverse of RTs (to preserve the original sign) for the raw values as dependent variable in the analysis. 2,208 observations remained after these cuts. I next checked for collinearity between the predictor variables, which proved to be unacceptably high (condition number $\kappa > 80$). First, the other predictors were residualized out of word frequency. Then, all predictors (except for word frequency) were residualized out of bigram surprisal. Substituting the residualized variables for the raw variables revealed that collinearity had been reduced to a moderate but acceptable level ($\kappa <$ 12).

Smoother terms were included for log trial number, log previous RT, naming diversity, objective visual complexity of the image, residual word frequency, residual bigram surprisal, LOR, the five diversity components, and the four atypicality components. Shared name, length in syllables, and the quality of the vowel in *the* were added as parametric terms. Finally, random intercepts were included for participants, images, and response lemmas.

An initial analysis revealed significant effects of the diversity component 2 (*p* =.01) and the atypicality component 2 (*p* = .03). However, the variables showed distant and sparse observations at the lower and upper extreme, respectively. To guard against the influence of outliers, I refit the model without these observations (7% of remaining observations). Both effects remained significant. I report this model (summarized in Table 4) as it represents a more conservative perspective on the nature of the effects.

Table 4: Significant predictors of response time at *the* in *the* + N picture naming

Smooth terms	***eDF***	***refDF***	***F* value**	***p* value**
previous RT (log ms)	6.49	7.49	74.99	<.001
age of acquisition	1.00	1.00	11.41	<.001
name diversity	2.93	3.43	34.18	<.001
objective visual complexity	1.00	1.00	9.37	.002
diversity component 2	1.00	1.00	8.26	.004
prototypicality component 3	2.10	2.44	3.76	.02
Random effects				
name lemma	0.92	1.00	11.95	<.001
subject	78.94	26.60	0.49	<.001

The significant control variables were all in the expected direction. Unlike Experiment 1, the model did not reveal effects of trial number or inflectional entropy. The former indicates that fatigue did not impact participants in this study to the same extent that it did participants in Experiment 1. To explore whether the lack of an inflectional entropy effect was due to a moderating effect of response type, I refit the model with factor smooths on inflectional entropy for dominant responses and non-dominant responses. The interaction was significant (corrected *p*=.006), but opposite of the one observed in Experiment 1: non-dominant responses showed no effect but dominant responses showed a slight facilitation at the upper registers of the entropy. When the most conventional response was available, participants were quicker to produce *the* for nouns with more even splits across singular and plural forms.

Two of the syntactic components surfaced as significant: diversity component 2 and atypicality component 3. These effects along with the component loadings are plotted as Figure 13 and Figure 14, respectively.

The upper panel of Figure 13 shows the component loadings of diversity component 2. This component loads positively on the as-modifier dimension with support from the rightward-facing as-modifier and leftward diversity. These dimensions are contrasted with that of rightward headship. The lower panel shows that this component has a facilitatory effect on RTs at *the*. This facilitation means that nouns that are used in the most diverse array of modifier and leftward-facing dependencies allow for faster production of the noun phrase. By contrast, nouns that are used in a diverse array of rightward-facing headship dependencies to the exclusion of modifier and leftward dependencies lead to longer production latencies.

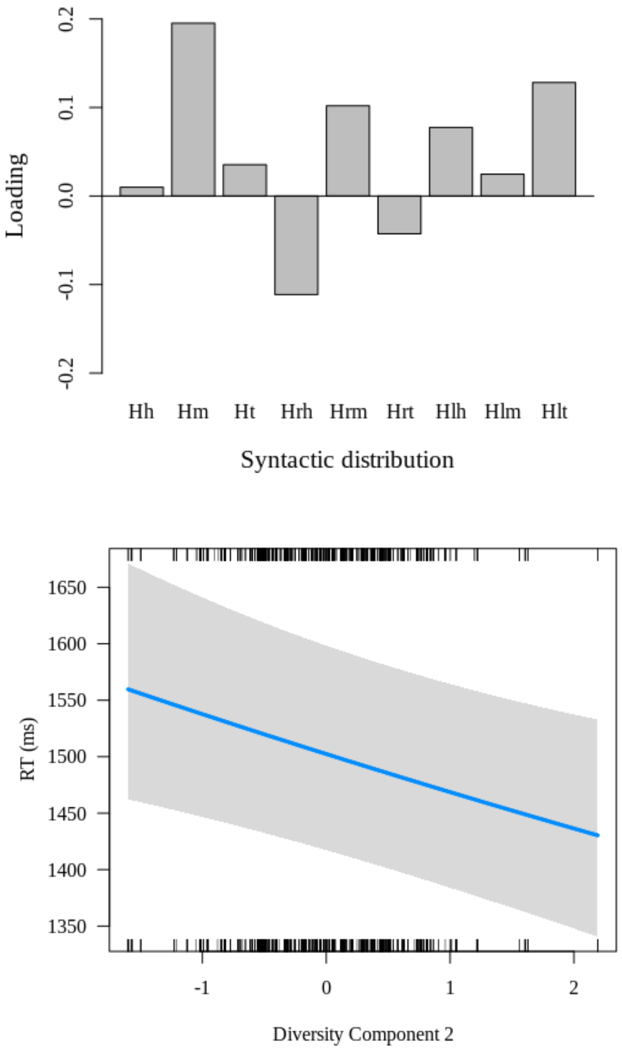


Figure 13: Effect of diversity component 2 (the). Top panel: Component loadings for diversity component 2. Positive values reflect increasing diversity as (rigthward-facing) modifier, with support from leftward diversity; negative values reflect increasing diversity as rightward head. Bottom panel: Effect of diversity component 2 on RTs measured at *the*. The solid line plots the regression curve. The shaded area represents the 95% confidence interval.

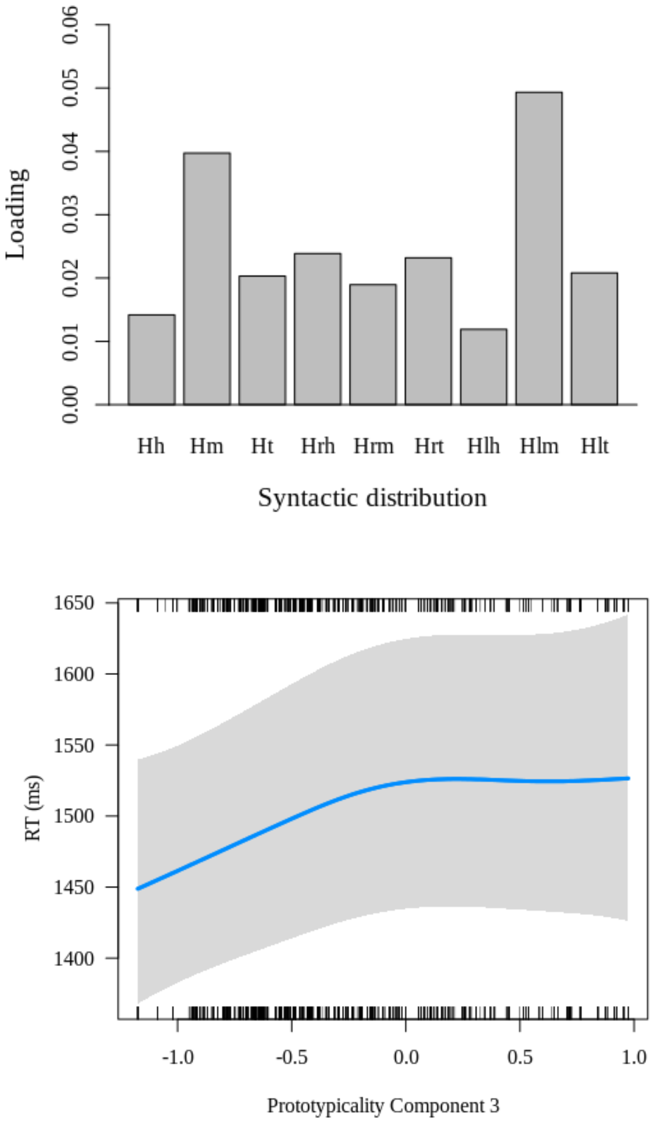


Figure 14: Effect of atypicality component 3 (the). Top panel: Component loadings for atypicality component 3. Loadings indicate that positive values reflect increasing distance from the noun prototype. Bottom panel: Effect of atypicality component 3 on RTs. The solid line plots the regression curve. The shaded area represents the 95% confidence interval.

Figure 14 plots the effect of atypicality component 3. The top panel shows that this component loads positively on all syntactic dimensions. It therefore reflects general divergence from the syntactic prototypes of nouns. The lower panel reveals a non-linear effect of this component on *the* RTs. For positive loadings, no effect was observed; however, for negative loadings, RTs become increasingly faster. Therefore, when the nouns are produced in a syntactic context, that is, a noun phrase, more prototypical nouns resulted in faster initializations of the phrase. Atypical nouns resulted in slower phrasal RTs than prototypical nouns, but the effect of increasing atypicality levels out around loadings of 0.

We now turn to RTs measured at the noun itself. Naturally, these RTs are strongly correlated with those measured at *the* (r = 0.97). However, several of the predictors may be more closely aligned with production of the noun. For example, the quality of the vowel in *the* may reflect additional planning that takes place during the production of *the*. Such mid-speech lexical selection processes should be expected given that prior research has shown that speakers rely on grammatical forms to mitigate on-line planning difficulties (e.g., Clark & Fox Tree, 1997). These effects may therefore arise only once the speaker has begun to produce the noun phrase. I also guard against possible task-based strategies. For example, participants entrained on producing *the* + N may use *the* as a crutch, producing it early while still searching for the noun. In this case, the noun-locked RTs may be more reliable indices of lexical selection processes.

6. Results: Noun RT analysis

I fitted a second GAMM predicting RTs at the noun. This model had the same structure as that fit for the RTs taken at *the*. Similar to the *the*-locked RTs, a Box-Cox analysis of the noun-locked RTs suggested the inverse transformation to normalize the distribution. I again removed outlier RTs falling more than two times the standard deviation from the mean. An initial attempt at modeling revealed a significant effect of diversity component 2 (*p* = .002), but as with the *the*-based analysis, this component contained a very sparsely populated region at the lower extreme. I removed these observations and refit the model, which actually increased the significance of the effect (*p*<.001). I report results for the refit model. The summary of significant effects is given in Table 5.

Table 5: Significant predictors of response time at N in *the* + N picture naming

Parametric terms	***β***	**SE**	***t* value**	***p* value**
intercept	-0.58	0.009	-63.15	<.001
number of syllables	-0.01	0.005	-2.20	.02
the vowel: stressed *uh*	0.03	0.004	7.70	<.001
the vowel: *ee*	0.03	0.007	3.71	<.001
Smooth terms	***eDF***	***refDF***	***F* value**	***p* value**
previous RT (log ms)	6.01	7.03	70.62	<.001
age of acquisition	1.30	1.46	8.78	<.01
name diversity	2.90	3.34	42.95	<.001
visual complexity (log file size)	1.00	1.00	6.15	.01
diversity component 2	1.00	1.00	8.26	<.001
Random effects				
image	0.95	1.00	21.39	<.001
subject	104.20	280.00	0.70	<.001

Unlike *the*-locked analysis, this analysis revealed two significant parametric terms. First, the number of syllables resulted in faster naming latency at the noun. This result was unexpected on general grounds, but may be accounted for as a function of the experimental design. Meyer, Roelofs, & Levelt (2003) find a numerical 18 ms advantage for multi-syllabic words over mono-syllabic words, but only when they used a mixed design that included both mono- and multisyllabic targets. Participants in the present study produced both monosyllabic and multisyllabic words, akin to the mixed condition in the experiment of Meyer and colleagues. The effect did not reach significance in that study, and I find inconsistent evidence for it here: the effect only arose for RTs taken at the noun in the *the* + N naming condition. Moreover, this study differs from that of Meyer and colleagues in several ways. First, in this experiment, participants were not familiarized with the picture names. Greater demands were therefore placed on the lexical retrieval system. This increased demand, coupled with the mixing of multiple word lengths, could have exacerbated the advantage for longer words (if indeed task demands drive this effect). Second, I analyzed all responses, regardless of whether they were the “intended” name. However, this is unlikely to have had an impact, as I find no evidence for an interaction between word length and whether the response matched the dominant response across participants for that image. Third, some participants in this sample produced three-syllable responses, hence providing a broader range of longer words. Again, this is unlikely to have had an impact. Removing these tokens and re-running the model did not alter the significance. Therefore, this result may be a function of the task demands, but further research is necessary to determine why it surfaces in some conditions and not others.

Second, the quality of the vowel correlated with the response times in the expected directions. When participants produced unstressed *uh*, the noun was produced earlier than when they produced a stressed *uh* or *ee.* Unstressed *uh* indicates that the speaker has cliticized the determiner to the noun, which may suggest that the noun was selected quickly relative to the highly available determiner, allowing the two to be integrated prosodically. Stressed *uh* and *ee* indicate a weaker prosodic bond between *the* and the noun, which could reflect either task-based early production of *the* or selection-based lags in accessing the noun.

All other control predictors had similar effects to those observed for *the*-locked RTs, and so will not be discussed further.

Only one of the syntactic components surfaced as significant. As with the *the*-locked analysis, diversity component 2 was a strong predictor of RTs. Unlike the *the*-locked analysis, no effect of syntactic atypicality was observed. The component loadings and fitted effect of diversity component 2 are presented in Figure 15.

The top panel of Figure 15 shows the same component loadings that appear in the top panel of Figure 13, but it is repeated here for convenience. Again, these loadings indicate that diversity component 2 loads positively for as-modifier diversity, with support from leftward diversity. It loads negatively for rightward headship. Exactly as was observed for *the*-locked RTs, diversity component 2 was negatively correlated with noun-locked RTs. Nouns with the most diverse as-modifier and leftward syntactic distributions are produced faster within the context of the noun phrase. By contrast, nouns with diverse rightward as-head distributions are produced relatively more slowly.

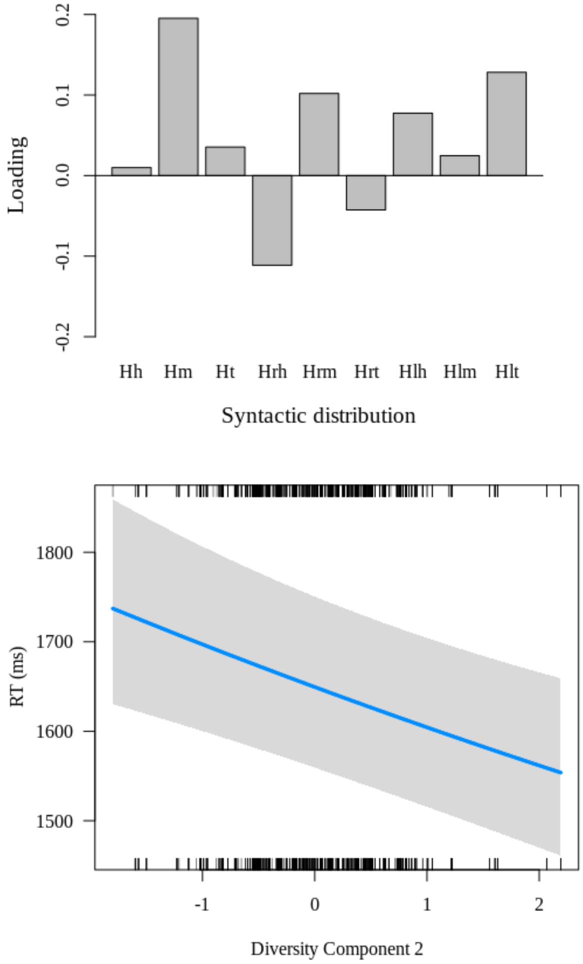


Figure 15: Effect of diversity component 2 (N). Top panel: Component loadings for diversity component 2. Loadings indicate that positive values of the component reflect increasing diversity as (rigthward-facing) modifier, with support from leftward diversity; negative values of the component reflect increasing diversity as rightward head. Bottom panel: Effect of diversity component 2 on RTs measured at NOUN. The solid line plots the regression curve. The shaded area represents the 95% confidence interval.

7. Discussion

Experiment 2 demonstrates that the effect found for bare-noun picture naming in Experiment 1 is not limited to the (rather artificial) bare-noun naming task. Even when names are produced in a fully lexically and syntactically predictable context, picture naming RTs vary in response to aggregate prior syntactic distributions. Notably, these effects hold where measures of the syntactic and syntagmatic predictability of the nouns in the *the* + NOUN frame do not. Also of interest is the fact that the syntactic measures affected onsets of both the determiner and the noun. This suggests that participants were not simply producing the and waiting for the noun to come to mind. Instead, there appears to be a process that proceeds as follows: lexical selection, mediated by prior syntactic distributions, integration with the determiner, then articulation of the NP.

Both *the*-locked and noun-locked RTs showed an effect of diversity component 2. This component was most strongly associated with modifiership and leftward relations. More diverse modifiers of leftward content were produced faster at *the* and the head noun. This effect is complex, and could arise from multiple factors. Because this study is largely exploratory, especially with respect to the individual contributions of the different syntactic dimensions, I offer several possible explanations.

On the one hand, this component contrasts leftward and rightward as-head diversities. Leftward as-head diversity, particularly in the absence of rightward as-head diversity, facilitated naming. A common parse of the determiner + noun complex identifies the noun as the head. By this analysis, the nouns that are distinctively leftward heads are produced faster when they appear in a syntactic construction in which they are leftward heads. Applying the connectionist metaphor, activation could spread between syntactic and lexical representations. Syntactic-level activation would stay primed from repeated production of the noun phrase. Moreover, some activation should spread from the syntactic nodes to various lexical nodes according to the strength of their association. Lexical-level activation would also be triggered bottom-up by the visual → conceptual → lexical pathway. Once the noun lemma is activated, it spreads activation back into the syntactic nodes. When the patterns of activation sync up between the two, the noun with the strongest resonance wins out. Nouns that make the strongest candidates for leftward headship have a better chance of syncing up with projected determiner relation. Such resonance is at its lowest when the syntactic resonance of the noun overlaps not at all with the currently active syntactic space.

