



**HAL**  
open science

## Weaving an ambiguous lexicon

Isabelle Dautriche

► **To cite this version:**

Isabelle Dautriche. Weaving an ambiguous lexicon. Linguistics. Université Sorbonne Paris Cité, 2015. English. NNT : 2015USPCB112 . tel-01541510

**HAL Id: tel-01541510**

**<https://theses.hal.science/tel-01541510>**

Submitted on 19 Jun 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITÉ PARIS DESCARTES

**École doctorale Frontière du Vivant (ED 474)**

*Laboratoire de Sciences Cognitives et Psycholinguistique*

*Département d'Études Cognitives*

*École Normale Supérieure*

## **Weaving an ambiguous lexicon**

**Par Isabelle DAUTRICHE**

Thèse de doctorat de Sciences cognitives

Dirigée par Anne CHRISTOPHE

and

co-encadrée par Benoît CRABBÉ

Présentée et soutenue publiquement le 18 septembre 2015

Devant un jury composé de :

Anne CHRISTOPHE	[directeur]	- Directrice de Recherche, CNRS, École Normale Supérieure
Benoit CRABBÉ	[co-encadrant]	- Maître de Conférences, Université Paris Diderot
Padraic MONAGHAN	[rapporteur]	- Professeur, University of Lancaster
John TRUESWELL	[rapporteur]	- Professeur, University of Pennsylvania
François PELLEGRINO	[examinateur]	- Directeur de Recherche, CNRS, Université de Lyon 2
Kim PLUNKETT	[examinateur]	- Professeur, University of Oxford



# Abstract

Modern cognitive science of language concerns itself with (at least) two fundamental questions: how do humans learn language? —the *learning problem* —and why do the world’s languages exhibit some properties and not others? —the *typology problem*. Though the relation between language acquisition and typology is not necessarily one of equivalence, there are many points of contacts between these two domains. On the one hand, children work on language through an extended period of time and their progression could plausibly reveal aspects of the cognitive blueprint for language. On the other hand, paying attention to the structural commonalities of languages can clue us in to what the human learning mechanism producing these preferences must look like. These questions, although complementary, represent different approaches of understanding the features of cognition underlying the language faculty and have often been dealt with separately by different research communities.

In this dissertation, I attempt to link these two questions by looking at the lexicon, the set of word-forms and their associated meanings, and ask why do lexicons look the way they are? And can the properties exhibited by the lexicon be (in part) explained by the way children learn their language? One striking observation is that the set of words in a given language is highly ambiguous and confusable. Words may have multiple senses (e.g., homonymy, polysemy) and are represented by an arrangement of a finite set of sounds that potentially increase their confusability (e.g., minimal pairs). Lexicons bearing such properties present a problem for children learning their language who seem to have difficulty learning similar sounding words and resist learning words having multiple meanings. Using lexical models and experimental methods in toddlers and adults, I present quantitative evidence that lexicons are, indeed, more confusable than what would be expected by *chance* alone (Chapter 2). I then present empirical evidence suggesting that toddlers have the tools to bypass these problems given that ambiguous or confusable words are constrained to appear in distinct *context* (Chapter 3). Finally, I submit that the study of ambiguous words reveal factors that were currently missing from current accounts of word learning (Chapter 4). Taken together this research suggests that ambiguous and confusable words, while present in the language, may be restricted in their distribution in the lexicon and that these restrictions reflect (in part) how children learn languages.





## Résumé

Il y a (au moins) deux questions fondamentales que l'on est amené à se poser lorsqu'on étudie le langage: comment acquiert-on le langage? —le *problème d'apprentissage* —et pourquoi les langues du monde partagent certaines propriétés mais pas d'autres? —le *problème typologique*. Bien que l'acquisition du langage n'explique pas directement la typologie des langues, et vice-versa, il existe de nombreux points de contacts entre ces deux domaines. D'une part, la manière dont les enfants développent le langage peut être informative sur les aspects cognitifs générant l'existence de telle ou telle propriété dans les langues. D'autre part, étudier les propriétés qui sont communes à travers les langues peut nous éclairer sur les spécificités du mécanisme de l'apprentissage humain qui conduisent à l'existence de ces propriétés. Ces deux questions ont souvent été traitées séparément par différents groupes de recherche, bien qu'elles représentent une approche complémentaire à l'étude des caractéristiques cognitives qui sous-tendent la faculté du langage.

Dans cette thèse, j'entreprends de relier ces deux domaines en me focalisant sur le lexique, l'ensemble des mots de notre langue et leur sens associés, en posant les questions suivantes: pourquoi le lexique est-il tel qu'il est? Et est-ce que les propriétés du lexique peuvent être (en partie) expliquées par la façon dont les enfants apprennent leur langue? Un des aspects les plus frappants du lexique est que les mots que nous utilisons sont ambigus et peuvent être confondus facilement avec d'autres. En effet, les mots peuvent avoir plusieurs sens (par exemple, les homophones, comme "avocat") et sont représentés par un ensemble limité de sons qui augmentent la possibilité qu'ils soient confondus (par exemple, les paires minimales, comme "bain"/"pain"). L'existence de ces mots semble présenter un problème pour les enfants qui apprennent leur langue car il a été montré qu'ils ont des difficultés à apprendre des mots dont les formes sonores sont proches et qu'ils résistent à l'apprentissage des mots ayant plusieurs sens. En combinant une approche computationnelle et expérimentale, je montre, quantitativement, que les mots du lexique sont, en effet, plus similaires que ce qui serait attendu par *chance* (Chapitre 2), et expérimentalement, que les enfants n'ont aucun problème à apprendre ces mots à la condition qu'ils apparaissent dans des *contextes* suffisamment distincts (Chapitre 3). Enfin, je propose que l'étude des mots ambigus permet de révéler des éléments importants du mécanisme d'apprentissage du langage qui sont actuellement absents des théories actuelles (Chapitre 4). Cet ensemble d'études suggère que les mots ambigus et les mots similaires, bien que présents dans le

langage, n'apparaissent pas arbitrairement dans le langage et que leur organisation reflète (en partie) la façon dont les enfants apprennent leur langue.

# Contents

<b>Abstract</b>	<b>i</b>
<b>Résumé</b>	<b>iii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Ambiguity in the lexicon . . . . .	3
1.1.1 The function of ambiguity in the lexicon . . . . .	6
1.1.2 Ambiguity in language processing . . . . .	8
1.2 Ambiguity: A challenge for language acquisition? . . . . .	10
1.2.1 The segmentation problem . . . . .	11
1.2.2 The identification problem . . . . .	12
1.2.3 The mapping problem . . . . .	15
1.2.4 The extension problem . . . . .	17
1.3 Summary . . . . .	19
<b>2 Quantifying word form similarity in the lexicons of natural languages</b>	<b>21</b>
2.1 Lexical clustering in an efficient language design . . . . .	23
2.1.1 Method . . . . .	27
2.1.2 Results: Overall similarity in the lexicon . . . . .	31
2.1.3 Results: Finer-grained patterns of similarity in the lexicon . . . . .	42
2.1.4 Conclusions . . . . .	49
2.1.5 References . . . . .	53
2.2 Wordform similarity increases with semantic similarity: an analysis of 101 languages . . . . .	58
2.2.1 Method . . . . .	60
2.2.2 Results . . . . .	62
2.2.3 Conclusions . . . . .	68
2.2.4 References . . . . .	70
2.3 Summary and Discussion . . . . .	75
<b>3 Learning confusable and ambiguous words</b>	<b>77</b>
3.1 Learning phonological neighbors: Syntactic category matters . . . . .	79
3.1.1 Experiment 1 . . . . .	80
3.1.2 Experiment 2 . . . . .	83

3.1.3	Experiment 3 . . . . .	84
3.1.4	Conclusions . . . . .	86
3.1.5	References . . . . .	87
3.2	Learning homophones: syntactic and semantic contexts matter . . . . .	89
3.2.1	Experiment 1 - Manipulating the syntactic and semantic distance . . . . .	91
3.2.2	Experiment 2 & 3 - Manipulating the semantic distance . . . . .	96
3.2.3	Experiment 4 - Manipulating the syntactic distance using gender . . . . .	106
3.2.4	Experiment 5 - Manipulating neighborhood density . . . . .	109
3.2.5	Conclusions . . . . .	113
3.3	Similar-sounding words and homophones in the lexicon . . . . .	115
3.3.1	Minimal pairs in the lexicon . . . . .	115
3.3.2	Homophones in the lexicon . . . . .	117
3.4	Summary and Discussion . . . . .	123
<b>4</b>	<b>Theories of word learning and homophony: what is missing, what do we learn</b>	<b>127</b>
4.1	What homophones say about words . . . . .	129
4.1.1	Experiment 1: gap in conceptual space and structure of the lexicon . . . . .	132
4.1.2	Experiment 2: linguistic manipulations . . . . .	139
4.1.3	Conclusions . . . . .	145
4.1.4	References . . . . .	151
4.2	Word learning: homophony and the distribution of learning exemplars . . . . .	153
4.2.1	Experiment 1 . . . . .	156
4.2.2	Experiment 2 . . . . .	163
4.2.3	Experiment 3 . . . . .	167
4.2.4	Conclusions . . . . .	171
4.2.5	References . . . . .	173
4.3	Summary and Discussion . . . . .	181
<b>5</b>	<b>General Discussion</b>	<b>183</b>
5.1	The lexicon: The arena of many functional constraints . . . . .	183
5.2	Ambiguity in context is not a challenge for language acquisition . . . . .	185
5.3	How did the lexicon become the way it is? . . . . .	188
5.4	The influence of external factors on the lexicon . . . . .	189
5.5	Insights into the link between language and mind . . . . .	193
5.6	Conclusion . . . . .	195
	<b>References</b>	<b>197</b>
	<b>A Word forms are structured for efficient use</b>	<b>213</b>
	<b>B Cross-situational word learning in the right situations</b>	<b>231</b>

# 1 Introduction

Language is such a common feature of our daily life that we rarely pause to think about it. It seems as natural for us to see children learn to speak as it is to see them learn to walk. Yet language may be more complex than one may think. While every child in the world walks in the same way, it is clearly the case that they do not speak in the same way: A child born in North India will (likely) learn Hindi while a child born in France will (more than likely) learn French. More over, walking has not evolved much across generations: There are limited variations between the way children and their parents are walking. Languages, on the contrary, seem to vary without established limits, so fast, that within the span of a human life, one can see novel words and expressions coming into daily usage.