This explanation accounts for the headship (and overall leftward diversity. However, it does not explain why these should align with modifiership. Looking into the distributions, rightward modifiership is almost exclusively reserved for three categories: the initial noun in NOUN + NOUN compounds (written with a space), subject of active verbs, and subject of passive verbs. The latter two are much more heavily populated. Now, subjects must be selected and articulated early in the production of a sentence. They are often seen as “hitching posts” to prior discourse more so than subordinates of the main verb (Chafe, 1994), and experimental results show that the syntactic structure of the clause tends to be built to accommodate whatever noun has been selected to initiate the clause (Tomlin, 1995; Myachykov, Garrod, & Scheepers, 2009). Therefore, words that are diverse rightward modifiers may have a history of rapid selection as subjects partially independent of the clausal structure that follows. As frequent subjects, they may establish a lasting processing benefit in production that surfaces even in contexts for which no sentential continuation is required.

This effect of modifiership could also be driven by the task. I asked participants to produce the same structure repeatedly. In each token, the participants produced the determiner prior to the head. For one, this means that the determiner was a stable component of every trial. Participants likely maintained a high level of activation of the determiner, searching only for the noun. This preparation could extend into the syntactic domain: on each trial, the participant must hitch a new noun to the trial-stable determiner, The task demands could therefore place *the* in an ad hoc position of hierarchical superiority – it functions as the syntactic head. Although the actual parses used to construct the syntactic distributions treat nouns as the heads of determiners, the opposite relationship has also been proposed – namely that determiners head a determiner phrase in which the noun is embedded. This approach is particularly popular within theories that use phrase-structure representations (Abney, 1987). Therefore, the hierarchical superiority of *the* might be a general – and not ad hoc – property of the syntactic system of English. However, if this were the case, we should have expected a stronger role from leftward modifiership, which is nearly absent from this component. Therefore, the role of modifiership (if any) may be to privilege nouns that are easily integrated into modifier space when participants are tasked with “modifying” the ad hoc experimental frame *the* + ___. At this stage, this proposals remain hypothetical. I leave it to future research to pin down the exact cause of the observed facilitation.

Turning to the atypicality effect, the more prototypical the noun across all syntactic distributions, the faster the production of *the*. However, I observed no such effect at the noun. One possibility is that the onset of the determiner is put off until a threshold of activation within the lexical system is reached, even if some competition remains among semi-active lexical forms. In other words, participants wait to begin articulating the noun phrase until they have a sense that some noun is available. Lexical selection processes then continue as the determiner is articulated, allowing the system to settle on a target name. By this account, the prototypicality of the head exerts its effect at the phrasal rather than lexical level. Or there could be some kind of push-and-pull whereby the facilitation for producing *the* – which should pull production of the noun up in time given the high degree of correlation between the *the*-locked and noun-locked RTs – cuts against the interference produced by activating lexemes that occupy densely populated areas of the lexico-syntactic space. The end result would then be a null effect atypicality at the noun.

F. General Discussion

In two picture naming studies, I find effects of prior syntactic distributions on lexical production latencies. Similar effects have been reported in earlier studies (e.g., Hendrix et al., 2017). The present study improves on that work in several important ways. First, prior syntactic distributions were measured across constructions rather than within a single construction. In this way, we arrive at much more complete picture of the syntactic behavior of nouns. Second, I explicitly attempt to remove information associated with lexical context when measuring cross-constructional information. I thereby ensure that the measures tap into the abstract syntactic space, maximally divorced from the surface context. These first two steps are necessary given claims that have been advanced recently about the fundamental role of surface patterns in determining word learning, hence word processing (e.g., Baayen, et al., 2011). Third, different dimensions of syntactic behavior were contrasted, including hierarchy and word order. Fourth, syntactic diversity, a measure of breadth and strength of the lexico-syntactic relationships, was contrasted directly with syntactic atypicality, a measure of how similar a noun is to other nouns in its syntactic behavior. This contrast is necessary given that research from comprehension suggests that the two have independent effects, as reflected in the electrophysiological signature (Linzen et al., 2013). Fifth, estimates of diversity and prototypicality were carefully corrected for underestimation bias. Other studies rely on maximum-likelihood estimates, which are known to be biased (Miller, 1955). The corrected estimates therefore better approximate the true syntactic behavior of nouns.

In Experiment 1, participants named pictures with isolated nouns. We saw only a weak effect of atypicality. Nouns that were more distant from the prototypical noun were produced faster. This result resembles that observed for inflectional prototypicality: nouns with inflectional distributions that differ from the typical noun given their class take longer to produce (Baayen, Levelt, Schreuder, & Ernestus, 2008). I interpret this as an interference effect: when the target name occupies a densely populated corner of the syntactic space, it shares its activation with many other lexical items, which makes it more difficult to isolate the target. This account invokes the notion of competition, which has been repeatedly challenged in recent years (Dhooge & Hartsuiker, 2010; Miozzo & Caramazza, 2003). The alternative explanation involves selection by exclusion of competitors. The difference between interference by competition and interference by exclusion lies primarily in the locus of the effect, that is, whether the competitors interfere with selection of the target (competition) or with the removal of competitors from the response buffer (after the target has already been selected; exclusion). Either explanation is compatible with the current findings, but future research might investigate interference via the syntactic space using, for example, the picture-word interference (PWI) paradigm, especially the delayed variant of PWI (e.g., Dhooge & Hartsuiker, 2010).

Importantly, the atypicality effect was driven by rightward relations. This specificity validates the fine-grained approach I adopt here. It also suggests a functional motivation. Words are rarely experienced outside of syntactic contexts in natural speech. A central principle of usage-based linguistic theory is that experience shapes how language is represented and processed (e.g., Barlow & Kemmer, 2000). The effects of the syntactic measures are therefore expected to be tuned to the needs of speakers who are producing connected speech, including sentences. Much evidence from sentence production suggests that the language production system is highly incremental (Allum & Wheeldon, 2009; V. Ferreira, 1996; F. Ferreira & Swets, 2002). Incrementality in one sense is concerned with the sequencing of elements one after another (i.e., word order). Each choice of a word in these sequences constrains the possible continuations, that is, the possibilities for how "downstream" (i.e., rightward) may be integrated into the unfolding structural template. Some planning happens in advance (e.g., Myachykov, Scheepers, Garrod, Thompson, & Fedorova, 2013), but a word at any stage in the production is still related to pre-planned structures via rightward-facing dependencies. This focus on the future could influence the strength of the relationships between words and syntactic forms, a residue of use that constrains production even in the absence of syntax. How this effect fits into the broader evidence base notwithstanding, the correlation of prior distributions with isolated noun production provides strong support for the probabilistic network models of the "constructicon" (e.g., Diessel, 2015).

When participants were asked to produce full noun phrases – *the* + NOUN – rather than bare nouns, an effect of atypicality was observed in only one of the two positions, and a stronger effect of syntactic diversity in both positions. Regarding atypicality, we saw a general distance effect: nouns that most closely approximated the noun prototype on all dimensions were produced faster. However, the effect only surfaced for RTs measured at the determiner *the*. Not only do the significant components differ in shape across Experiment 1 (rightward distance) and 2 (general distance), they correlate with RTs in opposite directions. I proposed that the bare-noun effect arose from interference within the densely populated space of the lexico-syntactic network. But this cannot explain why prototypical nouns produce faster onsets of *the* in the noun-phrase naming task. To explain this difference, we must consider two things: (1) facilitation was associated with general rather than right-facing distributions and (2) the facilitation was only observed at *the*, not at the noun itself. I propose that the advantage for prototypical nouns is based on a tendency to delay onset of the noun phrase until a threshold of activation has been reached within the lexical network. The system uses this activation threshold to determine whether a head noun will be available when needed. This could be a general strategy, or one dictated by the time-pressure of the task. Once this threshold is exceeded, the processor prepares the determiner for articulation. A greater general amount of activation within the lexico-syntactic network ensures that some form will be available, though perhaps not immediately. Interference from competing forms can delay articulation of the intended noun (as was observed in the bare-noun study). Thus, while the earlier production of *the* puts pressure on the system to articulate a noun soon after, the noun is simultaneously delayed by interference. Together, this push-and-pull produces a null effect at the noun. The present data do not allow us to evaluate this proposal. Nevertheless, the fact that the effect of prior distributions was observed at all is non-trivial.

Under any theory, we expect the syntactic system to switch on when the task requires the participant to produce a syntactic utterance. However, for theories that separate words from syntax, the syntactic system is expected to limit itself to the task at hand (i.e., to look for lexical entries labeled as nouns that are compatible with the semantic input and communicative intention; e.g., Borer, 2005; Pickering & Branigan, 1998; Bresnan, 2001; Chomsky, 1995; Ramchand, 2007). The present findings present a serious challenge to these theories.

An effect of diversity was also observed, and this effect was constant across both *the*- and noun-locked RTs. The precise source of the effect remains unclear, given the complex nature of the component loadings, as well as an uncertain relationship between these loadings and the demands of the task. I proposed three possibilities. First, increasing leftward as-head diversity, as well as total leftward diversity were associated with faster RTs. Nouns are often treated as the head of determiners. Therefore, nouns that are most associated with left-facing relationships, when produced in a left-facing relationship, are produced faster. The fact that the effect is observed at the determiner could be due to a mechanism similar to that which underlies the prototype effect. For example, the participants know that they must produce the noun phrase, in which the only unknown is the noun. The leftward syntactic space may be primed during selection, providing stronger feedback to diverse leftward heads, which could boost activation potentials above the threshold required for initiation of the determiner. Diversity also speeds production of the noun, unlike prototypicality, which generates interference. If this explanation is correct, then syntactic diversity plays a role in both lexical selection and integration into syntactic frames.

Second, the component loaded positively for rightward as-modifier diversity. Inspection of the vectors reveals that rightward as-modifier diversity is defined primarily relative to subjects of active and passive clauses. Subjects are produced early within sentences, often before the speaker has committed to the overall structure of the clause (e.g., V. Ferreira, 1996). Subjects are also highly topical and available, providing a bridge between prior discourse and the current clause (Chafe, 1994). Repeated topicality and availability may be signs of cognitively or culturally salience, which translates into faster processing (for experimental evidence, see e.g., Tomlin, 1995; Allum & Wheeldon, 2007). Therefore, nouns that are commonly used as subjects may be selected and produced more quickly outside of clausal contexts as a matter of general discourse-functional salience, in other words, some aspects of the measure provide an index of discourse-level organization . This benefit apparently only arises when the noun is produced in a minimally syntactic context; however, rightward as-modifier relations also played a role in the atypicality effect for bare-noun naming. Future research is needed to clarify whether this effect is truly restricted to syntactic contexts. This would involve isolating forms that serve distinctively as subjects relative to other syntactic relations.

Third, the component loaded positively on as-modifier diversity. The repetitive nature of the task – producing *the* + NOUN repeatedly in sequence – may have led participants to develop an ad hoc strategy of linking nouns to the determiner. The predictability of the determiner could establish it as an experiment-specific head for a response frame into which the noun is slotted. Another possibility is that the determiner really is the syntactic head of the noun (the “Determiner Phrase” hypothesis; Abney, 1987). In either case, the nouns that show high utility as modifiers might be more easily accommodated into the frame. This possibility could be tested in future research.

The main conclusions to be drawn from these data are as follows. First, cross-constructional distributions affect bare-noun production. Producing a noun inevitably involves some information being passed between lexical and syntactic representations, exactly as predicted by usage-based construction grammar. Moreover, these prior distributions impact noun phrase production, suggesting that local syntactic contexts interact with the aggregate syntactic behavior of individual words. Much work has shown that the predictivity of a word in a local syntactic context based on its prior association with that specific context affects production (e.g., Gahl & Garnsey, 2004). I extend this work to show that syntactic probabilities outside of the target construction can promote integration into that construction if certain properties overlap (such as the direction or hierarchical structure of the relations). The low-level measures I adopt here do not even begin to approximate the complexity of the true syntactic space of a language, which would include higher-order argument-structure constructions (e.g., X VERB Y), idioms (*kick the bucket*), partially-schematic structures (X puts up with Y), and so on. However, the dependency-based perspective is powerful enough to find effects in isolated word production. The dependency formalism thus represents an important new tool for understanding how prior contexts influence processing within a specific context. However, the system could be expanded to include composite dependency structures akin to traditional syntactic constituents (i.e., phrase-structural constituents) and the abstract constructional templates of construction grammar. Careful construction of such expanded syntactic spaces, along with the statistical techniques employed here, presents many opportunities for advancing our understanding of the lexico-syntactic interface and its functional architecture.

IV. Effects of syntactic distributions on language acquisition

A. Introduction

The linguistic input that children receive was argued by some to be insufficient for supporting language acquisition. How could a child extrapolate an infinitely generative combinatorial system on the basis of a handful of unsystematic exemplars? This argument, known as the “poverty of the stimulus” (e.g., Chomsky, 1980), has been applied with great force by those who support a theory of inborn linguistic ability, one driven by the twin engines of “Universal Grammar” and the presence of a “Language Acquisition Device”. However, much research has since pushed back against this notion, revealing that the information contained by linguistic signals experienced by children had been severely underestimated. For example, children as young as 8 months old are able to leverage distributional biases in sequences of sounds to segment the speech signal, even with only small amounts of input (e.g., Saffran, Aslin, & Newport, 1996). Moreover, they can apply similar strategies to induce grammatical categories such as the gender classes of Russian nouns (Gerken, Wilson, & Lewis, 2005). This research re-frames the problem of language acquisition. Instead of asking how an innately specified grammar unfurls within the child, irrespective of input, researchers in statistical learning ask what aspects of the input children can leverage to produce and comprehend utterances that are consistent with the input. This research not only provides a plausible explanation for the development of language without the weighty assumption of a uniquely human, inborn mental system of significant complexity; it offers a straightforward link between how language is acquired and processed (e.g., Romberg & Saffran, 2010; Seidenberg & MacDonald, 1999).

Early research on statistical language learning in young children focused on segmental transitions (e.g., Saffran et al., 1996). A common finding in this research is that children attend to the transitional probabilities of sounds, preferring to segment the speech stream at low-probability junctures. Since then, a host of more abstract distributional profiles have been proposed to impact the linguistic development of children, including those based on morphological (e.g., Baayen et al., 2006; Gerken et al., 2005; Stoll et al., 2012), as well as lexical co-distributions (e.g., Mintz, 2003), among others (e.g., those related to features of the interactional or prosodic context, and so on; see Romberg and Saffran, 2010, for a succinct review). In the present study, I propose two additional sources of information that young children may exploit when learning to produce words based on their syntactic distributions: the diversity of constructions in which they are observed, and the typicality of these distributions relative to other words. Recent work on language production and comprehension suggests that adults are sensitive to both types of information (Baayen, Milin, Filipović-Đurđević, Hendrix, & Marelli, 2011; Lester, Feldman, & Moscoso del Prado Martín, 2017; Lester & Moscoso del Prado Martín, 2016; Linzen, Marantz, & Pylkkänen, 2013). Experimental research on infants suggests that even children as young as 18-25 months have developed abstract syntactic knowledge (e.g., Gertner, Fisher, & Eisengart, 2006; Lidz, Waxman, & Freedman, 2003) and that they can use this knowledge to learn novel words (e.g., Lidz, White, & Baier, 2017). However, no study to my knowledge has examined how aggregate syntactic distributional information contributes to word learning in young children. I explore this possibility by correlating the syntactic diversity and typicality of nouns with the age at which they are first produced in the naturally occurring speech of English-speaking children. Following work on morphological distributions, I expect more diverse and more typical nouns to emerge earlier in child speech. I model lexical acquisition using a survival analysis technique (Cox Proportional Hazard Regression; Cox, 1972). These results confirm that more diverse and more typical words tend to be produced earlier, over and above a number of other control variables.

Statistical properties of the input that children receive support lexical acquisition on several levels. Biased relative positioning of phonological segments (phonotactics) can help children carve out candidate words from the continuous speech signal (Saffran et al., 1996), and children can rapidly map newly segmented words to meanings (Graf Estes, Evans, Alibali, & Saffran, 2007). Beyond single segments, the distribution of stems across their morphological variants also correlates with language development. Baayen et al. (2006) reanalyzed a previously published database of lexical decision and word naming latencies for English. As co-predictors, they include an information-theoretic measure of the morphological distributions of simplex English nouns and verbs: the inflectional entropy (e.g., Moscoso del Prado Martín, Kostić, & Baayen, 2004). Inflectional entropy captures the average uncertainty any particular inflectional exponent of a word. Higher entropies reflect stems that tend to occur relatively often across a larger number of exponents; stems with lower entropies tend not to display much morphological variability across tokens. orthographic words with higher inflectional entropies have been observed for adults in production and comprehension (though the effect is much reduced for production). Crucially, they also correlated inflectional entropy with subjective age of acquisition norms. They find that what adults process more efficiently, children learn earlier. High entropy forms have the earliest age of acquisition ratings. This finding demonstrates that children attend to more than how surprising segmental transitions are when learning words. They appear to track contextual variability at more abstract levels of linguistic structure. Common to the two situations is the fact that a higher degree of variability in the local context promotes word learning.

Effects of inflectional entropy have also been observed for languages with much more complex inflectional paradigms. Stoll et al. (2012) report a corpus analysis of child produced and child-directed speech in the Tibeto-Burman language Chintang, spoken by roughly 6,000 speakers in Nepal. Chintang is a strongly polysynthetic language, boasting a massive paradigm of obligatory verbal morphology, complicated by derivation and variable affix ordering. Stoll and colleagues find that where the inflectional entropies of the child productions match those of the adults more precisely, the children produce a greater share of verbs relative to nouns. Hence, children begin to produce more verbs once they have begun to master the distributional properties of the morphological system as deployed by adults.

These findings beg the question: what mechanism lies behind this learning? We know that children are sensitive to probabilistic distributions, but not yet how this distributional information is represented in the mind of the child. Several theories have been proposed. One could account for these findings by appealing to exemplar (or memory-based) models of linguistic knowledge. According to this family of models, children store specific examples of language use. Words, categories, and even syntactic constructions are built gradually through categorization processes that generalize across partially variable, partially stable exemplars. These exemplars may form “clouds” within a hyperdimensional space, organized around item-specific prototypes (e.g., Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Tomasello, 2003). As children generalize further, they develop hierarchies that bind representations at varying degrees of abstraction (e.g., see Goldberg, 1995, for a thorough discussion of how such a system could be organized). Evidence for this explanation comes from several sources. For example, children's early syntactic knowledge tends to be highly item-specific and dependent on input frequency (Tomasello, 1992), which suggests that children tend to use memory of highly frequent input structures as unanalyzed chunks to express complex meanings prior to breaking them down into their component parts. Furthermore, adults are sensitive to the frequencies of multi-word units (Arnon & Snider, 2010). Frequency effects are typically interpreted suggesting the independent status of mental representations. Thus, these frequency effects could arise from activation of specific exemplars, similar to the unanalyzed chunks that children rely on early in development. However, these models still need to explain the nature of the exemplars. As Baayen, Hendrix, & Ramscar (2013) point out, this model requires a massive capacity for memory, as well as high-speed retrieval operations capable of navigating such bloated networks. One way to get around this problem is to reconstrue the chukas as temporally linked patterns of activation within distributed networks. Phase transitions between the activation states of words in sequence could be facilitated by repeated exposure to multiword chunks, without need for an exemplar to be stored. Similar explanations have been offered to account for semantic priming and other psycholinguistic phenomena in adults (e.g., Plaut & Booth, 2000).

Recent computational models have successfully learned to mimic the effects of morphological and syntactic paradigms on adult language processing (e.g., Baayen et al., 2011). Specifically, a two-tier neural network with orthographic input nodes and semantic output nodes formed more stable relationships between word forms and meanings. Words that are disributed more typicallyacross the set of possible morphological exponents, based on the overall frequency of their exponents. The model uses a simple but powerful learning algorithm based on the Rescorla-Wagner equations (Rescorla & Wagner, 1972), which determine cue/outcome associations based on how frequently the cue, and not other cues, occurs with a specific outcome, and not other outcomes. This approach has been dubbed naive discriminative learning, as the model learns to discriminate lexical representations without any knowledge other than sequences of elements. The same style of model has also been able to reproduce phrasal frequency effects (Baayen, et al., 2013), suggesting that it generalizes beyond word-internal structures. Naive discriminative learning therefore provides an economical and plausible explanation for lexical and supra-lexical learning, and one that links acquisition to processing (in line with the statistical learning literature).

For present purposes, both types of models make the same general predictions about child behavior. First, they both predict that the diversity of syntactic contexts should support learning. When children experience words in highly variable contexts, the common points of contact between form and meaning stand out more clearly, allowing the child to form more stable lexical representations. Beyond that, diverse distributions provide more and stronger exemplars for how words should be used syntactically. The more exemplars of verb/structure pairings that children experience, the more likely it is that they have experienced a syntactic frame compatible with their communicative needs in any given situation. This prediction is given as H_1:

H_1: Children are more likely to produce nouns with high-diversity syntactic distributions earlier in their development.

Second, both approaches predict that more prototypical nouns should be produced earlier. Prototypical nouns occur in the set of syntactic contexts in which children will most often experience nouns. Assuming that these prototypes reflect the communicative needs or habits of speakers generally, prototypical distributions better equip the children to integrate nouns into syntactic structures that are most often needed for the encoding of nouns. The denser the network of exemplars, the more support the word receives from the syntactic system. This hypothesis is given as H_2:

H_2: Children are more likely to produce nouns with prototypical syntactic distributions earlier in their development.

The two classes of models do differ in at least one respect. Exemplar-based models have been specifically developed to handle linguistic representations at all levels of abstraction, from sequences of specific sounds or words to fully abstract argument structure constructions (Goldberg, 1995; Langacker, 1987). They assume direct connections between lexical and more abstract representations (Diessel, 2015). By contrast, naive discriminative learning has only been evaluated on surface sequences (i.e., for which meanings are directly related only to orthographic units in the input). In fact, one fo the key design elements of the model architecture is the lack of abstract representations besides surface forms and meaning (Baayen, et al., 2011). If we find support for H_1 and H_2, the naïve discriminative model would have to augment what it allows as input to include syntactic information beyond what is available on the surface.

Current evidence is inconsistent on whether diversity and typicality measure the same or different aspects of language learning. Milin, Filipović-Đurđević, and Moscoso del Prado Martín (2009) find that inflectional typicality swallows the effect of diversity in lexical decision for nouns. However, Linzen et al. (2013) find different electrophysiological signatures of the two variables for entropies taken over the subcategorization frames of verbs. Moreover, they found no effect of diversity on response latencies, while they did observe a typicality effect on behavior. I let the two compete in the present analysis to determine whether syntactic contexts of nouns likewise show a double sensitivity to diversity and typicality in the context of language acquisition.

In what follows, I introduce the measures of syntactic diversity and typicality. A corpus study based on data from a dense longitudinal sample of twelve English-speaking children is reported, and findings are discussed with relation to several current proposals for the nature of distributional statistical learning.

B. Methods

1. Child Data

Data were taken from the Manchester Corpus (Theakston, Lieven, Pine & Rowland, 2001), distributed through the CHILDES database (MacWhinney, 2000). The Manchester

Corpus contains a dense longitudinal sample of twelve middle-class English-speaking children (six girls and six boys) from the areas of Manchester and Nottingham, England. All children were monolingual and the oldest or only children in their respective families. Ages at the beginning of data collection ranged from 1;8.22 to 2;0.25; the earliest mean lengths of utterance (MLUs; i.e., the mean number of words per utterance for a given sample) ranged from 1.06 to 2.27. Children were recorded for one hour every three weeks for one year (with the exceptions of five missed sessions and two half-sessions across the twelve children). Each hour was broken into two 30-minute play periods. During the half hour, children played with their own toys. During the second half hour, the children played with a set of toys provided by the experimenters. The experimenters interacted only minimally with the children and caretakers, meaning that the bulk of the data reflect caretaker-child interactions. The sessions were recorded and transcribed. Predictable pieces of language-based games (e.g., nursery rhymes, songs, and so on) and proper nouns were treated as single units (e.g., *Thomas_Tank_Engine*; *row_row_row_your_boat*).

The Manchester data were selected because they focus on the earliest stages of grammatical development. The children have just begun to produce syntactic utterances but continue to add rapidly to their vocabularies. This range of ages and syntactic abilities is thus well suited to an analysis of how the syntactic information carried by words impacts when children will begin to produce those words. It also provides a strong test of the functional role of syntax. Any effect observed here would mean that children track syntactic information very carefully, even when they have only barely begun to produce multi-word utterances.

The present question concerns how syntactic diversity and atypicality impact lexical acquisition. Specifically, I am interested in when words are acquired for production (I acknowledge that the child surely comprehends more than they can produce themselves; e.g., Benedict, 1979). I treat a noun as having been acquired once the child produces it for the first time. To extract all and only nouns, I first tagged and lemmatized the entire Manchester corpus for part of speech using the English model from *spaCy*, an open-source natural language processing library for python (documentation available at http://spacy.io).[8] Next, I cycled through each file for each child, from the earliest to the latest, and extracted each unique noun lemma relative to what a given child had already produced. The names of the children, as well as the age and MLU at which the lemmas first appeared, were recorded.

2. Estimating syntactic diversity

Children demonstrate some lack of syntactic competence at the ages studied here, at least relative to adult intuitions (e.g., Gleitman, Gleitman, & Shipley, 1972). In other words, their syntactic systems are expected to be incomplete, still under construction. Therefore, they may only be able to take advantage of certain gross generalizations about the syntactic contexts of words. I attempt to accommodate for this imcompleteness by exploring the role of total syntactic diversity and prototypicality. That is, I do not distinguish the full array of syntactic features, which include (at higher levels of abstraction) word order, hierarchy (i.e., head vs. dependent status), and the crossing of those two variables. Such variables have been

[8] The Manchester corpus already contains part of speech annotation. However, in order to maximize comparability across our child and adult data, we derive part-of-speech labels for both samples using a single tagging algorithm.

shown to impact adult language comprehension and production (Lester & Moscoso del Prado Martín, 2016); I leave it to future research to explore their role during the emergence of syntax.

The Manchester data come from children learning British varieties of English. I therefore estimate the prior syntactic distributions of nouns from the British National Corpus (BNC; *The British National Corpus*, 2007). Prior research has focused on American English, and so has used the largest and most well-balanced corpus for that variety which is freely available in its entirety:, the Open American National Corpus (OANC; Reppen, Ide, & Suderman, 2005). The OANC contains approximately 15 million words of text from spoken and written modes, covering many genres, registers, and so on. To improve the comparability of the samples across these studies, estimates of diversity and prototypicality were collected for a random subset of the files totaling approximately 15 million words. This way, any differences in the effects observed between this and prior research on adults cannot be attributed to the size of the sample on which the estimates are based. I parsed this 15-million word sample of the BNC using the *spaCy* dependency parser (Honnibal & Johnson, 2015).

The *spaCy* dependency parser produces a dependency graph for each sentence. These graphs represent the syntactic structure of sentences as a set of binary relationships between pairs of words. Within each pair, one word – the head – is hierarchically superior, while the other – the modifier – depends on the head for its realization. For example, in the noun phrase *the bowl*, *bowl* is the head. It is the semantic core of the phrase; *the bowl* is more about *bowl* than *the*. It is also the syntactic core, in that it determines how the pair of words may be integrated into the broader syntactic frame (e.g., in *The bowl sat on the table, bowl* *c*an be the subject of the verb *sat* because it is a noun; *the* cannot fill this role). The head and modifier are linked by a typed functional relation. For example, in the *spaCy* conventions, the relation that binds *the* to *bowl* in *the bowl* is labeled *det* (for "determiner"). I refer to this triplet of head, modifier, and syntactic relation as a *bundle*.

The syntactic measures are defined using these dependency relations for each bundle in which the nouns occur. I take the syntactic distribution to be the frequency distribution of nouns across the possible relations, either as head or modifier. To measure syntactic atypicality, I create a "syntactic prototype" by summing the frequency distributions across all nouns. The distribution of each noun is then compared to the summed distribution using an information-theoretic measurement known as the Jensen-Shannon divergence (JSD), a symmetrical variant of the Kullback-Leibler divergence (KLD). The KLD from distribution *P* to distribution *T* is given in Eq. 8:

$$KLD(P\|T) = \sum P\mathrm{log}\frac{P}{T} \tag{8}$$

The KLD is asymmetric, meaning that the divergence from *P* to *T* is not necessarily equal to the divergence from *T* to *P*. However, for present purposes, there is no reason to prefer one direction (e.g., from target noun to prototype) to the other (e.g., from prototype to target noun). I get around this issue by using the JSD. The JSD is calculated in two steps. First, the two distributions *P* and *T* are averaged to create a new distribution *M* "midway" between them (i.e., the euclidean midway point). Then, the KLDs are taken from *P* to *M* and from *T* to *M* and averaged together. That way, JSD(*P* || *T*) = JSD(*T* || *P*). Nouns with high

JSD values are those whose distributions least resemble that of the prototype. Nouns with low JSD values have more prototypical distributions. I improve the accuracy of the frequency estimates by applying the James-Stein shrinkage smoother prior to taking the JSD (Hausser & Strimmer, 2009). This step guards against the bias on maximum-likelihood frequency estimates based on samples (i.e., using frequency counts as approximations of true probabilities). The James-Stein technique works best for distributions in which the number of cells is known. JSD requires a common space for all nouns, meaning that all noun distributions must share the same number of cells, and that this number must be equal to the size of the set of dependencies observed for any noun. While the true number of syntactic relations exceeds what I analyze here, I make the simplifying assumption that the set of relations encoded in the *spaCy* parser exhausts the possible dependency types.

I also measure the syntactic diversity of these distributions. However, the diversity of these distributions is affected by at least two distinct sources of information: the relations in which the noun appears and the other words with which the nouns are bundled. Often, the syntactic relation between two words can be read off of the words themselves. For example, knowing that *bowl* is related to *the* leaves only one possible relation: *det.* Furthermore, lexical co-distributions are known to capture non-syntactic information, for example, semantics (Bullinaria & Levy, 2012). Therefore, we need some way to ensure that we are dealing with abstract syntactic information that is decoupled from the surface aspects of the use of the nouns. I opt for an information-theoretic measure known as the conditional entropy. Eqs. 9-13 define the conditional entropy $H(D \mid L)$ of the syntactic distribution D of each noun given its lexical co-distribution L.

$$H(D|L) = H(D,L) \text{-}H(L) \tag{9}$$

where

$$H(D,L) = H(D) + H(L) \text{-}I(D;L) \tag{10}$$

so that *H*(*D* | *L*) reduces to

$$H(D|L) = H(D) \text{-}I(D;L), \tag{11}$$

where

$$H(X) = \text{-}\sum p(X) \log p(X) \tag{12}$$

and

$$I(D;L) = \sum_{d \in D} \sum_{l \in L} \quad p(d,l) \log \frac{p(d,l)}{p(d)p(l)}. \tag{13}$$

The conditional entropy is equivalent to the joint entropy of the two distributions *D* and *L* minus the entropy of *L* (Eq. 9). The joint entropy (Eq. 10) can be rewritten as the sum of the individual entropies of *D* and *L* (Eq. 12) minus the information shared between the two distributions (the mutual information; Eq. 13). The conditional entropy therefore reduces to the information carried by the syntactic distribution minus the mutual information (Eq. 13) shared between the syntactic and the lexical distributions (Eq. 11). Mutual information is similar to KLD in that it measures how well one distribution (i.e., the joint distribution of *D* and *L*) approximates another distribution (i.e., the fully random combination of *D* and *L*). However, unlike KLD, it is symmetrical. With the conditional entropy so defined, I can remove the information specific to the lexical component of the syntactic bundles from the information carried by the dependency relations, while accounting for any information that may be jointly carried by the two distributions.

Again, if we apply these measures to the raw frequency distributions observed in a corpus, we necessarily underestimate their “true” values. Unlike the JSD, we are here dealing with syntactic paradigms of potentially different sizes; we do not need to specify a common space for all nouns as we did before. This difference means that we can apply a different smoother, one better suited to situations in which the number of cells itself may be impacted by the underestimation bias. I select the technique proposed by Chao, Wang, & Jost (2013). This technique has been shown to perform well at correcting entropies based on the distributional profiles of words (Moscoso del Prado Martín, 2016).

I apply these measures of atypicality and diversity to all nouns from the 15-million word subsample of the BNC described above. I then annotate the nouns from the database of first appearances in the Manchester corpus with their corresponding atypicality and diversity scores.

3. Control variables

I further annotate the Manchester data with a number of variables that might influence when a child ventures to produce a word for the first time. These controls include

— word frequency (log)

— emotional valence (how positive or negative the word is)

— arousal (how “exciting” the word is)

— (conceptual) concreteness

— syllabic length

— phonological neighborhood density (PLD20)

Word frequencies were taken from the SUBTLEX-UK corpus (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), a corpus of approximately 200 million words of

British English based on the subtitles of BBC broadcasts. Higher frequency nouns are expected to be produced earlier by children (e.g., Goodman, Dale, & Li, 2008). Frequencies were log-transformed to correct for strong positive skew (words in the highest frequency ranges are few and far between). Emotional valence and arousal ratings were taken from the norming database provided by (Warriner, Kuperman, & Brysbaert, 2013). These norms reflect how emotionally positive and exciting the concepts expressed by English words are, based on the impressions of a large sample of adults. Recent cross-linguistic studies report that children tend to produce positive words earlier than negative words (Braginsky, Yurovsky, Machman, & Frank, 2016; Harmsen, 2017). Weaker effects have been found for arousal, with a slight trend for more exciting words to be learned later (Braginsky et al., 2016). Concreteness norms were extracted from the database published in Brysbaert, Warriner, and Kuperman (2014). Similar to the valence and arousal variables, these norms reflect the intuitions of adults. Prior work has shown strong negative correlations between concreteness and the age of acquisition of nouns: concrete nouns are learned earlier than abstract nouns (Braginsky et al., 2016; Harmsen, 2017). Syllabic lengths were extracted from the CELEX database (Baayen, Piepenrock, & Gulikers, 1995). Weak effects of word length have been observed for English, such that longer words are learned later (Braginsky et al., 2016; Harmsen, 2017; Lewis & Frank, 2016). Finally, I compute a measure of phonological neighborhood density known as PLD20. Phonological neighborhood density refers to the number of words that overlap with the target word in their phonological form. PLD20 operationalizes neighborhood density as the average Levenshtein distance (LD; the smallest number of single-character edits – insertion, deletion, addition, or substitution – to change the target into another word) between the target word and its twenty closest neighbors (words with smallest LD). I compute PLD20 for all words in the sample using the phonological representations from CELEX and the Levenshtein algorithm as implemented in the *vwr* library for *R* (Keuleers, 2013). Children as young as nine months preferentially attend to words with dense as opposed to sparse phonological neighborhoods (Jusczyk, Luce, & Charles-Luce, 1994). Therefore, phonological density supports early word learning. Moreover, research from picture naming suggests that children aged three to five are faster and more accurate when producing names from dense neighborhoods as opposed to words from sparse neighborhoods (Arnold, Conture, & Ohde, 2005). Therefore, words with lower PLD20 (words from dense neighborhoods) should be produced earlier.