Languages are thus complex systems, that not only differ greatly from one another at every level of description (sound, lexicon, grammar, meaning) but also evolve rapidly. Yet, children gain a good understanding of their native language even before they learn to dress themselves alone or brush their teeth. Two-year-olds are capable of learning an average of 10 new words per day without explicit training or feedback and can learn grammatical rules for which there is only scarce evidence in their environment. Thus, it is not surprising that people have spent decades thinking about the *learning problem*: how do children learn languages, despite languages being implemented so differently across the world?

Certainly, the presence of fundamental differences between languages does not imply that languages are unconstrained systems that vary freely. As several scholars have observed, languages also share important similarities. All languages are complex symbolic systems that combine the same units (phonemes, morphemes, words, sentences) to convey a potential infinity of meanings. These properties, inter alia, are listed as *design features* of languages (a term introduced by Hockett, 1969). Besides these core properties, languages also share important statistical tendencies in their surface patterns: properties that occur more often than chance. For instance, subjects tend to precede objects in simple declarative sentences (Greenberg, 1966). Such *universals*, although not observed in all languages<sup>1</sup>, indicate that languages may not be random samples of properties, at least not at an abstract level. A resulting question thus concerns why languages share some properties and not others. To date, this *typology problem* has received less attention than its corresponding

---

<sup>1</sup>Note that most "absolute universals", that is, properties that are universally represented in language, are contested (Evans & Levinson, 2009).

what-question (i.e., what are the properties languages share) that is still heavily debated (e.g., Evans & Levinson, 2009).

The learning and typology problems discussed above are mutually informative of one another. On the one hand, paying attention to the structural commonalities between languages would delineate necessary properties of human languages and thus characterize the constraints on cognitive capacities that humans may bring into the learning problem. On the other hand, the study of language acquisition has to provide mechanisms that will allow children to learn *any* of the world's language, and such general mechanisms could plausibly reveal aspects of the cognitive blueprint for language.

A particularly interesting illustration of the interaction between learning and typology is the distribution of grammatical encoding across languages. In order to interpret a sentence, we need to determine the grammatical roles of the words to understand who did what to whom. There are two major ways in which languages signal syntactic relationships and grammatical roles: word order and case-marking. Slobin & Bever (1982) found that Turkish, English, Italian and Serbo-Croatian children asked to act out transitive sentences of the type "the squirrel scratches the dog" differed in their ability to perform this task. Turkish-speaking children as well as English and Italian-speaking children had no problem determining the meaning of these simple sentences, most likely because of the presence of regular case-marking (in Turkish) and fixed word order (in English and Italian) indicated readily who is doing what to whom. In contrast, Serbo-Croatian children performed poorly, most likely because this language combines a flexible word order with a non-systematic case-marking system. These results show that some properties are harder to learn than others in line with typological data: Most of the world's languages display either a fixed word order or alternatively, a regular case-marking system. This suggests that language properties which are easily learnable proliferate, while others, not easily learnable, remain limited or die out.

Certainly, learning and typology are not directly related (see Bowerman, 2011). Learning is dependent on the maturation of the brain and of other cognitive functions (executive functions, memory, etc) that may shape the learning progress. As such, learning difficulties or learning facilities may not reflect solely cognitive constraints on the linguistic system but also maturational constraints. Conversely, language patterns are not only influenced by learning but also by language usage in mature speakers and by external environmental properties inherent to human culture. In sum, while language acquisition and language typology have a lot of potential to be informative about one another, disentangling their individual contributions is a delicate problem by itself.

There are certain cases, however, where typology and learning can give us insight into one another. These are cases where we know that a given property exists in languages and we also know that such a property is difficult to learn for children. Such cases are particularly

interesting as one can examine what exactly enables children to learn such a property anyway, what the distribution of this property is, both within and across languages, whether it correlates with children's abilities, as well as what other processes this property could account for. Thus, instead of trying to solely explain learning by typology, or typology by learning, one could study both in a complementary approach to understand the features of cognition that underly the language faculty.

In this thesis, I take this complementary approach to look at the lexicon, the set of word forms and their associated meanings, and concentrate on one puzzle of languages: the presence of ambiguity. Ambiguity in the lexicon can arise in two different ways: First because the same phonological form can have multiple meanings (e.g., homophones, "bat" refers to both flying mammals and sport instruments); Second, because word forms may sound very much alike (e.g., "sheep" and "ship") and can easily be confused during language usage. This gives rise to the idea that *phonological proximity* is a concern. The purpose of this dissertation is to examine why lexicons are ambiguous. For this, I will attempt to answer two-sub-questions: How prevalent are ambiguity and confusability in the lexicon? And how can children manage to learn such ambiguous and confusable words? I will then discuss whether the distribution of ambiguity in the lexicon can be (in part) explained by the way children learn their language.

The plan of this introductory chapter is as follows. I start by describing what could be the pros and the cons of an ambiguous lexicon and how this feature of lexicons impacts language processing (Section 1.1). I then turn to the domain of language acquisition and evaluate the impact of phonological proximity and phonological identity on different aspects of learning as well as how these factors challenge current word learning accounts (Section 1.2).

## 1.1 Ambiguity in the lexicon

Sganarelle: - Je veux vous parler de quelque chose  
Pancrace: - Et de quelle langue voulez vous vous servir avec moi?  
Sganarelle: - De quelle langue?  
Pancrace: - Oui.  
Sganarelle: - Parbleu! De la langue que j'ai dans la bouche, je crois  
que je n'irais pas emprunter celle de mon voisin.  
Pancrace: - Je vous dis de quel idiome, de quel langage?  
Sganarelle: - Ah!

— Molière, le mariage forcé

Everybody has once faced a situation where the meaning of an utterance or a word was

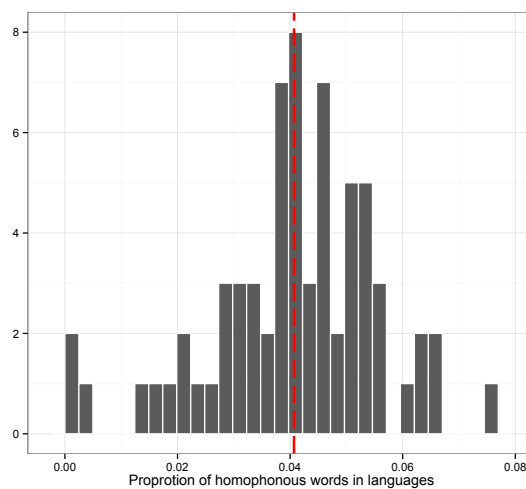


uncertain. We may ask an interlocutor to specify her intended meaning because there is an ambiguity about the meaning the speaker intended to convey (like Sgnararelle, we may wonder whether *langue* means "tongue" or "language" in that context.). However, the frequency of such interventions is typically quite sparse, simply because most of the time we are not aware of the ambiguity of our speech. Consider, for instance, the sentence in French "la grande ferme la porte" (*The big girl closes the door*), each word in this sentence can map onto several meanings: "grande" can refer both to a (*big*) girl and the adjective *big*, "ferme" can both mean the noun *farm* and the verb *to close* and "porte" could either mean the noun *door* or the verb *to carry*. Moreover, even a function word such as "la" is ambiguous in French as it could be an article *the*, or an object clitic as in "Je la ferme" *I close it*. Yet despite the availability of several interpretations for each word in this sentence, it is likely that French listeners processed the sentence without noticing any ambiguity. Keeping this in mind, paying attention to our own productions will cause us to realize that ambiguity is the norm rather than the exception; it is a pervasive property of natural language (Wasow, Perfors, & Beaver, 2005).

Languages are thus full of words that have multiple distinct senses (homophones). To have a rough idea of the magnitude of this phenomena, I used the details of a multilingual encyclopedic dictionary (BabelNet<sup>2</sup>, Navigli & Ponzetto 2012) which readily gives the number of word forms and the number of disjoint senses used in the dictionary. We can read that across the 67 languages represented in Figure 1.1, homophones cover approximately 4% of the words across languages. This number does not include polysemous words, which involve different but related senses (e.g., "café" *coffee* which means, in French, the coffee plant, the drink made of roasted seeds of this plant as well as the place where this beverage is served) nor grammatical morphemes that may be homophonous (e.g., the English morpheme "-s" is used for possessives as well as for plurals).

---

<sup>2</sup>BabelNet is a multilingual encyclopedic dictionary combining resources from WordNet, Wikipedia and other semantic networks. Details and data are available at <http://babelnet.org>. Note that BabelNet covers 271 languages but coverage is poor for most of them and thus do not appear in Figure 1.1

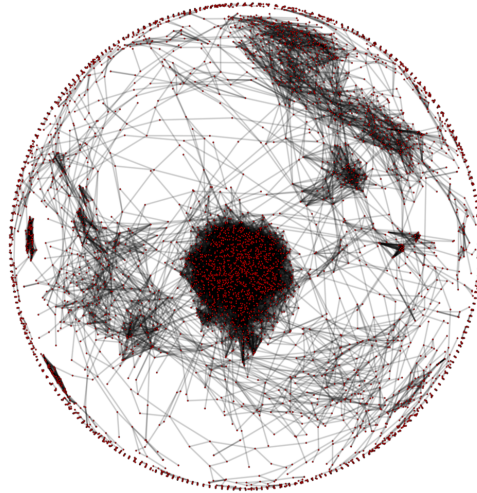


**Figure 1.1:** Distribution of the proportions of homophones across 67 languages; data obtained from BabelNet (Navigli & Ponzetto, 2012). The red dashed line indicates the mean proportion of homophones across languages.