Pairwise scatterplots show strong intercorrelation between the variables. In particular, frequency and concreteness were correlated with several other variables each. This situation, known as multicollinearity, creates problems for regression models (Baayen, 2008). A test for multicollinearity revealed that the degree of intercorrelation between the variables is unacceptably high (condition number $\kappa > 47$). To address this issue, I perform two generalized additive regression models (GAMs). First, I predict frequency on the basis of the other variables (allowing for non-linear relationships via spline-based smooth terms, as well as random intercept adjustments per child) and replace the raw frequency variable with the residuals of that model. The residuals reflect the part of frequency that cannot be explained by the other variables, which means that the new residualized frequency measure and the other predictors are fully decorrelated. I then do the same for concreteness, this time leaving out frequency. This process reduced multicollinearity to an acceptable level ($\kappa < 25$; see

Baayen, 2008).

C. Results

I performed a Cox Proportional Hazard Regression (CPHR) predicting the time of first occurrence of nouns in the Manchester corpus. CPHR is useful for modeling the time until some event is realized – for example, the time to death in some population (e.g., patients afflicted with some disease). CPHR predicts changes in the hazard rate, that is, the change in probability of an event occurring at a particular point in time when the predictor increases by one unit. In CPHR, the hazard rate is assumed to be constant across the period of time in which observations are made, which allows one to summarize the effect of a variable of interest with a single value. This assumption is known as the proportional hazard assumption (PHA). Usually, the rate is log-transformed, which centers the variable on 0, such that positive values indicate a higher "morbidity" (earlier observation of the event) and negative values indicate a "protective" quality (later observation). This approach has been fruitfully applied to lexical acquisition in several studies (e.g., Smolík, 2014; Smolík & Kříž, 2015), though the technique is perhaps underused in the field of child language acquisition generally.

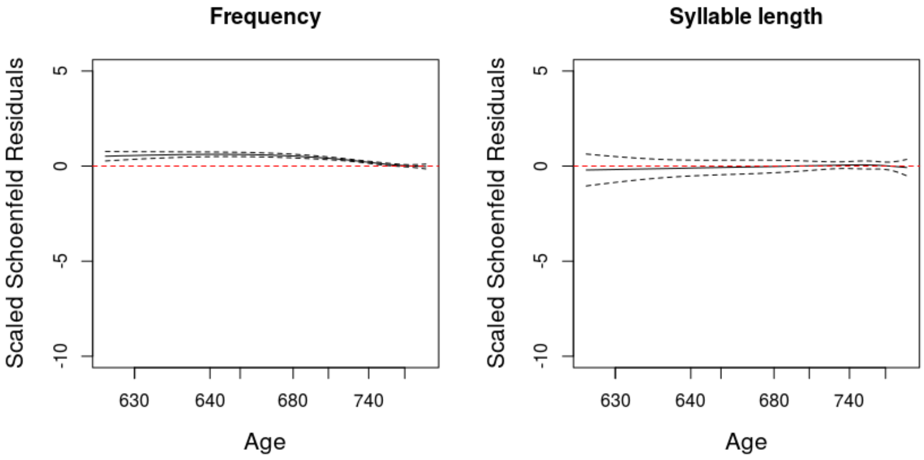
Cox regression is sensitive to a form of sampling bias known as truncation. The data show both left and right (random) truncation. Left truncation refers to the fact that words are only included in this analysis if they have "survived" (i.e., not been produced) until the window of time captured in the Manchester corpus. Right truncation means that words that survive beyond the window of time in Manchester are excluded (right censoring), as well as words that never happen to surface in the present sample (truncation). These truncations are random in the sense that we cannot identify which words we are missing – some children, even as they grow into adults, will never utter certain words that exist in the English language, and the children in the Manchester corpus would certainly have been saying words before the experimenters began collecting data, even if they do not repeat those words in any of the recordings. This point illustrates why these truncations are an unavoidable aspect of applying survival analysis to lexical acquisition data: we cannot define, much less track the entirety of the English lexicon relative to each child. Handling left truncation is simple, and only requires that we relativize the survival function against the earliest age at which the children entered the Manchester study. In the present case, right truncation may be impossible to address. I therefore make the simplifying assumption that words produced after the window captured by the Manchester recordings will follow a similar pattern to those produced within the window (others have implicitly made similar assumptions; e.g., Smolík, 2014). In other words, I assume that words tend to surface earlier in any given sample if the child has actually produced the word before (even if it has not yet been observed in the samples). Many studies that examine variables similar to those studied here rely on age of acquisition norms that have been averaged across children (e.g., the norms that some have derived from the MacArthur-Bates Child Development Inventory; e.g., Goodman, et al., 2008). However, such approaches overlook the crucial individual differences in acquisition. These differences play out in a number of ways, including earlier or later acquisition, but also with respect to precisely which words are successfully acquired within a given time frame. I account for these differences in the present study by including random adjustments to the baseline hazard rate per child. In this way, I directly model the random variability between children instead of computing some measure of central tendency of the enitre group.

I fit the CPHR with the time to first appearance (given the earliest age observed in the sample) as dependent variable. I restrict the analysis to word forms which occur at least 100 times in the BNC sample (~7 per million words) and which are predominantly used as nouns (based on the annotation in Brysbaert et al., 2013; n = 4206). Residualized frequency and concreteness ratings, as well as raw number of syllables, average phonological neighborhood density (PLD20), and adult-based valence and arousal ratings were included as control predictors. I further included the two critical variables, syntactic atypicality and diversity. Finally, I allowed for random intercepts per child[9]. An initial inspection of the model revealed that the several predictors violated the PHA. PHA states that the hazard ratio between words with different values for the various predictors should remain constant over time. If this assumption is violated, the overall hazard coefficient is a mean assessment of a time-evolving variable (Allison, 1995), which can be misleading. For example, given a simple positive linear relationship between time-to-event and the estimates of the hazard coefficient, the overall coefficient will be an underestimate for words that are first produced in the earlier age range and an underestimate in the later age range. Crucially, these under- (or over-)estimates may cross over the null-effect threshold, meaning that observations at either end of the age spectrum would generate opposed estimates of the overall hazard ratio.

Violations of PHA can be handled in several ways. Following Smolík (2014), I stratify

[9] We model random effects using the *frailty* function from the *R* package *survival* (Therneau, 2015).

the age variable into sub-groups and compute separate estimates of the hazard for each range so that no range (or the model as a whole) violates PHA. Smolík did not specify how he selected the age strata, so I follow a simple empirical heuristic. I first plot scaled Schoenfeld residuals against age (Grambusch & Therneau, 1994). Schoenfeld residuals represent the covariate values for each individual that "fails" (words produced for the first time) at time *t* minus the expected covariate value given all individuals at that time. The expected covariate value is the sum of covariate values for all individuals (words) in the hazard set weighted by their likelihood of failure β. These residuals can then be scaled by multiplying by the inverse covariance matrix of β. These scaled residuals can be summed across individuals at each time *t* and plot a smooth curve through the resulting points over time. If the slopes of these lines deviate from zero, then the PHA is violated. I fit nonlinear smooth terms through the scaled Schoenfeld residuals over time for each of the covariates and examined the trends. These smooth terms are plotted in Figure 16.



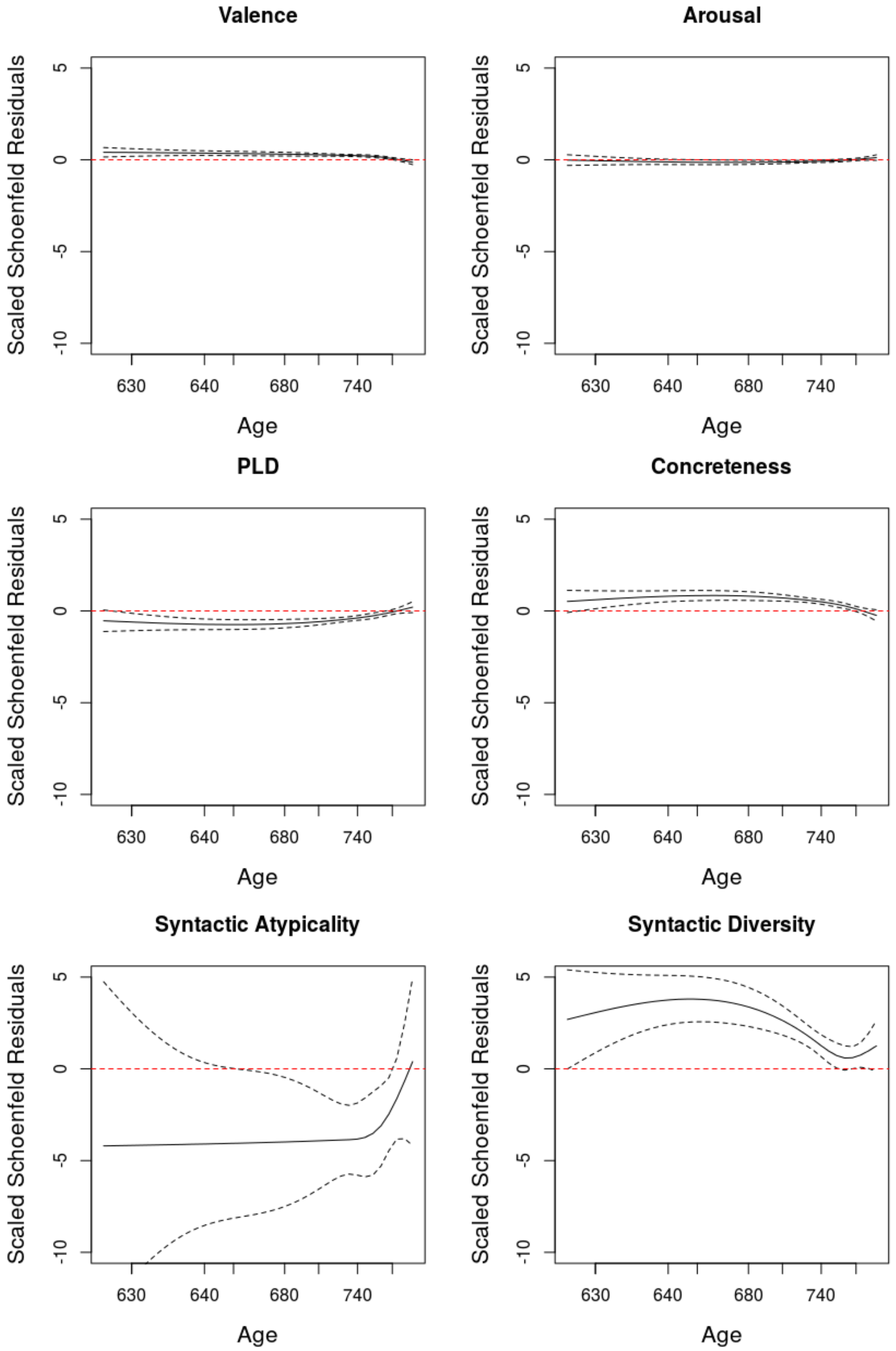


Figure 16: Schoenfeld residuals per predictor variable over time. The y-axis plots the estimated coefficient. The x-axis plots the age of first appearance of words (days).

The curves in Figure 16 reveal unacceptable nonlinear trends for frequency, valence, PLD20, concreteness, and syntactic diversity. These trends were confirmed via two-tailed tests of the correlation between the Schoenfeld residuals and age (all $p < 0.05$). For each variable, I noted the approximate ages at which the unacceptable curvatures appeared. I then carved the time scale based on a compromise between how many variables showed a deflection in a comparable range and how serious the deflection of any single variable was regardless of the behavior of other variables. Using this approach, we can identify the time chunks that align across variables while avoiding chunks that collapse too wide a range of coefficients estimates for any single variable. The four age groups (across all children) are as follows (in days): 626-700, 701-800, 801-950, and 951-1105. Smolík arrives at a very similar set of time chunks: 627-690, 691-800, 801-1000, and 1001-1105. These cutoffs correspond fairly well to the overall density of the first-mention observations, plotted in Figure 17.

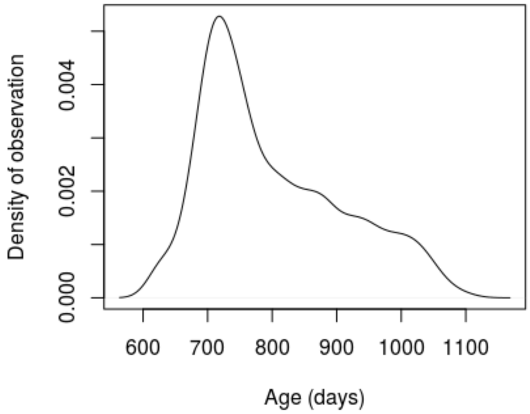


Figure 17: Density of observations for the first appearance of words

Figure 17 shows that the density of observations rises quickly from 626 days to reach a maximum at approximately 750 days, followed by a slight step function to a local minimum at around 950 days, followed by a precipitous drop to 1105 days. We can further see that the second time slice encompasses a period of rapid growth in the vocabulary of the children consistent with the much discussed “lexical burst.” The timing (around 24 months) suggests that this growth corresponds to the “second burst” that accompanies the emergence of morphosyntax (e.g., Bates & Goodman, 1999; Brown, 1973). Chunking time in this way removes all violations of PHA, as confirmed with another set of correlation tests (all $p >$ 0.19).

Results of the stratified model are presented in Table 6. The coefficients of the Cox regression (log-transformed hazard ratios) for each time stratum are given in the columns. The standard errors are given in parentheses.

No variable showed a significant effect in the final time window. This could be due to a lack of power, as very few nouns in this window that had not appeared already (roughly 25 observations per child, respectively). Furthermore, length in syllables showed no effect in any window. This is not surprising given that several studies have repeated only weak or no effect of this variable (e.g., Braginsky et al., 2016).

1. Significant controls

Residualized frequency shows a significant positive coefficient in the first two time chunks, meaning that increases in frequency correlate with increased chances of being produced early but not later in development. For every one unit increase in frequency, the word is approximately 1.57 ($e^{0.45}$) times more likely to be produced at any given time within the first chunk. Within the second time chunk, this effect weakens to 1.30 times more likely, disappearing completely by the third and fourth chunks.

Table 6: Results of stratified Cox proportional hazard regression

Fixed effects	**626-700 (*n*=1009)**	**701-800 (*n*=1935)**	**801-950 (*n*=958)**	**951-1105 (*n*=304)**
frequency	0.45 (0.05) ***	0.26 (0.03) ***	0.01 (0.04)	-0.01 (0.05)
length (syl.)	-0.17 (0.19)	0.07 (0.10)	0.05 (0.12)	0.10 (0.18)
PLD	-0.67 (0.15) ***	-0.40 (0.07) ***	-0.19 (0.08) *	0.04 (0.11)
valence	0.28 (0.06) ***	0.29 (0.04) ***	0.05 (0.05)	0.04 (0.05)
arousal	-0.09 (0.07)	-0.13 (0.04) ***	0.02 (0.04)	0.06 (0.06)
concreteness	0.67 (0.16) ***	0.64 (0.09) ***	0.17 (0.08) *	0.08 (0.10)
atypicality	-2.67 (2.09)	-3.48 (1.16) **	-2.69 (1.35) *	-0.53 (1.86)
diversity	2.69 (0.65) ***	1.77 (0.35) ***	0.55 (0.38)	0.90 (0.54)

Random effect	**Variance**
child	.32***

*p < .05. **p < .01. ***p < .001

Phonological neighborhood density was a significant predictor of initial use in the first three time chunks. In all chunks, the effect was negative: words from sparse neighborhoods tend to surface later than words from dense neighborhoods. In the first chunk, every one unit increase in PLD20 led to a 49% ($1\text{-}e^{-0.67}$) reduction in the chances of being produced for the first time. This effect weakens to approximately a 33% reduction in likelihood for the second time chunk and a 17% reduction in the third time chunk.

Valence was associated with earlier productions. Higher valence scores translate into more positive emotional content. In the first and second time chunks, words were 1.32 and

1.33 times more likely to be mentioned per unit increase, respectively. No effect was observed at later time chunks.

Arousal also surfaced as significant, but only in the second time chunk. Increasing arousal ratings reflect more exciting words, meaning that more exciting words were produced later. Each unit increase results in a 12% reduction in the odds of a word appearing in this window.

As expected, concreteness supports early production. More concrete words tend to appear earlier in each of the first three time slices. The effect is roughly equivalent in the first two time slices: words were 1.95 and 1.90 times more likely to appear per unit increase, respectively. This effect weakened in the third slice to a per-unit increase in likelihood of appearance of 1.19 times.

2. Critical predictors

Both of the critical predictors surfaced as significant: diversity promotes early production, while atypicality promotes later production. We therefore find support for both H_1 and H_2. Notably, they exert their effects at a slight offset, such that the diversity effect precedes but overlaps with that of atypicality. Furthermore, both variables have effect sizes more than double (in the log scale) those of the other predictors. The diversity effect begins early. In the first time slice, more diverse words were 14.73 times more likely per unit of diversity to be produced. By the second time slice, we see a large drop in the effect size to a factor of 5.87. Atypicality surfaces for the first time in the second period, greatly reducing chances of first production (97% per unit increase). The effect continues into the third time slice, diminishing slightly to decrease chances by 93%. While the effects are somewhat stronger, they also come with much larger confidence intervals (time slice two: {.71, .99}; time slice three: {.05, .99}).

D. Discussion

Statistical learning during early lexical acquisition depends on a host of factors across several layers of linguistic organization. Phonological and prosodic factors drive early word segmentation (e.g., Saffran et al., 1996). One layer up, morphological variability supports children's productive use of word classes (e.g.,Stoll et al., 2012). The present study shows that children likewise attend to variability at the level of syntactic distributions. Moreover, this distributional information is tracked at multiple levels: both individually (diversity per word) and paradigmatically (distribution measured against other words). Importantly, we observe these novel effects while simultaneously replicating previously reported effects for word frequency, emotional valence, arousal, concreteness, and phonological neighborhood density. The fact that the results on all these variables match the previously reported literature adds credence to the reliability of the novel analysis methods.

First, consider syntactic diversity. From the approximate ages of 1;8 to 2;2, words that occur more frequently in a wider array of syntactic relations are produced earlier than more syntactically constrained words. Crucially, this effect is independent of the specific words that manifest the abstract syntactic relations. Prior evidence suggests that children can learn syntactic constructions by generalizing over fully lexically specified exemplars (Goldberg, 2006; Tomasello, 1992). The results presented here extend this research: even at the earliest stages of syntactic acquisition, children appear to exploit abstract syntactic knowledge for purposes of word learning. This finding is consistent with the experimental evidence from infants (Gertner, et al., 2006; Lidz, et al., 2003; Lidz et al., 2017) and young children (Shimpi, Gámez, Huttenlocher, & Vasilyeva, 2007; Thothathiri & Snedeker, 2008; cf. Savage, Lieven, Theakston, & Tomasello, 2003). Going beyond these studies, I find that more variable syntactic contexts solidify children's lexical representations, irrespective of the individual syntactic functions in which they are observed. Importantly, I do not mean to imply that the contextualized functions of nouns within particular syntactic relationships do not play a role. Rather, I wish to convey that the syntactically constrained word learning outlined by, for example, Lidz et al. (2017) , could be strengthened when it applies across many different constructions simultaneously.

Second, children's early syntactic knowledge is paradigmatically organized relative to word class. Children aged 1;10 to 2;7 learn nouns earlier when the syntactic cues to their use overlap with those of over nouns. The measure of typicality can be interpreted as reflecting the syntactic density of nouns within the noun category. Density effects in the same direction have been uncovered for other linguistic domains in child language acquisition. For example, infants prefer to attend to words from dense phonological neighborhoods (Jusczyk et al., 1994), and older children process such words faster and with fewer errors (Arnold et al., 2004). Moreover, adults are faster at recognizing nouns from typical orthographic/phonological distributions (Ferrand, et al., 2011; Yarkoni, Balota, & Yap, 2008), as well as morphological and collocational distributions (Baayen, et al., 2011). One possible explanation for this effect is that the typical distribution serves as the baseline for processing (e.g., Linzen et al., 2013). Adjusting one's expectations to handle unusual nouns comes at a cognitive cost (e.g., Plaut & Booth, 2000). For children, this cost could be prohibitive, delaying acquisition for production. Another explanation is that dense neighborhoods produce “gang effects,” whereby a cluster of closely related words support recognition of the target through sympathetic activation or mutual inhibition of non-targets (e.g., McClelland & Rummelhart, 1981).

The typicality effect held over and above that of syntactic diversity, lending further support to the notion that diversity and typicality correspond to independent dimensions of language representation (e.g., Linzen et al., 2013). An unexpected finding concerned the fact that the diversity effect precedes the typicality effect. While this finding should be treated with caution, it suggests that children are sensitive to the syntactic distributions of single words before they are influenced by the distributions of nouns as a class. If found to be reliable, this effect would fit well with both exemplar-based models of language (e.g., Bybee, 2010; Diessel, 2015) and distributed-activation models (e.g., Plaut & Booth, 2000). First, the child becomes aware of the syntactic information carried by individual words. Over time, this experience builds up both within and across words, producing a form of prototype (whether built explicitly from exemplars or “burned” into an interactive-activation network). This explains the offset, and could be first evidence of how syntactic paradigms are established. These results are more difficult to account for in the context of a naïve discrimination model. Certainly, discrimination learning can account for distributional effects. However, it is not clear how it could explain the independent effects of diversity and atypicality, nor the temporal offset in the effects. Milin et al. (2009) find that typicality trumps diversity in adult processing for morphological paradigms. Perhaps the offset uncovered here has captured an in-process shift in the representation of nouns – one that moves from diversity to prototypes, and which endures into adulthood. Future research should further investigate the time course of the development of the two effects in early childhood, as well as the relationship between the developmental trajectory and adult performance.

Finally, two points regarding methodology. First, this study joins a handful of others which apply multifactorial methods to naturalistic child production data (e.g., Braginsky et al, 2016; Harmsen, 2017). By studying naturalistic production, I sacrifice the level of control achieved in laboratory experiments, but drastically increase the ecological validity of the study. By including many predictors, I accomplish two things: I maintain a high degree of statistical control of the analysis, and I compare the relative importance of different variables from many domains of linguistic representation. More importantly, the latter point allows us to test to what extent the variables of interest – syntactic diversity and atypicality – are independent of other types of information (for an example of this, see Moscoso del Prado Martín, 2007, who finds that erstwhile distributional effects of "morphology" might in fact reduce to semantics). Second, I employ an underused regression technique – Cox Proportional Hazard regression – that is well suited to studying the emergence of vocabulary in child speech. A particular advantage of this approach is that no the most critical variable, age, need not be transformed prior to modeling. Moreover, one can avoid the issues that come with substituting other measures for age, such as mean length of utterance. Such substitutions have often been deemed necessary to capture individual variation in the developmental trajectories of children; however, they create several problems for the analysis of longitudinal corpus data (e.g., Gries & Stoll, 2009). CPHR can naturally accommodate such variability through random effects. It therefore maximizes interpretability while minimizing common issues in the corpus-based analysis of lexical acquisition.

This study shows that syntactic distributions play a strong role in supporting early acquisition. The results further suggest a temporal relationship between syntactic diversity and typicality. The child starts out by discriminating words via diverse distributions. Soon after, they accumulate standard expectations for the syntactic behavior of word classes, which further support productive use of new vocabulary. These findings provide grist to the mill for research on early lexical acquisition, statistical learning, and the syntax-lexis interface.

V. General Discussion

A large body of research has established that the distributional properties of language use shape lexical production, comprehension, and acquisition at multiple levels of linguistic analysis (to name but a few: Baayen, 2010; Baayen et al., 2006; Baayen et al., 2011; Hendrix, et al., 2016; Jusczyk, et al., 1994; Kostić et al., 2003; Lester & Moscoso del Prado Martín, 2015, 2016; Lester et al., 2017; Lidz et al., 2017; Linzen et al., 2013; McDonald & Shillcock, 2001a,b; Milin et al., 2009; Mintz et al., 2017; Moscoso del Prado Martín et al., 2004; Newport, 2016; Saffran et al., 1996; Storkel, 2004). Over the last fifteen years, a promising new approach to the measurement of these distributions has been developed (e.g., Kostić et al., 2003; Moscoso del Prado Martín et al., 2004; Milin, et al., 2009) based on information theory (Shannon, 1948; for a technical reference, see Cover & Thomas, 1991; for a more accessible introduction, see Stone, 2015). Originally applied to morphology, these information-theoretic measures have recently been extended to analyze syntactic distributions (Baayen et al., 2011; Hendrix et al, 2016; Linzen et al., 2013). However, no standard measure has yet emerged. One approach measures the typicality (relative entropy) of the distribution of prepositions that co-occur with nouns within the prepositional phrase construction (e.g., Baayen et al., 2011). Linzen and colleagues take a different approach. They estimate the frequency distributions of different argument-structure (or sub-categorization) frames for verbs. Using these distributions, they compute measures of both diversity (entropy) and prototypicality (relative entropy). Importantly, the two approaches yield different results in adult behavior: the prepositional measure correlates significantly with response times in several tasks, but the constructional measures do not. In short, we have no standard measure of syntactic diversity, and no consensus on the relationship between syntactic diversity and behavioral response.

Beyond the lack of consensus, the approaches outlined above each come with their own set of issues. The prepositional relative entropy has limited application, in that it can only be applied to nouns. Second, it also only measures typicality of the nouns in a syntactically subordinate role (as objects of prepositions), and only for left-facing syntactic relationships. Third, this measure is not contrasted with a comparable measure of the diversity of the noun distributions; thus, we cannot be sure whether or to what extent typicality impacts processing independently of diversity. Fourth, the entropies are based on maximum-likelihood estimates of the probabilities (i.e., based on the raw frequencies as observed in a corpus). Entropies based on maximum-likelihood estimates are negatively biased (Miller, 1955), which reduces their reliability, hence interpretability. Finally, and more importantly, the prepositional relative entropy measures purely lexical co-occurrence in a small-scale co-occurrence window (prep + determiner + noun trigrams). Such co-occurrence windows are known to capture semantics (Bullinaria & Levy, 2012; McDonald & Shillcock, 2001a,b; Moscoso del Prado Martín, 2007), which draws the syntactic interpretation of the effect into question.

The constructional measures successfully avoid the semantic confound. They also allow for typicality and diversity to be compared directly. Indeed, this comparison turned out to be necessary, as the two measures impacted the electrophysiological signature differently. However, these measures only account for the frames projected by the verb, that is, syntactic relationships for which the verb is the head. Secondly, they take a holistic approach to argument structure, such that each syntactic type represents a complete subcategorization frame (e.g., for a simple transitive verb: VERB <subject, object>). Therefore, they do not directly capture the ordering of component syntactic relationships relative to the verb.

Finally, as with the prepositional measure, the constructional measures are based on maximum-likelihood probability estimates, and so suffer from the same problem of reliability.

A principle goal of this dissertation has been to introduce a set of measures capable of addressing all of these issues (see Appendix for a thorough discussion of the database). The measures I proposed are based on binary, asymmetric syntactic dependencies. Such dependencies directly distinguish syntactic relations on the two critical dimensions that are obscured in the previously proposed measures: hierarchical status (heads vs. modifiers) and word order (leftward vs. rightward facing dependencies). Second, the measures, like those proposed by Linzen and colleagues, are based on fully abstract syntactic relations. For the measures of diversity, I go one step further, explicitly removing information carried by the lexical content that fills out the abstract syntactic relations. Third, I carefully control for underestimation bias by smoothing the probability estimates on which the entropies are based. Taking this step improves the reliability of the estimates. Finally, although I only compute the measures for nouns, I have designed them such that, in principle, they may be extended to any lexical category. Moreover, although the present measures are rather low-level, they may be scaled up to approximate the types of constructional measures applied in Linzen et al. (2013). For example, one could count the arrays of dependencies projected by the target word as single units (e.g., the noun phrase *the stealthy owl* could be counted as an instance of OWL <*det, amod*>).

This list of improvements proved to be necessary. The effects of prior syntactic distributions were differentiated by hierarchical status and word order in both comprehension and production. Moreover, the improved measures uncovered purely syntactic effects where they had not been detected before. In what follows, I review these findings across the different tasks. I compare the structure of the syntactic effects to see what they reveal about the nature of the syntax-lexis interface in word processing. I then comment on the possible genesis of these effects in distributional statistical learning based on the findings from child language acquisition.

A. Comprehension

Chapter 2 reported re-analyses of two previously published databases of visual lexical decision, one simple (single lexical items) and one primed via overt lexical priming. Previous work linking prior constructional distributions to lexical decision latencies found no effect of either diversity or prototypicality on response times. However, it did uncover reliable and independent effects of the two measures in the electrophysiological signature. The dependency-based measures revealed significant effects on response times taken from the English Lexicon Project lexical decision mega-study (Balota et al., 2007) for diversity and typicality. This study therefore replicates the independence of the two effects, while extending it into the domain of behavior. Speed of recognition indeed depends on information drawn from fully abstract syntactic distributions. These effects held over and above a number controls from several linguistic domains (orthography, semantics, and frequency) known to influence response times, providing further support for their veracity.

The diversity and typicality effects showed different sensitivity to hierarchy and word order. Typicality played out uniformly across heads and modifiers, whether facing to the right or to the left. The role of diversity varied across the types of distributions: rightward diversity (particularly as-modifier diversity) was inhibitory, while leftward diversity, regardless of hierarchical status, was facilitatory.

In a follow-up study, I sought to cement the reality of the dependency relations underlying the distributional effects. To this end, I measured the overall syntactic similarity between nouns in the dependency space for all noun-noun prime-target pairs in the Semantic Priming Project mega-study database (Hutchison, et al., 2013). I made sure to clean any shared semantics between the nouns from this measure. I then correlated the syntactic similarities with response latencies for the target nouns. Results showed that syntactic similarity between prime and target facilitated recognition. Therefore, following the typical interpretation of priming effects (e.g., Branigan & Pickering, 2017), the dependencies which underly the distributional effects observed in the simple lexical decision task should have some form of representation (discrete or distributional; for the latter, see Plaut and Booth, 2000) and shared across nouns. I interpret these findings as evidence that the information carried by syntactic relationships is not simply an directly surface-driven side-effect of distributional learning (e.g., Baayen et al., 2011); instead, it suggests that adults have formed stochastic syntactic generalizations (which might indeed originate in distributional learning), and that these syntactic generalizations are intimately bound to lexical representations.

B. Production

The prepositional relative entropy has been tested in previous research using a hybrid comprehension/production task. In this task, a phrasal context was presented visually prior to an image of an object, whose name would complete the phrase (e.g., *in the* followed by a picture of a bucket). The participants then named the image aloud. Response times were faster for names with lower prepositional relative entropy (i.e., names that were more typically distributed across prepositional contexts). Crucially, this task involves a semi-predictable syntactic context, so that finding an effect of syntactic distributions might be expected a priori. No study to my knowledge has yet examined the role of these distributional measures in isolated word production. Moreover, no study has examined such effects on the basis of purely syntactic distributions (recall the lexical, hence semantic, confound of the prepositional relative entropy).

I address this gap in the literature with two picture-naming studies reported in Chapter 3. First, I asked participants to name images of concrete objects using isolated nouns (e.g., *banjo!*). Controlling for a number of conceptual and lexical factors, I find no effect of syntactic diversity. The reduced influence of diversity fits with prior work. Other distributional measures have also shown weak or null effects in production (e.g., Tabak et al., 2010, for inflectional entropy), and studies of syntactic gender (the "gender congruence effect") have found no effects in bare-noun naming using picture-word interference (La Heij et al., 1998). I did find a weak but significant effect of syntactic typicality. Therefore, isolated word production shows generally weaker effects of syntax than comprehension. Moreover, this effect was conditioned on the ordering of the dependency relation: only rightward dependencies played a role, irrespective of hierarchy. This rightward specialization could be explained in several ways. For example, it could be a result of the task. Participants were required to answer as quickly as possible, and produced many names one after the other. Nouns that align in their ability to open syntactic doors for upcoming words may facilitate rapid progress through the task – a sort of syntactically mediated, future-oriented priming from trial to trial. This could also be part of a general processing response to time pressure. Research on sentence production reports increased incrementality under time pressure (e.g., Ferreira & Swets, 2002), which could surface as increased sensitivity to the rightward-facing syntactic dependency space. Speculation aside, the finding demonstrates that purely syntactic distributions can affect isolated noun production, and that these effects should at least take account of word order.

I hypothesized that stronger effects might emerge if participants were required to name the pictures using a syntactic frame. I ran a second experiment, identical to the first, except that the participants named the pictures with a noun phrase of the form *the* + NAME (e.g., *the banjo!*). Indeed, both syntactic diversity and typicality surfaced as significant predictors at the onset of the determiner (only diversity predicted onsets of the noun within the noun phrase).

The determiner was produced earlier for nouns with generally more typical distributions. This effect deserves further exploration. I suggested that it could arise from a task-specific strategy whereby the participant waits to produce the determiner until a critical mass of activation builds up within the lexical network. The faster this activation builds up (e.g., via widespread sympathetic activation within a densely populated corner of the syntactic space), the more certain the participant that a lexical item will be available to produce.

For nouns with more diverse modifier and leftward distributions, *the* and the noun were produced earlier. For nouns that served as more diverse rightward heads, *the* and the noun were produced later. This effect could have several sources. On one analysis of the determiner relation, the noun is the head. Such is the case in the CLEAR labels on which we based the syntactic distributions. Taking this perspective, diverse leftward heads are produced faster in contexts in which they are indeed leftward head – a sort of symmetry effect. But total leftward diversity also played a role. Nouns are commonly elaborated by leftward content. For example, both determiners and adjectives typically precede nouns. These relationships could play a more critical role in carving out the links between form and meaning (see Baayen, 2010). Modifiership was also important. Nouns are frequently modifiers of a number of critical structures, including verbs (as subjects or objects) and prepositions. Further, some theories argue that the syntactic frames into which nouns are fit dictate their interpretation (e.g., Borer, 2005). Again, the critical finding at this stage is that syntactic distributions do not exert monolithic effects on processing; instead, they are sensitive to multiple dimensions of syntactic structure, including hierarchy and word order.

C. Comparing the effects of diversity in production and comprehension

The overall shapes of the significant diversity components from the lexical decision and naming studies are similar. For example, leftward diversity plays the same facilitative role in both tasks (with a possibly reduced contribution of leftward as-modifier diversity for naming. This could mean that the lexical representations of nouns are in general better discriminated from preceding context: one learns better when one can compare a token to the incremental expectations generated immediately prior. Such a pattern of effects could be supported by standard error-based learning (e.g., Fine & Jaeger, 2013), and fits well with prior work. For example, Baayen (2010) finds that more diverse lexical contexts to the left of a target word correlate with faster reaction times. Similar interpretations apply to the prepositional measures of Baayen et al. (2011) and (largely) prepositional measures of Lester and Moscoso del Prado Martín (2015; prepositions by definition stand to the left of the noun).