Ambiguity does not only appear in error-free sequences of words, but also arises during normal language use, i.e., conditions where the transmission of the intended message is not always perfect. Speakers may make errors in their productions because they are too emotional (stress, joy), because they do not yet master the language they are speaking in (young children) or simply because they did not plan sufficiently in advance what they were about to say. In addition, listeners may mishear because of inattention or because the message was corrupted due to external factors (e.g., receiving a phone call in the subway, being in a loud environment). In such situations it is likely to misperceive one word as a different, phonologically close, word. This is especially common in songs, for instance, a shared misperception of the song "Purple Haze" of Jimmy Hendrix is to hear "Excuse me while I kiss *this guy*" instead of "Excuse me while I kiss *the sky*".<sup>3</sup> Yet, given that noisy situations prevail in normal language use, it is surprising that we do not confuse words as often as opportunities arise.

To visualize the extent to which word forms may be similar to one another, Figure 1.2 shows a graph in which each word in the English lexicon is a node and any phonological neighbors (i.e., words that are one edit apart like "cat" and "bat") are connected by an edge (as in Arbesman, Strogatz, & Vitevitch 2010; Vitevitch 2008). Words with no or few neighbors tend to be clustered on the outside (the perimeter of the circle). Visually, many words tend to cluster together in the middle of the circle indicating that there are some regions of high phonological density (hence high confusability) in the English lexicon.

<sup>3</sup><https://www.youtube.com/watch?v=PQyGrPw8P50>



**Figure 1.2:** Phonological neighbor network of the English lexicon (constrained to the subset of monomorphemic words). Each red dot is a word, and any two connected words are phonological neighbors.

### 1.1.1 The function of ambiguity in the lexicon

At first sight, ambiguous and confusable words appear to be a great flaw of the linguistic system (Chomsky, 2002), especially for theorists arguing that the shape and properties of language have evolved to optimize communication (Hockett, 1969; Pinker & Bloom, 1990a). If lexicons are efficient solutions to the communicative problem of transmitting information, we expect that language should completely disambiguate meaning and avoid similar-sounding words to make sure that we never misunderstand each other (much like legal texts). At the extreme, this would lead to a language system which maximize *distinctiveness* where each word form would be paired with a single meaning and would be maximally distinct from all other word forms to optimize its recoverability. Suppose for instance a language with a limited phone inventory {b, p, a} in which the only allowed syllables are CV. Intuitively one may start to form words using the shortest forms possible, such that "ba". Yet because this language maximizes the recoverability of its words, "pa" will be disallowed as it is too close (one phoneme difference) to an existing word, leaving us only with one word of two phonemes instead of the two combinations possible within the constraints of that language. To express more meanings, such language needs longer words, for instance: "baba", "papa" but again not "bapa" and "paba" which are too close to existing words. And so on. It is easy to see that a language with a hard constraint for distinctiveness will have many words (as one word can have only one meaning) and long, therefore complex, words (as words need to be distinctive). Certainly, clarity of the signal is only one aspect of an efficient communicative system. An efficient communicative system

must also be composed of simple signals that are easily memorized, produced, processed and transmitted over generations of learners. Simple signals would be frequent, short and composed of common sound sequences. At its limit, the easiest language would be a language maximally *compressible* that uses only one simple word to express all meanings, such as "ba". Certainly, natural languages neither seem to be fully compressible nor distinctive and are likely to be situated on the scale of possible languages existing in-between these two extremes.

The idea that there should be a balance between clarity and simplicity, or distinctiveness and compressibility, is not new (e.g., Piantadosi, Tily, & Gibson, 2012; Shannon, 1947; Zipf, 1949). Zipf formalized the *principle of least effort* which advances that languages are a tradeoff between listeners' and speakers' interests. At the phonological level, listeners want words to be distinctive, while speakers want simple words that minimize articulatory effort and maximize brevity and phonological reduction. At the lexical level, listeners want a large vocabulary size such that each word maps onto a single meaning, while speakers want to reduce the size of the vocabulary to a limited list of simple words that map onto several meanings. This is, in essence, what Zipf's law is about: If one lists all the words of a language by how often they are used, the second most frequent word is about half as frequent as the most frequent one, the third most frequent is about a third as frequent as the most frequent one, the fourth is a fourth as frequent and so on. In addition, frequent words tend also to have more meanings (Zipf, 1949) and to sound more alike (Mahowald, Dautriche, Gibson, & Piantadosi, *submitted*, see Appendix A) than less frequent words. Thus, Zipf's law is consistent with a lexical tradeoff: when speakers tend to choose more frequent words this makes the listener's task harder (as these words are the most ambiguous). By contrast, when listeners find it easier to determine a word's meaning, this means that the speaker had to work harder (as these words will be less frequent and more numerous). Additionally frequent words in languages are simpler than infrequent words: They tend to be short, predictable and phonotactically typical (Mahowald et al., *submitted*; Piantadosi, Tily, & Gibson, 2011; Zipf, 1949). Similarly, by assigning shorter and phonotactically simple forms to more frequent and predictable meanings, and longer and phonetically more complex forms to less frequent and less predictable meanings, languages establish a trade-off between the overall effort needed to produce words and the chances of successful transmission of a message.

Previous quantitative analyses of the lexicon used word frequency as a tool to argue for the presence of functional trade-offs. By showing that frequent words carry different properties than less frequent words, they demonstrate that certain properties are non-uniformly distributed across the words of the lexicon in a way that is consistent with communicative optimization principles. Yet much previous work has focused on simply measuring statistical properties of natural language and interpreting the observed effects. This does not tell us whether the properties we observe are a by-product of chance or are really the manifes-

tation of communicative principles or other cognitive principles associated with language use and language acquisition. If we want to understand the processes that give rise to the observed structure of the lexicon, we need to simulate a range of possible processes to assess which aspects of natural language occur by chance and which are the result of constraining forces. This can be done in (at least) two ways: 1) by simulating the emergence of lexical structure in accelerated lab time (e.g., Kirby, Cornish, & Smith, 2008, using the iterated learning paradigm, I will discuss further this experimental method in the General Discussion) or 2) by using computing power to simulate the generation of lexical structure with different constraints, structures, and biases. In this work, I focus on the latter.

In **chapter 2.1**, I propose to investigate whether the pattern of word form similarity in the lexicon differs from chance and in what direction. As I outlined before, ambiguous and similar-sounding words may be a useful property of natural languages, simply because they allow the re-use of words and sounds which are the most easily understood and produced (Juba, Kalai, Khanna, & Sudan, 2011; Piantadosi, Tily, & Gibson, 2012; Wasow et al., 2005). However, ambiguous and similar-sounding words may be harmful to communication as they increase the chances of being misunderstood. The purpose of this chapter is thus to quantify where natural lexicons are situated along the continuum ranging from fully compressible to fully distinctive lexicons and whether this differs from what we would expect by chance alone.

### 1.1.2 Ambiguity in language processing

Il y a des verbes qui se conjuguent très irrégulièrement.  
Par exemple, le verbe "ouïr".  
Le verbe ouïr, au présent, ça fait : J'ois... j'ois...  
Si au lieu de dire "j'entends", je dis "j'ois",  
les gens vont penser que ce que j'entends est joyeux  
alors que ce que j'entends peut être particulièrement triste.  
Il faudrait préciser: "Dieu, que ce que j'ois est triste !"

— Raymond Devos

In practice, ambiguous and similar-sounding words may not harm communication as strongly as one might think. Indeed, words are rarely uttered in isolation but are part of the broader context in which they are used: the sentence in which they are pronounced, the discourse, the speakers involved, the register of language, the surroundings, etc. In other words, if we are able to integrate other sources of information to constrain the possible meaning of a word, then disambiguation may become (almost) free of cost.

Work on language processing supports this idea as many studies provided evidence that adults use various kind of information to constrain lexical access: verb selectional re-

striction (Altmann & Kamide, 1999), verb structural bias (Trueswell & Kim, 1998), semantic features (Federmeier & Kutas, 1999), event expectations (Kamide, Altmann, & Haywood, 2003), visual environment (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), speakers (Creel, Aslin, & Tanenhaus, 2008), prosody (Millotte, René, Wales, & Christophe, 2008; Millotte, Wales, & Christophe, 2007) or even discourse (Nieuwland & Van Berkum, 2006). As a result, it seems unlikely that one could be confused about the meaning of the word "bat" in a sentence such as "Bats are present throughout most of the world, performing vital ecological roles of pollinating flowers and dispersing fruit seeds." (extracted from Wikipedia).

Predictability has thus emerged as a pivotal factor in human language processing. Adults constantly form expectations about what might occur next in their environment. As a result, the cost of ambiguous and similar-sounding words may be sufficiently lowered, such that these words may not be detrimental for everyday speech comprehension. The fact that people hardly notice ambiguities proves the efficiency of our context-dependent language processing system. Of course, this does not mean that there is no cost associated with the processing of ambiguous and similar-sounding words. Most obviously, since these words' meanings are evaluated in the broader context of the utterance, there is a cost of integrating the context of the word and, perhaps, deciding which meaning is appropriate (e.g., Swinney, 1979; Tanenhaus, Leiman, & Seidenberg, 1979). Yet, the cost of integrating contextual information may be lower for adult listeners than the lexical competition generated by ambiguous words in the absence of context.

The costs and benefits of phonological proximity of words is often (if not always) evaluated in reference to language usage. This approach, however, overlooks language development. This is problematic as language systems are there to be learned. Specifically, previous work looking at the evolution of language highlighted the fact that there is a relationship between the ease with which a linguistic property is learnt and transmitted accurately, and its prevalence across languages (e.g., Boyer & Ramble, 2001; Culbertson & Newport, 2015; Culbertson, Smolensky, & Legendre, 2011; Hudson Kam & Newport, 2009; Kirby, Cornish, & Smith, 2008). Though this may not be the sole explanation of the prevalence of a given property in languages (see Rafferty, Griffiths, & Ettliger, 2013), learnability is a necessary condition for the observation of this property. As a result, the presence of ambiguous and similar-sounding words in the lexicon suggests that these words must be learnable by children – and of course they are, otherwise we would not find them in the lexicon. This does not mean that *any* kind of ambiguity can be learnt by children, but rather that the kind of ambiguous and similar-sounding words that are present in languages must exhibit properties that make them learnable, *inter alia*, and that other kinds of words lacking these properties may be eliminated in the course of language learning. In the following, I will review the challenges that are generated by ambiguous and similar sounding words with

regard to the word learning problem. I will also propose a few properties that would make such words easier to learn.

### 1.2 Ambiguity: A challenge for language acquisition?

Anyone who contemplates lexical acquisition will realize that ambiguous and similar-sounding words present a problem for children learning their language. Normally one can identify a pair of homophonous words if they sound the same but have different meanings. Similarly, two words form a minimal pair if their word forms differ by a single phoneme, but they have different meanings. Of course, this is a relatively easy task for adults because we already know which words are homophonous and which words sound similar to one another. By contrast, young language learners do not initially know which word forms have multiple meanings and which phonological elements are contrastive in their language: They must discover this during the course of language development. In order to examine what kind of challenge these words bring into the word learning game, we first need to define what a word is, what aspects need to be acquired, and how this is done.

What's in a word? By definition, a (content) word is composed of a phonological form that is paired with a concept. For instance the word "cat" is composed of a sequence of sounds, /kæɪt/, which is linked to the concept of CAT. Quite generally, the meaning of a word can be defined by its *extension*, that is the set of entities to which that word refers (e.g., the meaning of "cat" represents the set of all cats and only cats). Yet, we know much more about a word than its meaning: We know that "cat" is a noun, we also know that the meaning of "cat" is more similar to the meaning of "dog" than it is to the meaning of "chair" and we have stored information about contexts (linguistic and non linguistic) in which this word may occur. Thus the knowledge of a word also includes the knowledge of its syntactic properties, its relations with other words in the lexicon and some other types of non-linguistic knowledge, such as the situations in which that word typically occurs (e.g., Perfetti & Hart, 2002).

This brings us to the question of what aspects of words children learn when they "learn words". Recent evidence has shown that as early as the first year of life, infants acquire the meanings of basic nouns and verbs in their language (Bergelson & Swingley, 2012, 2013; Tincoff & Jusczyk, 2012). It is, however, unlikely that all dimensions of word meanings and all of their properties are in place this early on. For instance, it is only during the second year of life that toddlers start to realize how familiar and novel words are related to other words in their lexicon (Arias-Trejo & Plunkett, 2013; Wojcik & Saffran, 2013) and have accumulated knowledge of the grammatical environment in which a word can appear (Bernal, Dehaene-Lambertz, Millotte, & Christophe, 2010). These studies and many others provide growing evidence that children do not fast-map a dictionary-like definition at the



first encounter of a word (Carey, 1978a). Instead, word learning seems to be a slow process, gradually emerging through the accumulation of statistical, syntactic, semantic, and pragmatic fragmental evidence (e.g., Bion, Borovsky, & Fernald, 2013; Carey, 1978a; Gelman & Brandone, 2010; L. Smith & Yu, 2008).

How do children learn the meaning of words? There are, at least, four (non-sequential) problems that children must solve to in order to learn the meaning of words: extracting word forms from the speech signal (*the segmentation problem*), determining what counts as a novel word and what does not (*the identification problem*), determining what it refers to (*the mapping problem*) and determining the relevant concept to which it is associated (*the extension problem*). The fact that many phonological forms are alike or have several meanings potentially create a challenge at each level of the word learning process. I review these challenges in turn below.

### 1.2.1 The segmentation problem

Because words are generally not uttered in isolation, one of the first tasks for infants learning a language is to extract the words that make up the utterances they hear. This is not a trivial task because there is typically no pause between words. Adults are known to rely on their lexicon in order to segment continuous speech into words (e.g., Marslen-Wilson & Welsh, 1978). Because infants in their first year mostly lack such lexical knowledge, this procedure was thought to be unavailable to them, leading researchers to focus on alternative cues that can be recovered from the speech such as statistical regularities (e.g., Saffran, Aslin, & Newport, 1996), phrasal prosody (Gout, Christophe, & Morgan, 2004), phonotactics (e.g., Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993), but also knowledge about the environment such as the broader context of the learning situation (see Synnaeve, Dautriche, Börschinger, Johnson, & Dupoux, 2014, for a computational implementation of this idea).<sup>4</sup>

Certainly lexical ambiguities, such as homophones, are not a challenge for the segmentation problem since there is no ambiguity on the phonological form of the word. In fact, one could even imagine that a highly compressible lexicon with a small vocabulary size (thus with a lot of lexical ambiguities) would be easier to segment since all word forms would occur frequently in speech making it easier to spot their word boundaries. Indeed, when exposed to a stream of speech containing only 6 words (a rather compressible lexicon),

---

<sup>4</sup>Language occurs in context and is constrained by the events occurring in the daily life of the child. For example, during an eating event one is most likely to speak about food, while during a zoo-visit event, people are more likely to talk about the animals they see. These extra-linguistic contexts are readily accessible to very young children and could be used to boost the probability of specific vocabularies and constrain the most plausible segmentation of an utterance (e.g., at meal times, you expect food vocabulary).



infants distinguish between words and non-words after only 2 min of exposure (Saffran et al., 1996), suggesting that a compressible lexicon is easy to segment.

Regarding similar-sounding words, it has been suggested that they can *facilitate* the segmentation process. Indeed, studies have shown that hearing familiar words in the speech signal can help segmenting the speech into words: By 8 months of age, infants segment words more successfully when they are preceded by frequent function words (Shi & Lepage, 2008) or when they are preceded by a very familiar name such as "mommy" (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Accordingly, it might be easier to isolate a novel word (e.g., "tog") from the speech stream when it sounds similar to a known word ("dog"). Supporting that idea, Altvater-Mackensen & Mani (2013) showed that 7-month-old German infants familiarized with a word such as "Löffel" *spoon* in the lab had less difficulty segmenting novel words such as "Löckel" (similar in the onset) or "Nöffel" (similar in the offset) than phonologically unrelated words such as "Sotte". Thus, phonological overlap with a familiarized word may help infants recognize novel similar-sounding words in the speech stream.

Taken together, this suggests that compressible lexicons with reduced vocabulary size and a fair amount of phonological overlap may help infants in the segmentation process. Although this issue will not be further addressed in the main body of this thesis, it is certainly an interesting question to follow up for computational models of segmentation.<sup>5</sup> At any rate, compressible lexicons may be functionally advantageous for speech segmentation.

### 1.2.2 The identification problem

To learn the meaning of a word, children must be able to identify a phonological form as a potential candidate for a novel entry into the lexicon. In the case of similar-sounding words, they must identify that minimally different phonological forms (e.g., "sheep" and "ship") map onto two different meanings, and are not the by-product of the normal sound variability of their language. Indeed, one difficulty for learners is that the same sound categories are realized differently by different speakers of the same language (e.g., Labov, 1966) and differ depending on the phonetic context they appear in (e.g., Holst & Nolan, 1995). As a result, different instances of a given word do not all sound the same. Hence, children must learn to distinguish between the acceptable instances of a word and the instances that do not correspond to that word.

Clearly, the presence of similar-sounding words adds an additional complexity to the word

---

<sup>5</sup>One can imagine creating two artificial corpora: one using a compressible lexicon and the other using a distinctive lexicon, keeping the number and frequency of words constant. The idea would be to evaluate on which corpus the same segmentation model (for instance, Adaptor Grammars, M. Johnson, Griffiths, & Goldwater, 2006) would give the best segmentation result.

identification task. Indeed, several studies have shown that 14-month-old toddlers have a hard time differentiating between novel word forms that differ only in one phoneme (Pater, Stager, & Werker, 2004; Stager & Werker, 1997) (e.g., "bih" and "dih") despite being perfectly able to distinguish the phonemic contrast (e.g., /b/ and /d/). Yet, the same age group can succeed when the novel word forms are presented in a more supportive context (Yoshida, Fennell, Swingley, & Werker, 2009), i.e. if they are embedded in a sentence (Fennell & Waxman, 2006) or if the objects onto which the words map are presented prior to the experiment (Fennell, 2012). This suggests that young toddlers need more support to attend and encode minimally different forms. However, the presence of supportive context does not suffice in all conditions: Older 19-month-olds fail to use a single-feature phonological distinction to assign a novel meaning to a word form that sounds similar to a very familiar one (Swingley & Aslin, 2007) (e.g., learning a novel word such as "tog" when having "dog" in their lexicon). Again, perceptual discrimination between the familiar and the novel label was not an issue as children of the same age are able to distinguish familiar words from mispronounced variants (e.g., Mani & Plunkett, 2007; Swingley & Aslin, 2002). In sum, this suggests that young children have difficulty in attending to and encoding similar word forms compared to more distinct word forms.

In the case of ambiguous words, such as homophones, the problem is even more complex since the phonological form alone does not indicate whether a new lexical entry is appropriate for the phonological form. For example, to learn the word "bat", a child must observe several instances of the word "bat" referring to animal bats and several instances of the word "bat" referring to baseball bats. Let us assume, for the sake of simplicity, that a given child has parents whose passion is spelunking. It is likely that this child will encounter a greater proportion of animal-bat situations than baseball-bat situations early in life. As a result, (s)he may have already linked "bat" to animal-bat.<sup>6</sup> Yet, how does the learner know that a novel lexical entry for the word "bat" is appropriate during baseball-bat instances?

Previous research showed that preschoolers perform poorly on tasks requiring them to assign a different, unrelated meaning to a known word (e.g., learning that "snake" could also refer to a novel object that is not a snake) compared to learning a novel meaning for a novel word form (e.g., learning that "blicket" refers to a novel object) (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997). One possible interpretation of these findings is that homophones are learnt later in language development because children have yet to possess the required skills to learn them. However, this seems unlikely: In another study we indirectly<sup>7</sup> showed that children, as young as 28 months of age, know the meaning of several homophonous pairs (such as "porte" meaning either *door* or *carry*, de Carvalho,

---

<sup>6</sup>This situation may be true for a number of homophones which are often described as having a primary meaning (the most frequent) and (a) secondary meaning(s) (the least frequent(s)).

<sup>7</sup>The aim of the study was not to evaluate children's knowledge of homophones but used noun-verb homophonic pairs to show that children can use prosody to constrain their syntactic analysis of sentences.

Dautriche, & Christophe, 2014, 2015). Another possibility is that children need more evidence before accepting a new meaning for a known word than what is provided by experimental protocols because they already possess an established meaning for the tested word form (e.g., "snake"). At any rate, this suggests that identifying whether a new meaning is appropriate for a known word adds a degree of difficulty that children do not run into when learning non-homophones.