Despite the similarity of the comprehension and production effects, they do show three main differences: the loadings of general modifiers, rightward as-head and rightward as-modifier distributions are reversed, leading to opposite effects on response times. The difference in relationships means the following: (a) general modifier diversity inhibits lexical decision but facilitates naming; (b) rightward as-modifier diversity inhibits lexical decision but facilitates naming; (c) rightward as-head diversity facilitates lexical decision but inhibits naming. Points (a) and (b) suggest inhibition in lexical decision for nouns whose total as-modifier diversity mainly comes from the rightward direction. Inspection of the rightward as-modifier distributions shows that the categories with the greatest frequencies in the rightward as-modifier direction are active and passive sentential subjects. Perhaps the general indeterminacy of the subject role in English, which can assume many thematic functions relative to the verb, leads to a processing response that slows access until more information (i.e., from the main verb) becomes available. Increased indeterminacy at the syntactic level could exacerbate the conservativism of this approach. This "slowing of the clock" could burn into the system to produce slower responses to nouns in isolation.

The facilitation of rightward as-head diversity in comprehension may reflect the inverse of the rightward as-modifier diversity. I have proposed that the latter arises from a conservative processing strategy, based on the naturalistic standard of processing in context, which states, "when nouns project a more uncertain array of possible integrations with upcoming heads. suppress access until more information becomes available." Apparently, the strategy reverses for nouns that show greater uncertainty for how upcoming material may be bound to them. Under these circumstances, the noun is accessed more quickly, possibly to support the rapid integration of upcoming modifiers given expectations that the noun will be further elaborated. This situation mirrors that of the as-modifier distributions in that both prioritize the head as a means of settling the parse. Again, through repeated experience with contextualized processing, these strategies may tune the links between nouns and syntactic nodes to produce the observed effects in isolated and minimally syntactic contexts. The effect reverses in picture naming, which could be due to challenges specific to production, namely, the fact that speakers must commit to only one out of several possible continuations at each “choice point” in the utterance (e.g., Jaeger, 2010; Kuperman & Bresnan, 2012). One such choice point would be the decision of whether to elaborate a noun phrase with additional rightward structures (e.g., relative clauses or prepositional phrases). Some evidence suggests that when speakers face more choices for how to encode an utterance, they take longer to initiate the utterance (Hwang & Kaiser, 2014; Myachykov, Scheepers, Garrod, Thompson, & Fedorova, 2013; but cf. Ferreira, 1996). This effect has been attributed to planning: more choices require more careful, hence longer planning latencies. Perhaps this effect also plays out at the lexical level, such that words that introduce more possibilities for structural elaboration require more careful planning. Importantly, this effect would have to hold even when no such continuation is pursued. Interestingly, the effect was only observed for the syntactic naming task, suggesting that planning-oriented syntactic effects requires that the speaker intends to produce an overtly syntactic structure.

The above discussion demonstrates a deeper level of complexity in the nature of syntactic-distributional effects than has previously been proposed. The specific details regarding how each layer of syntactic organization impacts processing, and how these layers relate to one another remains unclear. I have proposed several possible explanations for the effects observed in the studies presented here, but ultimately future research is needed to unravel this complex tapestry of relationships. Nevertheless, several general conclusions are warranted. First, adults are highly sensitive to the purely syntactic distributions of words, even after information attributable to other distributional sources has been carefully stripped away (cf. Baayen et al., 2011). Moreover, this sensitivity surfaces in behavior, that is, in response latencies (cf. Linzen et al., 2013). Second, the behavioral response to syntactic distributions is differentiated according to hierarchy and word order. Finally, the shape of the response depends on the task, specifically whether one is reading or speaking, and whether the word is processed in isolation or within a syntactic context. Properly controlled (e.g., within carefully orthogonalized experimental designs), these measures have great potential for illuminating how experience shapes lexical representation and processing.

D. Word Learning in Children

In Chapter 4, I explored the genesis of the syntactic effects that were observed in adult lexical processing. Most research on distributional effects in adults assume that they arise during language learning, either through the accumulation of exemplars in memory (e.g., Bybee, 2010; Goldberg, 2006), “burnt-in” patterns of activation (e.g., Plaut & Booth, 2000) or through discrimination learning (e.g., Baayen et al., 2011; see Rescorla and Wagner, 1972). I therefore expected that the first appearance of words in child speech would be supported by the diversity and typicality of the syntactic relationships in which they appear. Some evidence from the inflectional morphology suggests that more diverse inflectional distributions lead to earlier acquisition. For example, Baayen, Feldman, and Schreuder (2006) find a negative correlation between inflectional entropy and subjective age-of-acquisition ratings. When children begin to master these inflectional distributions, they also begin to produce more tokens of words that belong to those paradigms (Stoll et al., 2012). I hypothesized that similar learning mechanisms would apply at the syntactic level, such that more diverse distributions would support earlier acquisition. I further predicted that more typical syntactic distributions would support earlier acquisition. Nouns that meet the expectations for the syntactic behavior the class as a whole should (a) be experienced more often in the environments that best fit the communicative needs of speakers and (b) allow for analogical extension into novel environments based on the behavior of other nouns. Point (b) receives additional empirical support from the results on priming in adult lexical decision. As a first step, I tested these predictions using the overall syntactic distributions, ignoring hierarchy and word order. Results confirmed the hypotheses, but revealed an unexpected temporal offset. The diversity and typicality effects overlapped in time, but diversity preceded typicality, and typicality extended beyond diversity. Thus, nouns are first learned through repeated exposure in diverse contexts. Given a certain threshold of experience, class-wide expectations begin to emerge, something like the accumulation of a Bayesian prior. Further research is needed to see what role, if any, hierarchy and word order may play at these early stages. Another open question is whether older children (e.g., 4-6 year olds) respond to the same syntactic dimensions as adults during online lexical processing (e.g., in an auditory lexical decision task).

E. Conclusions

Taken as a whole, the results reported here provide crucial evidence connecting the behavior of adults to its source in child language learning. They also mirror those observed for inflectional morphology in both adults and children, suggesting a general mechanism for all types of grammatical processing, whether morphological or syntactic. A possible candidate is implicit learning, which has so far been invoked to account for isolated word processing (Baayen et al., 2011) and syntactic priming (e.g., Chang, Dell, & Bock, 2006). However, in the latter case, lexical priming has been argued to manifest via a separate mechanism, namely explicit memory. Evidence for this difference comes from the fact that lexical priming is short-lived, while syntactic priming lasts much longer. However, the results reported here suggest that lexical acquisition and priming both depend on stochastically weighted, implicit relationships between words and syntax. Thus, lexical priming may at least in part depend on the same learning mechanisms that produce structural priming effects.

The primary take-away from this series of studies is that lexicon and syntax are intimately connected. These relationships are direct, probabilistic, and interactive, such that even when syntactic processing is precluded through carefully controlled experimental conditions, it nevertheless guides comprehension and production of words. These findings draw the well-established division between syntax and lexicon into question. It seems that words carry with them their entire history of use across syntactic environments. Moreover, they hint that syntactic structures behave more as timing mechanisms for the realization of words rather than abstract scaffolds into which arguments are slotted. This conclusion comes from the fact that different types of syntactic associations may facilitate or inhibit lexical access. Thus, while structural generalizations have strong support in the theoretical (e.g., Goldberg, 1995) and experimental literature (e.g., Branigan & Pickering, 2017), the functional characterization of these generalizations may need to be revised to arrive at a proper understanding of how they arise and how they are implemented online. At this point, these ideas remain purely speculative. However, given that other aspects of language production have recently been described in terms of mental clocks known as oscillators (Nam, Goldstein, & Saltzman, 2009). Furthermore, neurophysiological work has revealed timing-based entrainment affects that are sensitive to syntactic structure (Ding, Melloni, Zhang, Tian, & Poeppel, 2016). Perhaps the measures introduced here tap into distinct mechanisms for tracking/accessing words and integrating/building them into structures during comprehension and production respectively. At the very least, these measures constitute the most sophisticated and fine-grained analysis of prior syntactic distributions put forth so far.

F. Limitations

My measure of diversity explicitly discounted the role of lexical information. However, the measure of atypicality did not. Instead, the latter was based purely on the distribution of nouns across syntactic relations, irrespective of the attendant lexical context. Information theory does not provide straightforward means for decoupling the two sources of information when measuring distance between two distributions. This means that the independence of the diversity and atypicality effects that were observe here could actually be driven by the lexical information bound up in the latter. One possibility would be to compute independent JSDs for the purely lexical distributions, or $\mathrm{JSD}(L_{\mathrm{Target}} \| L_{\mathrm{Prototype}})$, and the joint lexico-syntactic distributions, or $\mathrm{JSD}(LD_{\mathrm{Target}} \,||\, LD_{\mathrm{Prototype}})$. One could then residualize the former out of the latter in a manner similar to the present treatment of frequency and concreteness.

A more general point concerns the size of the corpora used to estimate these measures. I have relied on the largest and best balanced corpus of American English that is freely available. However, this corpus is only 15 million words. While other distributional studies have achieved strong and replicable results based on even smaller samples (e.g., the spoken component of the British National Corpus, which comprises only 10 million words; McDonald & Shillcock, 2001a), the reliability of the estimates will necessarily improve given a larger sample (e.g., the 250-million-word British National Corpus). Crucially, these estimates should only be computed on dialects that are appropriate for the subjects under study.

Another important next step will be to extend this analysis to other word classes, in particular, verbs and adjectives. Statistical learning accounts predict that the same general learning mechanisms should apply to all word classes. Finding similar effects across word classes would provide strong evidence for the generality of these mechanisms. Further, these mechanisms should apply to broader linguistic structures, as well, such as argument structure constructions (e.g., Goldberg, 2006). One could invert the measures deployed here to test how aggregate lexical variability within constructions impacts the ease of acquisition of these argument frames. However, the specific details about how to implement such measures have yet to be worked out. For example, should only the immediate children of the root node be counted, or should grandchildren, great grandchildren, and so on also be counted? The answers to such questions will provide a more fine-grained perspective than the typical verb-oriented approach to constructional learning (e.g., Tomasello, 1992). Positive results will provide further empirical support for models of language that assume co-representation of lexical and syntactic information (e.g., Diessel, 2015).

Finally, the shapes of the effects in comprehension and production are complex.

Therefore, future studies should orthogonalize the different distributional dimensions when selecting stimuli (e.g., words that have high leftward diversity but low rightward diversity, and so on). Careful selection of stimuli could provide deeper insights into which of the dimensions contribute most strongly to the effects observed here. The nature of these effects so distilled should illuminate the types of information that structure the language processing architecture. Put simply, I expect the general approach followed here to provide a scalable and fruitful diagnostic for exploring several crucial debates in language acquisition, processing, and representation.

References

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

2. Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23,* 275-290.

3. Abney, S. P. (1987). The *English Noun Phrase in its Sentential Aspect*. Unpublished Doctoral Dissertation, MIT.

4. Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A., J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., & Davis, C. J. (2014). A behavioral database for masked form priming. *Behavioral Research Methods*, *46,* 1052-1067.

5. Allison, P. D. (1995). *Survival Analysis Using the SAS System: A Practical Guide*. SAS Institute, Cary NC.

6. Allum, P. H., & Wheeldon, L. R. (2007). Planning scope in spoken sentence production: The role of grammatical units. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 791-810.

7. Almeida, J., Knobel, M., Finkbeiner, M., & Caramazza, A. (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review*, *14*, 1177-1182.

8. Arnold, H. S., Conture, E. G., & Ohde, R. N. (2005). Phonological neighborhood density in the picture naming of young children who stutter: Preliminary study. *Journal of Fluency Disorders*, *30*, 125-148.

9. Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science*, *15*, 578-582.

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

11. Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *CELEX2*. Philadelphia: Linguistic Data Consortium.

12. Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using* R. Cambridge: Cambridge University Press.

13. Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*, 436-461.

14. Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*, 290-313.

15. Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naïve discriminative learning. *Language and Speech, 56*, 329-347.

16. Baayen, R. H., Levelt, W. M. J., Schreuder, R., & Ernestus, M. (2008). Paradigmatic structure in speech production. In: M. Elliott, J. Kirby, O. Sawada, E. Staraki, & S. Yoon (eds.), *Proceedings of the Chicago Linguistics Society 43, Volume 1: The Main Session* (pp. 1-29). Chicago: University of Chicago Press.

17. Baayen, R. H., Milin, P., Filipović-Đurđević, D., Hendrix, P. & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118,* 438-482.

18. Baayen, R. H., Wurm, H. L., and Aycock, J. (2007). Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419-463.

19. Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283-316.

20. Balota, D. A., Paul, S., & Spieler, D. H. (1999). Attentional control of lexical processing pathways during word recognition and reading. In S. Garrod & M. Pickering (Eds.), *Language Processing*, (pp. 15-57). East Sussex: Psychology Press Ltd.

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

22. Barlow, M. & Kemmer, S. (Eds.). (2000). *Usage based models of language*. Stanford: CSLI.

23. Bates, E., & Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney, (Ed.), *The Emergence of Language* (pp. 29-79). New Jersey: Lawrence Erlbaum Associates.

24. Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E.Bates (Eds.), *The crosslinguistic study of sentence processing* (3-76). New York: University of Cambridge Press.

25. Bates E, Devescovi A, Pizzamiglio L, D'Amico S, Hernandez A. (1995). Gender and lexical access in Italian. *Perception and Psychophysics*, *57*, 847–862.

26. Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, L. Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., & Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin and Review, 10*, 344-380.

27. Benedict, H. (1979). Early lexical development: Comprehension and production. *Journal of Child Language*, *6*, 183-200.

28. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*, 1165-1188.

29. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol: O'reilly Media.

30. Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341-345.

31. Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*, 158-167.

32. Borer, H. (2005). *Structuring sense volume 1: In name only*. Oxford: Oxford University Press.

33. Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B, 26*, 211–252.

34. Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of

the 38th Annual Conference of the Cognitive Science Society (pp. 1691–1696). Austin: Cognitive Science Society.

35. Branigan, H. & Pickering, M. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*, E282.

36. Bresnan, J. (2001). *Lexical-functional syntax*. Oxford: Blackwell.

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

38. Brysbaert, M. & New, B. (2009) Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods, 41, 977-990*.

39. Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412-424.

40. Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904-911.

41. Bullinaria, J. A. & Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, *39*, 510-526.

42. Bullinaria, J.A. & Levy, J.P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods, 44*, 890-907.

43. Bybee, J. (2010). *Language, usage, and cognition*. Cambridge: Cambridge University Press.

44. Chafe, W. L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

45. Chao, A., Wang, Y. T., & Jost, L. (2013). Entropy and the species accumulation curve: a novel estimator of entropy via discovery rates of new species. *Methods in Ecology and Evolution*, *4,* 1091–1110.

46. Choi, J. D., & Palmer, M. (2012). Guidelines for the Clear Style Constituent to Dependency Conversion. *Technical Report 01-12*, University of Colorado Boulder: Institute of Cognitive Science.

47. Chomsky, N. (1957). *Syntactic structure*. Cambridge: MIT Press.

48. Chomsky, N. (1970). Remarks on nominalization. In Jacobs, R. & Rosenbaum, P. (Eds.), *Readings in English Transformational Grammar* (pp. 184-221). Waltham: Ginn.

49. Chomsky, N. (1980). On Cognitive Structures and their Development: A reply to Piaget. In M. Piattelli-Palmarini (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky* (pp. 35-52) Cambridge: Harvard University Press.

50. Chomsky, N. (1995). The minimalist program. Cambridge: MIT Press.

51. Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

52. Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *The Quarterly Journal of Experimental Psychology*, *60*, 1072-1082.

53. Costa, A., Kovacic, D., Fedorenko, E., & Caramazza, A. (2003). The gender congruency effect and the selection of freestanding bound morphemes: Evidence from Croatian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1270-1282.

54. Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society B*, *34*, 187-220.

55. Cubelli, R., Lotto, L., Paolieri, D., Girelli, M., & Job, R. (2005). Grammatical gender is selected in bare noun production: Evidence from the picture-word interference paradigm. *Journal of Memory and Language*, *53*, 42–59.

56. Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

57. Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review*, *117*, 713-758.

58. de Simone, F., & Collina, S. (2015). The picture-word interference paradigm: Grammatical class effects in lexical production. *Journal of Psycholinguistic Research*, DOI: 10.1007/s10936-015-9388-9.

59. Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes*, *23*, 583-608.

60. Dhooge, E. & Hartsuiker, R. J. (2010). The distractor frequency effect in picture-word interference: Evidence for response exclusion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 878-891.

61. Diessel, H. (2015). Usage-based construction grammar. In E. Dabrowska and D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 295-321). Boston: De Gruyter.

62. Duràn, C. P. & Pillon, A. (2011). The role of grammatical category information in spoken word retrieval. *Frontiers in Psychology*, *2*, 1-20.

63. Feldman, L. B., Milin, P., Cho, K. W., Fermín Moscoso del Prado Martín, F., & O'Connor, P. (2015). Must analysis of meaning follow analysis of form? A time course analysis. *Frontiers in Human Neuroscience*, *11*, 1-19.

64. Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from Chronolex. *Frontiers in Psychology*, *2*, 1-10.

65. Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, *46*, 57–84.

66. Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language*, *35*, 724–755.

67. Fillmore, C. J. (1986). Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Linguistics Society* (pp. 95–107). Berkeley: Berkeley Lingusitics Society.

68. Fillmore, C. J., Lee-Goldman, R., & Rhodes, R. (2012). The FrameNet constructicon. In H. C. Boas & I. A. Sag (Eds.), *Sign-based Construction Grammar* (pp. 309-372). Stanford: CSLI.

69. Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing “the” as “thee” to signal problems in speaking. *Cognition*, *62*, 151-167.

70. Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, *80*, 748-775.

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

72. Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*, 684-691.

73. Gleitman, L. R., Gleitman, H., & Shipley, E. F. (1972). The emergence of the child as grammarian. *Cognition, International Journal of Cognitive Psychology, 1,* 137-164.

74. Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure constructions*. Chicago: University of Chicago Press.