Both similar-sounding and ambiguous words challenge learners with the same basic problem: Identifying what counts as a novel word and what does not. In the case of similar-sounding words, children must not only be able to distinguish different phonological forms (e.g., "tog" vs. "dog") but they must also be able to interpret that such a phonological distinction could be indicative of a novel word. In the case of homophones, the problem is even more complex since the same phonological form is used to label several meanings and phonemic information thus cannot be used as a cue to distinguish them (see however Conwell & Morgan, 2012; Gahl, 2008). So how do children eventually succeed at learning these words? As I highlighted earlier, there are many important factors other than phonology in interpreting speech. Adults are sufficiently attuned to their language to attend to the relevant context (i.e., linguistic, visual, pragmatic, etc.) for disambiguating ambiguous words and minimizing the risk of confusing similar-sounding words. Yet, it is an open question whether children, like adults, also take a broader range of contextual information into consideration when judging the likelihood that a word form is attached to a novel meaning.

In **chapter 3**, I investigate whether toddlers' ability to learn similar-sounding words (learning "tog" when "dog" is already in their lexicon) and homophones (learning a second meaning for "dog") depends on the context these novel words are presented in. In particular, I propose that phonological proximity and at the extreme end, phonological identity, with a known word does not impede learning, as long as the novel and the known words appear in distinct *syntactic* or *semantic* contexts. For instance, homophones may be easier to learn when their meanings are sufficiently distant syntactically (e.g. "an eat" may be a good label for a novel animal), or semantically (e.g. "a potty" as a new label for a novel animal), but not when they are close (e.g. "a cat" for a novel animal). In other words, presenting a similar or a known word form in a context that is distinct from its original use may eliminate the possibility that the original meaning was intended and thereby boost the likelihood that a novel meaning was intended. To test this, French 18- to 20-month-old toddlers were taught novel words that are phonologically similar or identical to familiar words and manipulated the words' syntactic and semantic distance to their familiar competitors. The results of this study were subsequently used to evaluate whether the dimensions that make similar-sounding words and homophones easier to learn are indeed reflected in the organization of these words in the lexicon (which would suggest that a learning pressure may have been applied to the evolving lexicon).

### 1.2.3 The mapping problem

Upon hearing a novel word (and having identified it as such), children must determine what the referent of the word is. Imagine that the child is exposed to the word "apple" in a situation where (s)he's eating one in the kitchen. What can the child infer about the word in such a situation? The response is straightforward: not much in the absence of other information. There are a number of objects, relations, properties that could be a potential match for "apple" in that single situation. This raises the question of how children learn to map the meaning of a word onto its label. One possible mechanism that has been proposed to reduce referential ambiguity involves keeping track of semantic properties that are constant across all contexts in which that word occurs. Imagine now that the learner hears the word "apple" while (s)he is eating one in the kitchen, but also when (s)he sees some in a grocery store and then, as (s)he's looking at a picture of an apple in a book. The basic idea is that different situations will help the learner to keep track of what remains invariant across these situations (i.e., the apple), a process referred to as *cross-situational learning* (Akhtar & Montague, 1999; Pinker, 1989; Siskind, 1996).

Experimental evidence has shown that both adults and infants successfully converge toward the correct meaning of the word after several individually ambiguous exposures (L. Smith & Yu, 2008; Trueswell, Medina, Hafri, & Gleitman, 2013; Vouloumanos & Werker, 2009; Yu & Smith, 2007) though the time needed to converge depends on the referential ambiguity of each learning situation (Medina, Snedeker, Trueswell, & Gleitman, 2011; K. Smith, Smith, Blythe, & Vogt, 2006) and the type of label considered (e.g., object vs. action labels Monaghan, Mattock, Davies, & Smith, 2014). Thus, information about the word's referent can be extracted from the environmental statistics of its use. Yet, exactly *how* learners use these statistics is a subject of debate.

Some work suggests that learners accumulate evidence about multiple candidate referents for a given word (*accumulative account*, e.g., L. Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu & Smith, 2007). That is, each time a new word is uttered, children entertain a whole set of situationally plausible referents and learning entails pruning the potential referential candidates as new instances of the word cause some of these candidates to be implausible by the situation. Other evidence suggests that learners maintain a single hypothesis about the likely referent of the word (*hypothesis testing account*, e.g., Medina et al., 2011; Trueswell et al., 2013). Based on a single exposure to a given word, children select the most plausible interpretation of this word. As new information becomes available in subsequent learning situations, this hypothesis may be confirmed or falsified. In the case of falsification, the old candidate referent is replaced by a new one.

That being said, the crucial question for the present work is whether the presence of similar-sounding or ambiguous words affects such learning process(es). Regarding similar-

sounding words, the mapping problem is very much linked to the problem of identification. As long as children have the means to identify that "tog" and "dog" are distinct words (see **chapter 3** for an exploration of what can make two words more distinct) then this cross-situational process should work just as well as for phonologically unrelated words (e.g., "dog", "spoon") (Escudero, Mulak, & Vlach, 2015, for evidence that adults learn minimally different words such as "pix"/"pax" just as well as words that are more different, across multiple ambiguous situations.).

The presence of ambiguous words, by contrast, induces an additional difficulty, not only for the learner but also for cross-situational word learning accounts. Imagine a child learning the word "bat." (S)he might be observing several situations involving an animal-bat and several situations involving a baseball-bat (and probably also some situations where neither of these items is available). At the end of the day, an accumulative learner will end up with a lot of evidence that "bat" could be associated with both animal-bats and baseball-bats, since such a learner can entertain several possible referents for a given word.<sup>8</sup> Yet, if word learning is best explained by a hypothesis-testing account, there is no possible way to explain how the meaning of homophones might eventually be learned. Recall that according to this view, word-referent mapping involves a one-to-one association which gets updated until it reaches a stable adult stage. Thus, in one particularly informative situation of animal-bat, the child may guess that the most likely referent for "bat" is animal-bat, yet when encountering baseball-bat situations it is likely that the child will have to change its best guess for a baseball-bat, and so on. In sum, the learner will keep on oscillating between the two referents of the word without ever forming a stable word-referent association.

Logically speaking, the presence of homophony suggests that learners *must* be able to entertain at least a few plausible referents for a word. One may imagine other learning strategies to accommodate the finding that learners encode more than a single meaning hypothesis. For instance, Koehne, Trueswell, & Gleitman (2013) proposed a multiple-hypothesis tracking strategy, according to which learners may memorize not only one hypothesis, but all past hypotheses for a given word (see also Stevens, Trueswell, Yang, & Gleitman, submitted). Importantly, this suggests that both strategies, accumulative and hypothesis testing, may be two extreme cases along a continuum of learning strategies, between remembering every possible occurrence and remembering only the one that is being entertained (see also Yurovsky & Frank, under review). Including more and more complex word learning phenomena, such as homophony, will be thus help us advance towards more realistic word learning accounts (see also the *extension problem*).

Simply looking at co-occurrence statistics between words and their potential referents may

---

<sup>8</sup>Note, however, that it does not mean that they have formed two separate lexical entries that happen to be homophones – I come back to this problem when tackling the *extension problem* of word learning below.

not be enough to account for homophony.<sup>9</sup> As discussed before, learners must be able to distinguish between different meanings of the same word form (*the identification problem*). This might be possible if learners encode not only form-referent co-occurrences but also other kinds of information. In previous work, I suggested that cross-situational learning is informed by the type of learning context (see Appendix B, Dautriche & Chemla, 2014). This idea rests on the observation that in the kitchen, one is more likely to speak about food than in the bathroom and the opposite holds for bath items. As such, the extra-linguistic context, which is naturally available to young children, may help learners in constraining the likely referent of the word in a given situation. If, as I suggested earlier, a sufficiently large semantic distance is a major characteristic of homophone pairs in language, then we expect homophones to appear in clearly different contexts. For instance, animal-bats and baseball-bats cover distant concepts, and it is likely that the situations in which animal-bats are mentioned (caving, garden at night) are quite different from the situations in which baseball-bats are mentioned (sport event). Thus, if learners retain contextual information, they may be able to "tag" different instances of the homophonous word, which may help them realize that they should track two separate referents instead of just a single one.

In sum, homophony challenges current word learning accounts on important grounds that need to be addressed to tackle more complex, and more ecologically valid, word learning phenomena. For the learner, the presence of homophones, as well as similar-sounding words, presents similar challenges for mapping as it does for identification: Finding the correct referent(s) of a word may be facilitated if learners are not overwhelmed by phonological identity, or proximity, and use other types of information, for instance the broader context of the learning situation, to constrain their word-referent hypotheses. Whether or not the distribution of ambiguous and similar-sounding words in the lexicon allows for such distinction will be addressed in **chapter 3**.

#### 1.2.4 The extension problem

At the same time as children determine the likely referent of a word, they must also determine the relevant *extension* associated with the word (i.e. the subset of entities to which a given word refers). In general, children will not observe *all* the entities that exhaust the set of candidate exemplars, but will rather observe only a subset of those. Suppose for instance, that the child observes that the word "cat" is used to refer to their pet. Ideally, the child should be able to extend the word to other cats, not just their own one. Yet, even assuming that the child understood which object the word refers to, in that context (*the mapping problem*), the meaning of "cat" is still underspecified after this single learning

---

<sup>9</sup>Nor for word learning in general. Recall that learning the meaning of a word is more than just establishing a link between a form and a referent, but also involves gaining knowledge about its syntactic properties and the situations in which this word may occur.

situation (Quine, 1960). Certainly "cat" could refer to the set of all cats and only cats but many other extensions are compatible with that one experience: Felix the cat, one of its body parts, Felix the cat at 3pm in the kitchen, the set of all black cats, the set of all pets, the set of all animals and so on. In sum, the learner must decide between a number of nested and overlapping possibilities.

Many theoretically plausible meanings can be ruled out simply because children do not, or cannot, consider them (e.g., Felix the cat at 3pm in the kitchen). Yet, many plausible hypotheses still remain. This has led researchers to hypothesize that children come equipped with constraints or biases about which hypotheses are more likely than others. A sensible way to characterize the constraints necessary for word learning is to observe how children extend words in controlled environments. In experiments testing this, children are usually shown several learning exemplars (e.g., "these are blickets") and asked whether the word could be extended to new objects (e.g., "which one of these is a blicket?"). Using such a paradigm, it has been shown that children are more likely to treat novel labels as referring to objects of the same kind (*the taxonomy constraint*, Markman & Hutchinson, 1984), or of the same shape (*the shape bias*, Landau, Smith, & Jones, 1988). Moreover, even young infants of 10 months expect that a word labels a group objects that share a common property (Plunkett, Hu, & Cohen, 2008). Accounts of word learning (associative learning accounts, e.g., Regier, 2005; Yu & Smith, 2007; hypothesis elimination accounts, e.g., Pinker, 1989; Siskind, 1996; Bayesian accounts Frank, Goodman, & Tenenbaum, 2009; Piantadosi, Tenenbaum, & Goodman, 2012; Xu & Tenenbaum, 2007) accordingly presuppose that the extension of a word is *convex* in conceptual space. That is, if two objects A and B are both labelled by the word "blicket", then A and B are exemplars of a single concept whose members are contiguous in conceptual space (Gärdenfors, 2004).

However this approach will fail as soon as identity of form does not imply identity of meaning, that is when the language contains words that have multiple meanings such as homophones. For example, if a child is learning the word "bat", (s)he might be observing several exemplars of animal-bats and several exemplars of baseball-bats. Thus, if both animal-bats and baseball-bats are thought to be exemplars of a single concept (the category including animal-bats and baseball-bats), then such a concept should encompass the common properties of both flying mammals and sport instruments, leading to very broad interpretations such as "thing" or "stuff". Thus, reasoning across different exemplars based on forms that are homophonous will lead to an overextension of the label, since all "things" are not "bats". Note that contrary to the *mapping problem*, here the issue is not about knowing that a word can apply to several referents but knowing that these referents belongs to distinct meanings and not a single one.

There are two possibilities that may explain why current word learning accounts have difficulty in dealing with homophony. The first possibility is that they simulate children's prior knowledge: Children may start with an expectation that the structure of the language is



*clear*, that is, involving transparent, uniform mappings between forms and meanings, leading to one-to-one correspondence across these domains (Slobin, 1973, 1975). The second possibility is that these priors reflect only what we know about very narrow word learning phenomena (learning unambiguous object labels) and cannot possibly explain the learning of other phenomena, such as homophones. At any rate, the existence of homophony in languages suggests that children must be able to entertain the possibility that other form-meaning representations, besides a one-to-one correspondence between forms and meanings, are possible.

In **chapter 4**, I provide the first careful examination of what homophones can tell us about word learning from a theoretical and an experimental standpoint, using both an adults and a child population. On the experimental side, I explore the circumstances under which learners accept several meanings for the same word form. In particular, I explore whether a word is more likely to yield homophony when it is learnt from exemplars clustered at two different positions in conceptual space than when exemplars form a uniform group. For instance, to learn "bat", learners will observe several exemplars of animal-bats and several exemplars of baseball-bats, but no non-bat exemplars. I hypothesize that such a distribution of learning exemplars will cue learners that they are in the presence of a homophonous word. Note that if homophones cover distant concepts (see *the identification problem*), we expect a significant gap in conceptual space between the exemplars of a homophone, which may increase the ease with which learners are able to learn homophones. In essence, **chapter 4** formalizes the intuition that *distinctiveness* in meaning is an important factor for the discovery of homophony. On the theoretical side, I illustrate, with the example of homophony, how making the prior assumptions of word learning accounts explicit provides the best means to identify irreducible assumptions the learning system may rely on.

## 1.3 Summary

The presence of ambiguity in languages is a problem at most levels of word learning (see, however, a potential advantage during *the segmentation problem*). At first sight, this may be taken as a demonstration that languages are not influenced at all by children's learning difficulties. Yet, as many have suggested, languages are a trade-off between several functional pressures competing for opposite properties. This suggests two non-mutually exclusive explanations: 1) the presence of ambiguity in languages may be the consequence of other functional pressures not related with the acquisition of words (e.g., cognitive constraint on word usage) and 2) the kind of ambiguity that is present in the lexicon is *learnable*; that is learning may exercise some more fine-grained influence on the distribution of ambiguity in the lexicon by keeping only ambiguous words of the learnable kind. I explore these ideas by first quantifying the amount of word form similarity in the lexicon (**chapter**



2). I then explore the question of what kind of ambiguity may be learnable by children (**chapter 3**), and I finish by formalizing how children can learn ambiguous words (**chapter 4**).

## 2 Quantifying word form similarity in the lexicons of natural languages

An important question is whether language is designed such that it can be reduced to properties of the cognitive system. Perceptual distinctiveness has been shown to play an important role in shaping the phonology of languages. For instance, Wedel, Kaplan, & Jackson (2013) show that phoneme pairs that have been merged in the course of language change distinguished fewer minimal pairs than pairs of phonemes that remained contrastive in the language. This suggests that there are constraints favoring less confusable contrasts over more confusable contrasts. These constraints have been argued to derive from communicative efficiency: successful transmission of a message requires listeners to be able to recover what is being said, therefore the likelihood that two distinct words would be confusable should be minimized.

The importance of perceptual distinctiveness in phonology has been widely reported (e.g., Flemming, 2004; Graff, 2012; Lindblom, 1986; Wedel et al., 2013). Yet very few studies have looked at 1) perceptual distinctiveness in the lexicon (i.e., at the word form level); 2) the manifestation of other functional pressures, in particular those which are not only beneficial for the *listener* but also serve the interest of the *speaker* and 3) the interaction of phonological distinctiveness with other factors that are relevant for language processing and acquisition, such as semantic distinctiveness. Here we precisely tackle these three points and ask whether the structure of word form similarity in the lexicon is the result of communicative and cognitive pressures associated with language acquisition and use. On one hand, one might expect that a well-designed lexicon should avoid confusable word forms to satisfy communicative constraints. On the other hand, one might expect that a well-designed lexicon should favor word form similarity to make the lexicon easier to produce, learn and remember.

**section 2.1** proposes a new methodology to investigate whether the structure of word form similarity in the lexicon differs from chance and in what direction. This methodology compares real lexicons against "null" lexicons by creating random baselines that provide a null hypothesis for how the linguistic structure should be in the absence of communicative and cognitive pressures. By simulating the generation of lexical structure without any communicative or cognitive constraint, it is thus possible to quantify whether there is

## 2 Quantifying word form similarity in the lexicons of natural languages

---

*more or less* word form similarity in the lexicons of natural languages compared to the chance level.

In **section 2.2**, I looked at the interaction of word form similarity and semantic similarity in relation with possible functional advantages. To date, with 101 languages in the sample, this is the largest cross-linguistic analysis that offers insight into the processes that govern language learning and use across languages.

## Lexical clustering in efficient language design

Kyle Mahowald\*<sup>1</sup>, Isabelle Dautriche\*<sup>2</sup>, Edward Gibson<sup>1</sup>, Anne Christophe<sup>2</sup> and Steven T. Piantadosi<sup>3</sup>

<sup>1</sup>Department of Brain and Cognitive Science, MIT

<sup>2</sup>Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France

<sup>3</sup>Department of Brain and Cognitive Sciences, University of Rochester

### Abstract

Recent evidence suggests that cognitive pressures associated with language acquisition and use could affect the organization of the lexicon. On one hand, consistent with noisy channel models of language (e.g., Levy 2008), the phonological distance between wordforms should be maximized to avoid perceptual confusability (a pressure for *dispersion*). On the other hand, a lexicon with high phonetic regularity would be simpler to learn, remember and produce (e.g., Monaghan et al., 2011) (a pressure for *clumpiness*). Here we investigate wordform similarity in the lexicon, using measures of word distance (e.g., phonological neighborhood density) to ask whether there is evidence for dispersion or clumpiness of wordforms in the lexicon. We develop a novel method to compare lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or sparse wordforms would be as the result of only phonotactics. Results for four languages, Dutch, English, German and French, show that the space of monomorphemic wordforms is clumpier than what would be expected by the best chance model by a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. This suggests a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonetically distinct as possible.

**Keywords:** linguistics, lexical design, communication, phonotactics,

---

\*These authors contributed equally to this work. For correspondence, e-mail [kylemaho@mit.edu](mailto:kylemaho@mit.edu) or [isabelle.dautriche@gmail.com](mailto:isabelle.dautriche@gmail.com)

### 1 Introduction

de Saussure (1916) famously posited that the links between wordforms and their meanings are arbitrary. As Hockett (1960) stated: “The word ‘salt’ is not salty, ‘dog’ is not canine, ‘whale’ is a small word for a large object; ‘microorganism’ is the reverse.” Despite evidence for non-arbitrary structure in the lexicon in terms of semantic and syntactic categories (Bloomfield, 1933; Monaghan et al., 2014), the fact remains that here is no systematic reason why we call a dog a ‘dog’ and a cat a ‘cat’ instead of the other way around, or instead of ‘chien’ and ‘chat.’ In fact, our ability to manipulate such arbitrary symbolic representations is one of the hallmarks of human language and makes language richly communicative since it permits reference to arbitrary entities, not just those that have iconic representations (Hockett, 1960).

Because of this arbitrariness, languages have many degrees of freedom in what wordforms they choose and in how they carve up semantic space to assign these forms to meanings. Although the mapping between forms and meanings is arbitrary, the particular sets of form-meaning mappings chosen by any given language may be constrained by a number of competing pressures and biases associated with learnability and communicative efficiency. For example, imagine a language that uses the word ‘feb’ to refer to the concept HOT, and that the language now needs a word for the concept warm. If the language used the word ‘fep’ for WARM, it would be easy to confuse with ‘feb’ (HOT) since the two words differ only in the voicing of the final consonant and would often occur in similar contexts (i.e. when talking about temperature). However, the similarity of ‘feb’ and ‘fep’ could make it easier for a language learner to learn that those sound sequences are both associated with temperature, and the learner would not have to spend much time learning to articulate new sound sequences since ‘feb’ and ‘fep’ share most of their phonological structure. On the other hand, if the language used the word ‘sooz’ for the concept WARM, it is unlikely to be phonetically confused with ‘feb’ (HOT), but the learner might have to learn to articulate a new set of sounds and would need to remember two quite different sound sequences that refer to similar concepts.