75. Goldberg, A. E. (2006). *Constructing a language: The nature of generalization in language*. Oxford: Oxford University Press.

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

77. Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: A Tool for Forced Alignment of Laboratory Speech. *Canadian Acoustics*, *39*, 192–193.

78. Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*, 254-260.

79. Grainger, J. & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518-565.

80. Grambsch, P., & Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81,* 515-26.

81. Gregory, E., Varley, R., Herbert, R. (2012). Determiner primes as facilitators of lexical retrieval. *Journal of Psycholinguistic Research*, *41*, 439-453.

82. Gries, S. Th., & Stoll, S. (2009). How to measure development in corpora: An association strength approach. *Journal of Child Language, 36*, 1075-1090.

83. Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

84. Harmsen, W. N. (2017). *Predicting word learning order in Dutch and English using different word frequencies and other word attributes*. Unpublished Bachelor Thesis, School of Psychology and Artificial Intelligence, Radboud University.

85. Hausser, J. & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*, 1469-1484.

86. Hendrix, P., Bolger, P. and Baayen, R. H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 128-149.

87. Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1373-1378). Lisbon, Association for Computational Linguistics.

88. Hudson, R. (2007). *Language networks: The new Word Grammar*. Oxford: Oxford University Press.

89. Hutchison, K.A., Balota, D.A., Neely, J.H., Cortese, M.J., Cohen-Shikora, E. R., Tse, Chi-Shing, Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The Semantic Priming Project. *Behavior Research Methods*, *45*, 1099-1114.

90. Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, *13*, 411-430.

91. Jackendoff, R. (1977). *X-bar syntax: A study of phrase structure*. Cambridge: MIT Press.

92. Jackendoff, R. (2013). Constructions in the parallel architecture. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 70-92), Oxford: Oxford University Press.

93. Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630-645.

94. Kay, P. & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The 'What's X doing Y?' construction. *Language*, *75*, 1-33.

95. Kay, P. (2013). The limits of (construction) grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 32-48). Oxford: Oxford University Press.

96. Keuleers, E. (2013). *vwr: Useful functions for visual word recognition research*. *R* package version 0.3.0. https://CRAN.R-project.org/package=vwr.

97. Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287-304.

98. Kostić, A. Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In R. H. Baayen & R. Schreuder (Eds.), *Morphological Structure in Language Processing* (pp. 1-44). New York: Mouton de Gruyter.

99. Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*, 978-990.

100. La Heij, W., Mark, P., Sander, J., & Willeboorsde, E. (1998). The gender congruency effect in picture word task. *Psychological Research*, *61*, 209–219.

101. Lam, K. J. Y., Dijkstra, T., & Rueschemeyer, S-A. (2015). Feature activation during word recognition: action, visual, and associative-semantic priming effects. *Frontiers in Psychology*, *6*, 1-8.

102. 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.

103. Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. I: Theoretical prerequisites*. Stanford: CSLI.

104. Lester, N. A. & Moscoso del Prado Martín, F. (2016). Syntactic flexibility in the noun: Evidence from picture naming. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2585-2590). Austin, TX: Cognitive Science Society.

105. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, *163*, 845-848.

106. Lewis, M., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182-195.

107. Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*, B65-B73.

108. Lin, J. (1991). Divergence measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, *37*, 145-151.

109. Linzen, T., Marantz, A., & Pylkkänen, L. (2013). Syntactic context effects in visual word recognition. *The Mental Lexicon*, *8*, 117-139.

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

111. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55-60). Baltimore, Association for Computational Linguistics.

112. Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania Working Papers in Linguistics*, *4*, 201-225.

113. Marchini, J. L., Heaton, C., & Ripley, B. D. (2013). *fastICA: FastICA algorithms to perform ICA and projection pursuit*. *R* package version 1.2-0. https://CRAN.R-project.org/package=fastICA.

114. Markoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971-979.

115. Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314-324.

116. McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375-407.

117. Mel'čuk, I. (1988). *Dependency syntax: Theory and practice*. Albany: The SUNY Press.

118. Meyer, A. S., Roelofs, A., & Levelt, W. J. M. (2003). Word length effects in object naming: The role of a response criterion. *Journal of Memory and Language*, *48,* 131-147.

119. Milin, P., Filipović-Đurđević, D. & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*, 50–64.

120. Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology* (pp. 95–100). Glencoe, IL: Free Press.

121. Mintz, T. H. (2003). Frequent frames as a cue to grammatical categories in child-directed speech. *Cognition*, *90,* 91-117.

122. Miozzo, M., & Caramazza, A. (2003). When more is less: A counterintuitive effect of distractor frequency in the pcture-word interference paradigm. *Journal of Experimental Psychology: General*, *132,* 228-252.

123. Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 116-133.

124. Morton, J. (1978). Word recognition. In J. Morton and J.C. Marshall (Eds.), *Psycholinguistics* (pp. 109-151), London: Elek Scientific Books.

125. Moscoso del Prado Martín, F. (2007). Co-occurrence and the effect of inflectional paradigms. *Lingue e Linguaggio*, *6,* 247-263.

126. Moscoso del Prado Martín, F. (2016). Vocabulary, grammar, sex, and aging. *Cognitive Science*, 1-26. DOI: 10.1111/cogs.12367.

127. Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94,* 1-18.

128. Myachykov, A., Garrod, S., & Scheepers, C. (2009). Attention and syntax in sentence production: A critical review. *Discours*, *4,* 1-17.

129. Myachykov, A., Scheepers, C., Garrod, S., Thompson, D., & Fedorova, O. (2013). Syntactic flexibility and competition in sentence production: The case of English and Russian. *The Quarterly Journal of Experimental Psychology*, *66*, 1601-1619.

130. Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theory. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-336). Hillsadale, NJ: Erlbaum.

131. Neidle, C. (1994). Lexical-Functional Grammar (LFG). In R. E. Asher (Ed.), *Encyclopedia of Language and Linguistics* (pp. 2147-2153). Oxford: Pergamon Press.

132. New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review, 13*, 45-52.

133. Nivre, J. 2005. *Dependency grammar and dependency parsing*. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.

134. Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327-357.

135. Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Science, 17*, 517-524.

136. Novick, J. M., Kim, A., Trueswell, J. C. (2003).Studying the grammatical aspects of word recognition: Lexical priming, parsing, and syntactic-ambiguity resolution. *Journal of Psycholinguistic Research*, *32,* 57-75.

137. Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology*, *17*, 273–281.

138. O'Regan, J. K., & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 185-197.

139. Pace-Sigge, M. (2013). *Lexical priming in spoken English usage*. New York: Palgrave.

140. Pickering, M. J., & Branigan H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.

141. Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 767-808.

142. Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*, 786-823.

143. Ramchand, G. C. (2008). *Verb meaning and the lexicon: A first phase syntax*. Cambridge: Cambridge University Press.

144. Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182.

145. Reppen, R., Ide, N., & Suderman, K. (2005). *American National Corpus (ANC) Second Release*. Philadelphia: Linguistic Data Consortium.

146. Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64–99). New York: Appleton-Century-Crofts.

147. Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*, 348-379.

148. Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews, Cognitive Science*, *1*, 906-914.

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

150. Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: Lexical and structural priming of syntactic constructions in young children. *Developmental Science*, *6*, 557-567.

151. Schiller, N.O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, *48*, 169-194.

152. Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language*, *37*, 118-139.

153. Schuster, S., & Manning, C. D. (2016). Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 2371-2378).

154. Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23*, 569-588.

155. Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523-568.

156. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *28*, 379-423.

157. Shimpi, P. M., Gámez, P. B., Hutternlocher, J., & Vasilyeva, M. (2007). Syntactic priming in 3- and 4-year-old children: Evidence for abstract representations of transitive and dative forms. *Developmental Psychology, 43,* 1334-1346.

158. Smolik, F. (2014). Noun imageability facilitates the acquisition of plurals: Survival analysis of plural emergence in children. *Journal of Psycholinguistic Research, 43,* 335-350.

159. Smolík, F., & Kříž, A. (2015). The power of imageability: How the acquisition of inflected forms is facilitated in highly imageable verbs and nouns in Czech children. *First Language*, *35*, 446-465.

160. Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8,* 209-243.

161. Stoll, S., Bickel, B., Lieven, E., Paudyal, N. P., Banjade, G., Bhatta, T. N., Gaenszle, M., Pettigrew, J., Rai, I. P., Rai, M., & Rai, N. K. (2012). Nouns and verbs in Chintang: Children's usage and surrounding adult speech. *Journal of Child Language*, *39*, 284-321.

162. Tabak, W., Schreuder, R., and Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon, 5*, 22-46.

163. Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

164. *The British National Corpus, version 3* (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

165. 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.

166. Therneau, T. (2015). *A Package for Survival Analysis in S*. version 2.38, https://CRAN.R-project.org/package=survival.

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

168. Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

169. Tomlin, R. S. (1995). Focal attention, voice, and word order: An experimental cross-linguistic study. In P. Downing & M. Noonan (Eds.), *Word order in discourse* (pp. 517-554). Philadelphia: John Benjamins.

170. van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111-122.

171. Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176-1190.

172. Veldre, A., & Andrews, S. (2018). Beyond close probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language, 100*, 1-17.

173. 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.

174. Wood, S. N. (2016). Just Another Gibbs Additive Modeler: Interfacing JAGS and mgcv. *Journal of Statistical Software*, *75*, 1-15.

175. Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression models does (and what it does not do). *Journal of Memory and Language*, *72,* 37-48.

176. Yarkoni,T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review*, *15*, 971-979.

177. Zeldes, A. (2013). *Productivity in argument selection: From morphology to syntax*. Boston: De Gruyter.

178. Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Jornal of Memory and Language*, *47*, 1-29.

179. Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G., M. (2009). *Mixed effects models and extensions in ecology with R*. New York: Springer.

Appendix

A. A database of the syntactic diversity of English Nouns: SynDi-EN

I introduce a number of syntactic measures aimed at capturing a fine-grained perspective on the diversity and typicality of the distributions of English nouns. I then compute these measures for two varieties of English – American and British – and package them into a database of more than 5000 distinct nouns suitable for integration with several previously published norming datasets. Exemplars at the high and low ends of the measurement scales are presented.

1. Selecting a syntactic formalism

Linguistic theorists have produced many formalisms aimed at describing the relationship between words and syntactic structure. Prominent competing formalisms include phrase-structure grammars (e.g., Chomsky, 1957), dependency grammars (e.g., Tesnière, 1959), and construction grammars (e.g., Boas & Sag, 2012). These formalisms each have strengths and weaknesses. I propose four desiderata based on the needs of the present study to identify the ideal formalism. First, the ideal formalism must provide a means for unambiguously determining the relative sequential position of a target word relative to syntactically related words. Second, it must capture which other words in a given syntactic domain depend on the target noun for their realization and/or interpretation. Third, it should provide ready labels for the syntactic functions served by a given noun. By function, I mean that the formalism should discriminate at some degree of granularity the various types of syntactic relationships that nouns may enter into with respect to other words. Fourth, it should allow for straightforward computer-automated processing.

The first criterion, related to encoding of word order, is satisfied by almost all contemporary formalisms. However, some systems designed to handle variable discontinuities in syntactic dependency (*scrambling*) avoid explicitly specifying linear order outside of the linguistic token itself (Mel'čuk, 1988). This approach therefore precludes any investigation of the representation of word-order asymmetries in the syntactic information carried by nouns. The second and third criteria, which concern which words link up syntactically and through what kind of relation, respectively, are more useful for discriminating between alternative formalisms. The three formalisms mentioned above vary with respect to how transparently they reflect both kinds of information. Consider perhaps the most widely used syntactic formalism: the phrase-structure (PS) tree. PS trees consist of typed nodes and (non-typed) arcs. Nodes represent lexical items (terminal nodes) and groups of words (non-terminal or phrasal nodes). These nodes are connected via vertical arcs that indicate which lower-level nodes are bound to which higher-order nodes (immediate constituency; Bloomfield, 1933; Chomsky, 1957). To determine which words are related to a target, one must traverse a potentially complex path via the set of intervening arcs and nodes. While not computationally intractable, the complexity of these paths makes the PS tree formalism a somewhat cumbersome choice, if only for purposes of exposition.

Furthermore, because the connecting arcs are untyped, information about functional relationships between words must be distributed throughout the tree. Identifying a given relationship requires one to consider at least the types of nodes intervening between the words (if one were to trace a path along the arcs that connect them), as well as the relative positioning of the words with respect to those nodes. For example, the word *stealthy* in the noun phrase *the stealthy owl* can only be identified as an adjectival modifier of *owl* in a typical PS tree by (1) its subordination to an AdjP (adjective phrase) node (2) that is left-sister to the word-class non-terminal N node (3) that dominates *owl* and (4), where both AdjP and N are eventually dominated by a single higher-order NP (noun phrase) node. The distributed nature of PS grammars thus presents a somewhat of a challenge for my fourth criterion.

An increasingly popular alternative to the PS notation is the *dependency graph* (DG; Tesnière, 1959; Mel'čuk, 1988, 2011; Nivre, 2005). DGs are trees whose nodes represent lexical items and whose arcs represent typed syntactic relations. By convention, DG formalisms tend to include only binary relationships between words, with one privileged relationship linking a word (usually the verb in a finite clause) to the utterance-generating root node.[10] These relationships are known as dependencies. Dependencies are asymmetrical, in that one word – the head – licenses the presence of the other word – the modifier (*governor* and *subordinate* in language of Tesnière, 1959). Heads, their modifiers, and the dependencies that bind them are together known as constructions. To avoid confusion between this and other, more widespread uses of the term construction (e.g., the “construction” of Construction Grammar; Goldberg, 1995), I will refer to the head-modifier-

[10] Root nodes are common to dependency grammars and phrase-structure grammars. However, the notion of root differs across the two formalisms. In DGs, the root connects directly to a lexical item, in agreement with the *lexicalist hypothesis* (Levelt, 1989; Bresnan, 2001). The lexicalist hypothesis states that words project their own syntactic structures to be unified with other co-active words. Syntactic projections are often referred to as *subcategorization frames* (Chomsky, 1965).

dependency trio as a bundle. Much more could be said about the nature of dependencies, including how headship is decided, what types of dependencies are allowed, whether dependencies can only hold between words, or whether supra-lexical (i.e., 'phrasal') nodes should be allowed, and so on. Such questions are addressed from theoretical (e.g., Mel'čuk, 1988; Hudson, 2007) and practical (e.g., Nivre, 2005) perspectives elsewhere. Here, I rely on the standards described by Choi and Palmer (2012) for English.

The CLEAR DG formalism (CDG; Choi & Palmer, 2012) readily meets the four criteria laid out above. While other DG formalisms ignore word order (Mel'čuk, 1988; Criterion 1), CDG supplies for each bundle the sequential position of the head and modifier within the overall string. The question of which words are related (Criterion 2) is replaced by a simpler question – which words are directly functionally related. For any target, the set of related words consists of those that are bundled with the target as head or modifier. The nature of these relationships (Criterion 3) is plainly indicated by a functional tag (e.g., *nsubj* for the clausal subject relation). The direct representation of each of these pieces of information within CDG means that it allows for straightforward computer-automated processing (Criterion 4).

CDG has several limitations. Perhaps chief among them is that it cannot directly associate meaning to complex structures (i.e., constructions involving more than one dependency). Consider the so-called *caused-motion construction*, which in English takes the form X[*agent*] VERB[*cause* + *move*] Y[*theme*] PREP[*path*] Z[*ground*], as in *Claude flicked the letter through the mail slot.* Fully generalized frames of this sort can be associated directly with various types of meaning. These meanings are revealed through phenomena such as semantic coercion (e.g., intransitive *laugh* takes a force-dynamic interpretation in *The audience laughed the speaker out of the session*), selectional restrictions (e.g., what types of words may surface in each syntactic "slot"), and so on. Importantly, the interpretations derived from instances of the caused-motion construction are not reducible to the content of the component phrases or words. Therefore, any realistic grammatical formalism must be able to account for the non-decompositional meanings that attach to syntactic templates at the phrasal, clausal, and supra-clausal levels (for a compelling discussion of such meanings, see Goldberg, 1995; see Linzen et al., 2013, for an operationalization of syntactic diversity using such structures). While I acknowledge the importance of such constructions for our understanding of the true syntactic space of any language, I set these concerns aside for later research. I do so for four reasons.

First, many theories of grammar acknowledge the independent status of structural representations at multiple levels of specificity (e..g, Culicover & Jackendoff, 2005; Goldberg, 1995; Langacker, 1987). For example, the simple transitive construction <$\text{NP}_{\text{subject}}$ VERB $\text{NP}_{\text{object}}$> involves two types of relationship between verb and NP, indicated here by subscripts on the NPs. Both of these syntactic relationships are also attested outside of simple transitive construction (subjects also appear in intransitive and ditransitive clauses, and objects also appear in ditransitive clauses). Therefore, the holistic schema (the construction) can be broken into subschemas (the syntactic dependencies). This subdivision can be pursued further, for example, to the individual words that fill out the NPs, along with the syntactic dependencies that bind them into their respective subunits. According to the theories introduced above, each of these subunits becomes activated as a function of their relation to the whole (e.g., as weighted by distributional biases; Stefanowitsch & Gries, 2003). In particular, any given word should be co-distributed across multiple nested tiers of syntactic abstraction, including the low-level structures studied here. The question is whether these distributions are functionally relevant to word production.

Second, no grammar of any language purports to account for every construction in that language. This is no more true of words than it is of higher-order constructions. In fact, the constructionist approach has spurred generations of construction-hunters to identify and catalog their quarry with greater and greater levels of precision. The more constructions are uncovered, the further removed seems the goal of an exhaustive taxonomy. Add to this the fact that languages do not sit still, but change constantly under internal and external pressures (Thomason & Kaufman, 1992), and the “true” syntactic space of a language becomes a moving target. Therefore, even if I wanted to derive syntactic measures from parses reflecting the true syntactic space, I should always face the possibility – in truth, the inevitability – of incompleteness. I tackle this necessary incompleteness by assuming only the reality of dependency bundles and the set of typed relations specified by CDG. This assumption comes with the caveat that I model only a 'toy' representation of the full grammar. Future improvements may replace or elaborate the relations considered here.