Here, we investigate how communicative efficiency and learnability trade off in the large-scale structure of natural languages. We have developed a set of statistical tools to characterize the large-scale statistical properties of the lexicons. Our analysis focuses on testing and distinguishing these two pressures in natural lexicons: a *pressure for dispersion* (improved discriminability) versus a *pressure for clumpiness* (re-use of sound sequences). Below, we discuss each in more detail.

#### *A pressure for dispersion of wordforms*

Under the noisy channel model of communication (Gibson et al., 2013; Levy, 2008; Shannon, 1948), there is always some chance that the linguistic signal will be misperceived as a result of errors in production, errors in comprehension, inherent ambiguity, and other sources of uncertainty for the perceiver. A lexicon is maximally robust to noise when the expected phonetic distance among words is maximized (Flemming, 2004; Graff, 2012), an idea used in coding theory (Shannon, 1948). Such

dispersion has been observed in phonetic inventories (Flemming, 2002; Hockett & Voegelin, 1955) in a way that is sensitive to phonetic context (Steriade, 1997, 2001). The length and clarity of speakers' phonetic pronunciations are also sensitive to context predictability and frequency (Bell et al., 2003; Cohen Priva, 2008), such that potentially confusable words are pronounced more slowly and more carefully. Applying this idea to the set of wordforms in a lexicon, one would expect wordforms to be maximally dissimilar from each other, within the bounds of conciseness and the constraints on what can be easily and efficiently produced by the articulatory system. The pressure for dispersion can be illustrated by noting that languages avoid long but similar-sounding words. While English has words like 'accordion' and 'encyclopedia', it does not also have 'accordiom' and 'encyclofedia.' Indeed, a large number of phonological neighbors (i.e., words that are one edit apart like 'cat' and 'bat') can impede spoken word recognition (Luce, 1986; Luce & Pisoni, 1998), and the presence of lexical competitors can affect reading times (Magnuson et al., 2007).

#### *A pressure for clumpiness of wordforms*

Well-designed lexicons must also be easy to learn, remember, and produce. What would such a lexicon look like? In the extreme case, one could imagine a language with only one wordform. Learning the entire lexicon would be as simple as learning to remember and pronounce one word. While this example is absurd, there are several cognitive advantages for processing words that are similar to other words in a speaker's mental lexicon. There is evidence that adults and preschoolers learn novel words occupying phonologically dense areas of the lexicon more readily than novel words from sparser phonological spaces (Storkel et al., 2006). Also, words that have many similar sounding neighbors in the lexicon are easier to remember than words that are more phonologically distinct (Vitevitch et al., 2012) and facilitate production by reducing speech error rate (Stemberger, 2004; Vitevitch & Sommers, 2003) and naming latencies (Vitevitch & Sommers, 2003) (but see Sadat et al. (2014) for a review of the sometimes conflicting literature on the effect of neighborhood density on lexical production). Additionally, words with many phonological neighbors tend to be phonetically reduced (shortened in duration and produced with more centralized vowels) in conversational speech (Gahl, 2015; Gahl et al., 2012). This result is expected if faster lexical retrieval is associated with greater phonetic reduction in conversational speech as it is assumed for highly predictable words and highly frequent words (Aylett & Turk, 2006; Bell et al., 2003). In sum, while words that partially overlap with other words in the lexicon may be difficult to recognize (Luce, 1986; Luce & Pisoni, 1998), they seem to have an advantage for learning, memory and lexical retrieval.

Another source of regularity in the lexicon comes from a correspondence between phonology and semantic and/or syntactic factors. For example, there is evidence that children and adults have a bias towards learning words for which the relationship between their semantics and phonology is not arbitrary (Imai & Kita, 2014; Imai et al., 2008; Monaghan et al., 2011, 2014; Nielsen & Rendall, 2012; Nygaard et al., 2009). Furthermore, it may be preferable for words of the same syntactic category to share phonological features, such that nouns sound like nouns, verbs like verbs, and so on (Kelly et

al., 1992). The presence of these natural clusters in semantic and syntactic space therefore result in the presence of clusters in phonetic space. Imagine, for instance, that all words having to do with sight or seeing had to rhyme with ‘look’. A cluster of ‘-ook’ words would develop, and they would all be neighbors and share semantic meaning. One byproduct of these semantic and syntactic clusters would be an apparent lack of sparsity among wordforms in the large-scale structure of the lexicon.

Another source of phonological regularity in the lexicon is *phonotactics*, the complex set of constraints that govern the set of sounds and sound combinations allowed in a language (Hayes & Wilson, 2008; Vitevitch & Luce, 1998). For instance, the word ‘blick’ is not a word in English but plausibly could be, whereas the word ‘bnick’ is much less likely due to its implausible onset *bn-* (Chomsky & Halle, 1965).<sup>1</sup> These constraints interact with the human articulatory system: easy-to-pronounce strings like ‘ma’ and ‘ba’ are words in many human languages, whereas some strings, such as the last name of Superman’s nemesis *Mister Mxyzptlk*, seem unpronounceable in any language. Nevertheless, the phonotactic constraints of a language are often highly language-specific. While English does not allow words to begin with *gn*, French does. Phonotactic constraints provide an important source of regularity that aids production, lexical access, memory and learning. For instance, words that are phonotactically probable in a given language (i.e., that make use of frequent transitions between phonemes) are recognized more quickly than less probable sequences (Vitevitch, 1999). Furthermore, infants and young children seem to learn phonotactically probable words before learning less probable words (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer listening to high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013).

The upshot of this regularity for the large-scale structure of the lexicon is to *constrain* the lexical space. For instance, imagine a language called *Clumpish* in which the only allowed syllables were those that consist of a nasal consonant (like *m* or *n*) followed by the vowel *a*. Almost surely, that language would have the words ‘ma’, ‘na’, ‘mama’, ‘mana’, and so on since there are just not that many possible words to choose from. The lexical space would be highly constrained because most possible sound sequences are forbidden. From a communicative perspective, such a lexicon would be disadvantageous since all the words would sound alike. The result would be very different from the lexicon of a hypothetical language called *Sparsese* in which there were no phonotactic or articulatory constraints at all and in which any phoneme was allowed. In a language like that, lexical neighbors would be few and far between since the word ‘ma’ would be just as good as ‘Mxyzptlk’.

---

<sup>1</sup>There are many existing models that attempt to capture these language-specific rules. A simple model is an n-gram model over phones, whereby each sound in a word is conditioned on the previous n-1 sounds in that word. Such models can be extended to capture longer distance dependencies that arise within words (Gafos, 2014) as well as feature-based constraints such as a preference for sonorant consonants to come after less sonorant consonants (Albright, 2009; Goldsmith & Riggle, 2012; Hayes, 2012; Hayes & Wilson, 2008).

*Assessing lexical structure*

In this work, we ask whether the lexicon is clumpy or sparse. But, because of phonotactics and constraints on the human articulatory system, a naive approach would quickly conclude that the lexicon is clumpy. Natural languages look more like *Clumpish* than they do like *Sparse* since any given language uses only a small portion of the phonetic space available to human language users.<sup>2</sup> We therefore focus on the question of whether lexicons show evidence for clumpiness or sparsity above and beyond phonotactics in the *overall* (aggregate) structure of the lexicon.

The basic challenge with assessing whether a pressure for dispersion or clumpiness drives the organization of wordform similarity in the lexicon is that it is difficult to know what statistical properties a lexicon should have in their absence. If we believe, for instance, that the wordforms chosen by English are clumpy, we must be able to quantify clumpiness compared to some baseline. Such a baseline would reflect the *null hypothesis* about how language may be structured in the absence of cognitive forces. Indeed, our methods follow the logic of standard statistical hypothesis testing: we create a sample of null lexicons according to a statistical baseline with no pressure for either clumpiness nor dispersion. We then compute a test measure (e.g., string edit distance) and assess whether real lexicons have test measures that are far from what would be expected under the null lexicons. We present a novel method to compare natural lexicons to phonotactically-controlled baselines that provide a null hypothesis for how clumpy or scattered wordforms would be as the result of only phonotactics.<sup>3</sup> Across a variety of measures, we find that natural lexicons have the tendency to be clumpier than expected by chance (even when controlling for phonotactics). This reveals a fundamental drive for regularity in the lexicon that conflicts with the pressure for words to be as phonetically distinct as possible.

## 2 Method

Assessing the extent to which the lexicons of natural languages are clumpy or sparse requires a model of what wordforms should be expected in a lexicon in the absence of either force. Prior studies looking at the statistics of language—in particular Zipf’s law (Mandelbrot, 1958; Miller, 1957)—have made use of a *random typing* model in which sub-linguistic units are generated at random, occasionally leading to a word boundary when a “space” character is emitted. However, this model makes unrealistic assumptions about the true generative processes of language (Howes, 1968; Piantadosi et al., 2013) as the sequences of sounds composing words are not generated randomly but follow complex constraints (Baayen, 1991; Hayes, 2012). To more accurately capture the phonotactic processes

<sup>2</sup>As an illustration, English has 44 phonemes so the number of possible unique 2-phone words is  $44^2 = 1936$ , yet there is only 225 unique 2-phone words in English, thus only 11% of the space possible for two -phone words is actually used in English.

<sup>3</sup>Using a similar approach, Baayen (1991) studied wordform similarity in relation to words’ frequency by simulating lexicons (see also Baayen (2001)’s implementation of the Simon-Mandelbrot model.)



at play in real language, here we built several generative models of lexicons: ngrams over phones, ngrams over syllables, and a PCFG over syllables. After training, we evaluated each model on a held-out dataset to determine which most accurately captured each language. The best model was used as the statistical baseline with which real lexicons are compared. We studied monomorphemes of Dutch, English, German and French. Because our baseline models capture effects of phonotactics, we are able to assess pressures for clumpiness or dispersion over and above phonotactic and morphological regularities.