Third, while the set of dependencies in CDG is rather small, even a conservative estimate of the total number of syntactic constructions in English is much larger. Moreover, the frequency distributions of these structures should follow a Zipf-Mandelbrot distribution (e.g., Zipf, 1935). This means that the expected frequencies for the vast majority of structures are extremely low. Therefore, we cannot expect finite samples of the size typical for syntactically annotated English corpora to produce reliable frequency estimates for the bulk of these constructions. By paring the space of alternatives down, I increase the likelihood of observing a sufficient number of tokens for each type to support – at the defined granularity – reliable estimates of frequency, hence diversity.

Fourth and finally, some work on associative learning has argued that low-level representations may play a more significant role in adult processing than higher-level collocational or constructional units (Baayen et al., 2011; Ramscar et al., 2010). According to this argument, the links between low-level (lexical and sub-lexical) and high-level (collocational and collostructional) units will tend to weaken over time as speakers experience an increasingly diverse set of ways in which the two may combine. Evidence that the higher-order relationships are not necessary comes from the fact that models that lack explicit representations for syntax learn associations between words and meanings that predict adult behavior in word production and comprehension tasks (Baayen et al., 2011; Baayen et al., 2013; Hendrix et al., 2017). However, these models do include grammatical information in the input, such as labels for inflectional categories (e.g., case labels). Therefore, the associations depend on syntactic information even if the connectionist architecture does not include a dedicated tier of syntactic nodes. Assuming that syntax matters, and assuming that lower-level relationships should dominate adult linguistic processing, distributions within the CDG space provide the best chances of identifying synax-specific effects in word processing.

For these reasons, I consider CDG a desirable formalism for beginning my investigation of syntactic diversity within the lexicon. The dependency types define a syntactic distributional space across which all occurrences of nouns are distributed. Based on the prior findings, I expect the shapes of these distributions to impact processing. To measure the shape of a given word's distribution in this space, I turn to information theory.

B. Syntactic diversity as entropy

The syntactic diversity of a noun *w* has two key components: (1) the set of possible syntactic constructions *C* and (2) the probability that *w* occurs in each construction c in *C*. Nouns should be considered more diverse to the extent that *C* increases. This relationship captures the common sense intuition that nouns that occur in a greater variety of constructions are more syntactically diverse. But this is only half the story. To see this more clearly, imagine that we extract 1000 instances of some noun from a corpus. We find tokens embedded in each of the possible syntactic structures defined in *C*. According to the metric just introduced, the noun exhibits maximal diversity. However, looking more closely, we notice that 900 tokens occurred in a single construction c_1, while the remaining 100 tokens are distributed relatively evenly across the remaining constructions. Now consider a different noun that also occurs in every available construction, but which occurs equiprobably in each construction. According to our first metric, the two nouns are equally diverse. And yet, our intuition suggests that the second noun is much more diverse than the first. The optimal measure of diversity should take both sources of information into account: *instance* (did it occur?) and *rate* (how frequently?). The measure should also account for the full distribution simultaneously (i.e., by taking some central tendency of the noun's distribution across all constructions). One measure that satisfies all of these criteria is the entropy *H* (Shannon, 1948). Entropy measures the average amount of uncertainty within a distribution. Applied to syntactic distributions as defined above, it represents how uncertain we are about assigning a given noun to any of the available constructions. In processing terms, it measures the richness of the spreading activation between lemmas and syntactic structures, with high uncertainty translating into richer patterns of activation. More specifically, entropy increases as the number of possible constructions increases and as the frequency distribution across these constructions approaches maximum uncertainty (equiprobability, or equally strong sets of connections). Therefore, holding the dimensionality of the syntactic space constant, the most diverse noun is the one that is least biased towards a particular subset of the possible constructions. Entropy has proved useful for estimating the syntactic diversity of full grammars (probabilistic context-free grammars) induced from treebanks (Moscoso del Prado Martín, 2014), as well as for estimating the morphological diversity of words (Moscoso del Prado Martín et al., 2004).

The entropy requires a probability distribution defined over a syntactic space (i.e., sets of possible constructions). I distinguish between nine such spaces. First, nouns may register distributional information about all dependencies to which they belong, irrespective of whether they serve as heads or modifiers. Therefore, I define a total syntactic distribution for which each cell contains the joint probability $p(w, d)$ of target noun w in dependency d. I refer to the (joint) entropy of this distribution as H_t for *total entropy*. But this measure may be decomposed further. By taking hierarchical status into account, we can dissociate the diversity of relations for which the noun is a head or a modifier. This decomposition may be necessary given evidence that heads and modifiers are treated differently by the syntactic machinery (e.g., Bürki et al., 2016), which may have consequences for lexical access more generally. Therefore, I define two new distributions consisting of the joint probabilities $p(w, d, h=\text{head})$ and $p(w, d, h=\text{modifier})$ for target noun *w* occurring in dependency *d* in hierarchical role *h* of head or modifier, respectively. I refer to the (joint) entropies of these distributions as H_h and H_m, for *as-head diversity* and *as-modifier diversity*. Finally, these distributions can be conditioned for word order: does the target follow or precede the word with which it is bundled? The former I refer to as a *rightward-facing* dependency, the latter as a *leftward-facing* dependency. Adding this dimension produces six additional distributions: *rigthtward* (H_{rt}), *rightward as-head* (H_{rh}), *rightward as-modifier* (H_{rm}), *leftward* (H_{lt}), *leftward as-head* (H_{lh}), *and leftward as-modifier* (H_{lm}). See Figure 3, Chapter II, for a schematization.

C. Accounting for lexical confounds and estimation biases

As already mentioned, syntactic dependencies are partially redundant given the lexical items that instantiate them. For example, if the word *this* is found bundled with *café*, then one knows immediately that the syntactic relation is *determiner.* In such cases, the information carried by the syntactic relation is partially or wholly reducible to that carried by the lexical context. We must therefore clean the syntactic information of the lexical confound if we are to produce a truly syntactic measure of distributional diversity. For this purpose, we can use another information-theoretic measure known as the conditional entropy. Conditional entropy of the dependencies *D* given the lexical items *L*, or $H(D \mid L)$, is defined formally as in Eq. 14.

$$H(D|L) = H(D,L) \text{ -} H(L) \tag{14}$$

The conditional entropy requires that we take the joint entropy of the dependencies and lexical items and subtract the entropy of the lexical items alone. What remains is the information carried by the dependencies without that of the words or the mutual information carried between the dependencies and words (a more thorough account of this relationship is given in Chapter 4). I compute the conditional entropy rather than the simple entropies for each of the nine distributions. Otherwise, I risk confounding syntax and semantics. This is a serious problem for interpreting correlations between these measures and human behavior. Semantics plays a powerful role in human processing, and without some means of distinguishing semantic from syntactic information, one risks gross misinterpretation of semantic effects as “syntactic” in nature.

When based on raw probability estimates, the entropies in Eq. 14 constitute maximum-likelihood estimates. Such estimates are biased downward (Miller, 1955). To correct for this bias, I apply the smoothing technique of Chao, Wang, and Jost (2013), which has been shown to perform well at handling linguistic distributions (Moscoso del Prado Martín, 2016). Specifically, I correct each of the component entropies $H(D\ , L)$ and $H(L)$ prior to the subtraction.

D. Prototypicality

Unlike for entropy, distance measures such as the relative entropy do not provide straightforward means for cleaning lexical confounds. For that reason, I rely on the dependency-only distributions outlined above in Figure 3. One issue with prior operationalizations of prototypicality (Baayen et al., 2011; Hendrix, et al., 2017; Linzen et al., 2013; Milin et al., 2009) is that the relative entropy is asymmetric: the value differs depending on whether one measures the divergence of the prototype from the target distribution or vice versa. I see no theoretical reason to prefer one direction over the other. Therefore, I use an alternative, symmetrical measure of the distance between distributions known as the Jensen-Shannon divergence (JSD; Lin, 1991). JSD gets around the asymmetry problem of relative entropy by first taking the midpoint between the two distributions to be compared. Then, the relative entropy is calculated separately from each distribution to the midpoint, and the resulting values are averaged together. The raw probabilities again underestimate the true probabilities, which negatively affects the accuracy of the JSD estimate. I therefore smooth the probability distributions prior to computing JSD. When comparing two distributions with JSD, the number of possible outcomes must be held constant between them. Under these conditions, the optimal smoothing strategy is known as the James-Stein shrinkage estimator (Hausser & Strimmer, 2009).

E. Data

Probability estimates for the diversity and prototypicality measures were drawn from two large corpora. The two corpora contain American and British English, respectively. The American data come from the Open American National Corpus (OANC; Reppen, Ide, & Suderman, 2005), which contains roughly 15 million words of writing and transcribed speech. The British data come from the British National Corpus (BNC; *British National Corpus*, 2007), which contains nearly 250 million words. However, to achieve a greater degree of comparability between the two corpora, I sample 15 million words by shuffling the corpus files and extracting the first 15 million words of running text (respecting sentence boundaries)[11]. While this represents only a small proportion of the overall BNC, reliable and replicable results for lexical co-distributions have been achieved with much smaller samples (e.g., the 10-million-word spoken component of the BNC; McDonald & Shillcock, 2001a,b).

I parse both corpora using the *spaCy* parser (Honnibal & Johnson, 2015) to derive CDG-style dependency graphs. For each noun lemma, I track the frequency of tuples containing the co-bundled word, the dependency type, the hierarchical status of the target, and direction of the relation.[12] For example, given the phrase *the stealthy owl*, I would isolate *owl* and extract two tuples: (*the, determiner,* head, leftward) and (*stealthy, adjectival modifier,* head, leftward). Using these tuples, along with their associated frequencies, I compute the conditional entropies and prototypicalities for the nine distributions laid out in Figure 3. I repeat this process for each corpus to create two sets of estimates. Naturally, the parses produce some degree of noise, such that many of the items returned are not nouns. Moreover, English nouns frequently undergo zero-conversion to function as other part-of-speech classes, resulting in a high degree of homography. To guard against faulty parses and homography, I restrict the sample to only those forms that are annotated as functioning primarily as nouns in a previously published database of lexical norms (Brysbaert, Warriner,

[11] Practical considerations also guide this decision. Time estimates to process the entire 250 million words of the BNC run into the range of several months (24-hour continuous processing) given my computational resources. Any researcher with greater resources interested in achieving more reliable estimates by processing the BNC or other massive corpora is welcome to the extraction code.

[12] Not discussed here are two additional parameters that are available in the full database: number of the noun, singular or plural (for calculation of inflectional entropy, or conditioning on morphological form), and mode of the production, spoken or written. Code is provided with the database to condition frequency distributions and entropies on any of these dimensions. Lemma-based token frequencies are also available.

& Kuperman, 2014). The final tally was 5727 distinct noun lemmas for the American English sample, and 5699 lemmas for the British sample. Tables 7-8 show the five most and least diverse noun lemmas per distribution for the OANC data, respectively. Tables 9-10 show the most and least prototypical noun lemmas. Tables 11-14 show the same for the BNC data.

Table 7: Most diverse nouns: OANC

Distribution	1	2	3	4	5
H_h	sequence	serum	index	matrix	glucose
H_m	sequence	serum	alpha	glucose	intake
H_t	fortress	discharge	spacing	palace	cleavage
H_{rh}	sequence	serum	index	dose	growth
H_{rm}	sequence	serum	intake	passage	index
H_{rt}	passage	fortress	plateau	gas	plasma
H_{lh}	sequence	serum	transport	index	matrix
H_{lm}	sequence	alpha	core	growth	access
H_{lt}	yeast	uptake	glucose	alpha	intake

Table 8: Least diverse words: OANC

Distribution	1	2	3	4	5
H_h	diva	junk	tantrum	axe	dock
H_m	wizard	rapist	cone	pitfall	craftsman
H_t	cracker	hostess	statehood	tofu	campground
H_{rh}	handful	stair	boom	cord	proof
H_{rm}	advent	impulse	bracelet	goat	pitfall
H_{rt}	meadow	morale	keeper	cracker	tic
H_{lh}	diva	fable	pulpit	cement	junk
H_{lm}	symptom	clinic	context	council	gourmet
H_{lt}	fable	tick	dawn	pulpit	lifer

Table 9: Most prototypical nouns: OANC

Distribution	**1**	**2**	**3**	**4**	**5**
$\mathbf{H_h}$	patch	base	option	building	brand
$\mathbf{H_m}$	machine	rain	wave	castle	block
$\mathbf{H_t}$	route	band	building	model	plot
$\mathbf{H_{rh}}$	code	trial	movement	building	service
$\mathbf{H_{rm}}$	rain	palace	style	string	project
$\mathbf{H_{rt}}$	band	product	code	network	theme
$\mathbf{H_{lh}}$	race	song	option	party	flight
$\mathbf{H_{lm}}$	machine	money	head	race	belt
$\mathbf{H_{lt}}$	race	crime	cow	beer	option

Table 10: Least prototypical nouns: OANC

Distribution	**1**	**2**	**3**	**4**	**5**
$\mathbf{H_h}$	forte	whatnot	kneecap	deathbed	stead
$\mathbf{H_m}$	totem	whatnot	screwball	excise	lymph
$\mathbf{H_t}$	whatnot	chevron	wedlock	plantain	cleaver
$\mathbf{H_{rh}}$	postman	huntsman	quintet	whatnot	turncoat
$\mathbf{H_{rm}}$	whatnot	bonkers	lasso	ditto	quitter
$\mathbf{H_{rt}}$	whatnot	dreamworld	bluebird	puzzler	smokescreen
$\mathbf{H_{lh}}$	forte	deathbed	kneecap	croquet	wader
$\mathbf{H_{lm}}$	totem	rookie	amber	instant	sham
$\mathbf{H_{lt}}$	chevron	codpiece	wader	stead	phlegm

Table 11: Most diverse nouns: BNC

Distribution	**1**	**2**	**3**	**4**	**5**
$\mathbf{H_h}$	blowout	twig	oxen	gal	opal
$\mathbf{H_m}$	ferret	rye	mafia	smuggler	peeler
$\mathbf{H_t}$	freelance	amber	rye	ferret	ripeness
$\mathbf{H_{rh}}$	eyesore	lamppost	ripeness	rave	triplet
$\mathbf{H_{rm}}$	loco	hairpin	polyp	freelance	flap
$\mathbf{H_{rt}}$	shrink	minnow	picker	freelance	rescuer
$\mathbf{H_{lh}}$	cologne	bedtime	bingo	brie	llama
$\mathbf{H_{lm}}$	mallard	warbler	finder	broiler	softie
$\mathbf{H_{lt}}$	puss	rashness	governess	brushwood	beaker

Table 12: Least diverse words: BNC

Distribution	**1**	**2**	**3**	**4**	**5**
$\mathbf{H_h}$	exhaust	syrup	tribesman	nutmeg	drainage
$\mathbf{H_m}$	hoard	sweetheart	fury	pamphlet	jaguar
$\mathbf{H_t}$	barman	herdsman	scum	marshal	audience
$\mathbf{H_{rh}}$	monk	footpath	eagle	boxer	shrub
$\mathbf{H_{rm}}$	teen	tablet	men	trustee	bureau
$\mathbf{H_{rt}}$	juror	spark	helper	poacher	whore
$\mathbf{H_{lh}}$	diver	roofing	fusion	mint	rush
$\mathbf{H_{lm}}$	lantern	blacksmith	monk	trifle	pepper
$\mathbf{H_{lt}}$	diver	bloodshed	sage	junk	clone

Table 13: Most prototypical nouns: BNC

Distribution	1	2	3	4	5
$\mathbf{H_h}$	target	trial	record	note	branch
$\mathbf{H_m}$	machine	horse	wheel	shadow	shell
$\mathbf{H_t}$	race	game	house	rule	word
$\mathbf{H_{rh}}$	theory	camp	force	game	movement
$\mathbf{H_{rm}}$	life	machine	soul	paper	pleasure
$\mathbf{H_{rt}}$	soul	game	door	race	ball
$\mathbf{H_{lh}}$	note	song	size	meal	tune
$\mathbf{H_{lm}}$	row	shadow	response	traffic	price
$\mathbf{H_{lt}}$	size	film	design	house	track

Table 14: Least prototypical nouns: BNC

Distribution	1	2	3	4	5
$\mathbf{H_h}$	backache	highness	godson	airmail	airway
$\mathbf{H_m}$	backroom	gab	cretin	scuba	sinker
$\mathbf{H_t}$	polka	sinker	linseed	teargas	blackjack
$\mathbf{H_{rh}}$	boatman	sniffer	quail	abscess	ahoy
$\mathbf{H_{rm}}$	backroom	quotient	ditto	hoot	grandpa
$\mathbf{H_{rt}}$	backroom	polka	sinker	vantage	centaur
$\mathbf{H_{lh}}$	hertz	airway	amber	armband	arson
$\mathbf{H_{lm}}$	lifeblood	gab	throwback	ahoy	airmail
$\mathbf{H_{lt}}$	gab	rye	butane	piggy	beep

Tables 7 through 14 suggest many things. First, the OANC and BNC contain different types of texts. In particular, the diversity estimates from the OANC reveal a strong bias for scientific journal writing. Second, typicality for nouns is likely strongly related to concreteness for the OANC, but not necessarily the BNC. There is also very little overlap in the top/bottom words for diversity or typicality across the two varieties. These differences reinforce the necessity of distinguishing the measures by dialect, a point which has been echoed repeatedly in the British tradition of corpus linguistic research on variation (e.g., Pace-Sigge, 2013).