### 2.1 Real Lexicons

We used the lexicons of languages for which we could obtain reliably marked morphological parses (i.e., whether a word is morphologically simple like ‘glad’ or complex like ‘dis-interest-ed-ness’). For Dutch, English and German we used CELEX pronunciations (Baayen et al., 1993) and restricted the lexicon to all lemmas which CELEX tags as monomorphemic. The monomorphemic words in CELEX were compiled by linguistic students and include all words that were judged to be nondecomposed. For French, we used Lexique (New et al., 2004), and I.D. (a native French speaker) identified monomorphemic words by hand. Note that, for Dutch, French and German, these monomorphemic lemmas include infinitival verb endings (*-er* in French, *-en* or *-n* in German and Dutch).<sup>4</sup> Since it is not clear how to separate homophones from polysemy, we chose to focus on surface phonemic forms: when two words with different spellings shared the same phonemic wordform (e.g., English ‘pair’ and ‘pear’ are both pronounced /per/), we included that phonemic form only once. In order to focus on the most used parts of the lexicon and not on words that are not actually ever used by speakers, we used only those words that were assigned non-zero frequency in CELEX or Lexique, including these words in the simulation however, does not change the results observed. All three CELEX dictionaries were transformed to turn diphthongs into 2-character strings in order to capture internal similarity among diphthongs and their component vowels. In each lexicon, we removed a small set of words containing foreign characters and removed stress marks. Note that since we removed all the stress marks in the lexicons, noun-verb pairs that differ in the position of stress were counted as a single wordform in our lexicon (e.g., in English the wordform ‘desert’ is a noun when the stress is on the first vowel ‘désert’ but is a verb when the stress is on the last vowel ‘desért’ but we use only the wordform /desert/ once). This resulted in a lexicon of 5343 words for Dutch, 6196 words for English, 4121 words for German and 6728 words for French.

### 2.2 Generative models of Lexicons

In order to evaluate each real lexicon against a plausible baseline, we defined a number of lexical models. These models are all generative and commonly used in natural language processing (NLP) applications in computer science. The advantage of using generative models is that we can use the

---

<sup>4</sup>Removing these verb endings and running the same analysis on the roots did not change the results observed for these 3 languages (but see section 4.2 for an analysis where verb endings matter)

set of words of real lexicons to construct a probability distribution over some predefined segments (phones, syllables, etc.) that can be then used to generate words, thus capturing phonotactic regularities.<sup>5</sup> These models are all lexical models, that is, their probability distributions are calculated using word types as opposed to word tokens, so that the phonemes or the syllables from a frequent word like *the* are not weighted any more strongly than those from a less frequent word.<sup>6</sup> We defined three categories of models:

- **n-phone models:** For  $n$  from 1 to 6, we trained a language model over  $n$  phones. Like an  $n$ -gram model over words, the  $n$ -phone model lets us calculate the probability of generating a given phoneme after having just seen the previous  $n-1$  phonemes:  $P(x_i|x_{i-(n-1)}, \dots, x_{i-1})$ . The word probability is thus defined as the product of the transitional probabilities between the phonemes composing the word, including symbols for the beginning and end of a word. For example, the word ‘guitar’ is represented as  $\blacktriangleright$  g i t a: r  $\blacktriangleleft$  in the lexicon where  $\blacktriangleright$  and  $\blacktriangleleft$  are the start and the end symbols. The probability of *guitar* considering a bigram model is therefore:

$$P(\text{g}|\blacktriangleright) \times P(\text{i}|\text{g}) \times P(\text{t}|\text{i}) \times P(\text{a:}|\text{t}) \times P(\text{r}|\text{a:}) \times P(\blacktriangleleft|\text{r})$$

These probabilities are estimated from the lexicon directly. For example  $P(\text{a:}|\text{t})$  is the frequency of *ta:* divided by the frequency of *t*.

- **n-syll models:** For  $n$  from 1 to 2, we trained a language model over syllables. Taking the same example as above, ‘guitar’ is represented as  $\blacktriangleright$  gi ta:r  $\blacktriangleleft$  and its probability from a bigram novel over syllable is:

$$P(\text{gi}|\blacktriangleright) \times P(\text{ta:r}|\text{gi}) \times P(\blacktriangleleft|\text{ta:r})$$

In order to account for out-of-vocabulary syllables in the final log probabilities, we gave them the same probability as the syllables appearing one time in the training set.

<sup>5</sup>Fine-grained models of phonotactics exist for English (e.g., Hayes (2012)) yet adapting them to other languages is not straightforward and there is no common measure that will allow us to compare their performances.

<sup>6</sup>Admittedly, the experience speakers have of real language is token-based, and not type-based. Yet, using token-based probability estimates instead of type-based probability estimates to capture phonotactic regularities does not change the pattern of results for the 4 languages.

- **Probabilistic Context Free Grammar** (Manning & Schutze, 1999, PCFG;): Words are represented by a set of rules of the form  $X \rightarrow \alpha$  where  $X$  is a non-terminal symbol (e.g., Word, Syllable, Coda) and  $\alpha$  is a sequence of symbols (non-terminal and phones). We defined a word as composed of syllables differentiated by whether they are initial, medial, final or both initial and final.

$$\begin{aligned}
 \text{Word} &\rightarrow \text{SyllableI} (\text{Syllable})^+ \text{SyllableF} \\
 \text{Word} &\rightarrow \text{SyllableIF} \\
 \text{Syllable} &\rightarrow (\text{Onset}) \text{Rhyme} \\
 \text{Rhyme} &\rightarrow \text{Nucleus} (\text{Coda}) \\
 \text{Onset} &\rightarrow \text{Consonant}^+ \\
 \text{Nucleus} &\rightarrow \text{Vowel}^+ \\
 \text{Coda} &\rightarrow \text{Consonant}^+
 \end{aligned}$$

These rules define the possible structures for words in the real lexicon. They are sufficiently general to be adapted to the four languages we are studying, given the set of phonemes for each language. Each rule has a probability that determines the likelihood of a given word. The probabilities are constrained such that for every non-terminal symbol  $X$ , the probabilities of all rules with  $X$  on the left-hand side sum to 1:  $\sum P(X \rightarrow \alpha) = 1$ . The likelihood of a given word is thus the product of the probability of each rule used in its derivation. For example, the likelihood of ‘guitar’ is calculated as the product of all probabilities used in the derivation of the best parse (consonant and vowel structures are not shown for simplification):

$$\begin{aligned}
 \text{Word} &\rightarrow \text{SyllableI}(\text{Onset}(\text{g}) \text{Rhyme}(\text{Nucleus}(\text{r}) \text{Coda}(\text{t}))) \\
 &\text{SyllableF}(\text{Rhyme}(\text{Nucleus}(\text{a:}) \text{Coda}(\text{r})))
 \end{aligned}$$

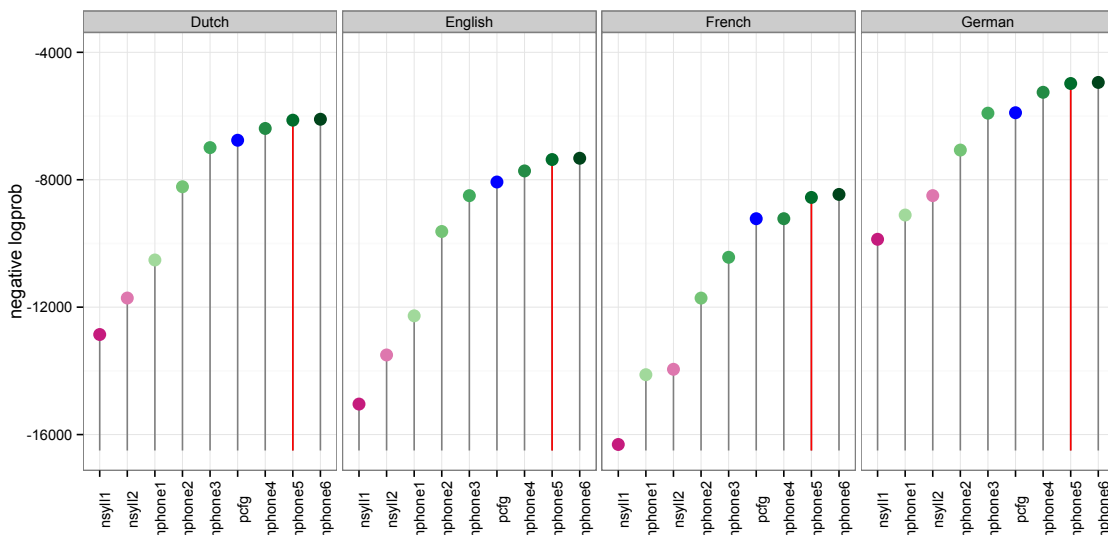
The probabilities for the rules are inferred from the real lexicon using the Gibbs sampler used in Johnson et al. (2007) and the parse trees for each word of the held-out set are recovered using the CYK algorithm (Younger, 1967).

### 2.3 Selection of the best model

To evaluate the ability of each model to capture the structure of the real lexicon, we trained each model on 75% of the lexicon (the training set) and evaluated the probability of generating the remaining 25% of the lexicon (the validation set). This process was repeated over 30 random splits of the dataset into training and validation sets. For each model type, we smoothed the probability distribution by assigning non-zero probability to unseen ngrams or rules in the case of the PCFG. This was to allow us to derive a likelihood for unseen but possible sequences of phonemes in the held-out set. Various smoothing techniques exist, but we focus on Witten-Bell smoothing and Laplace smoothing

which are straightforward to implement in our case.<sup>7</sup> All smoothing techniques were combined with a backoff procedure (though not for the PCFG), such that if the context  $AB$  of a unit  $U$  has never been observed ( $p(U|AB) = 0$ ) then we can use the distribution of the lower context ( $p(U|B)$ ). The smoothing parameter was set by doing a sweep over possible parameters and choosing the one that maximized the probability of the held-out set. The optimal smoothing was obtained with Laplace smoothing with parameter .01 and was used in all models described.

In order to compare models, we summed the log probability over all words in the held-out set. The model that gives the highest log probability on the held-out data set is the best model, in that it provides a “best guess” for generating random lexicons that respect the phonotactics of the language.



**Figure 1:** Each point represent the mean log probability of one model to predict the held-out data set. The nphone models are represented in green, the nsyll models in pink and the PCFG in blue. The 5-phone model has the highest log probability (indicated by a red segment) for all languages. Standard deviation of the mean are represented but too small to be visible at this scale.

As shown in Figure 1, the 5-phone model gives the best result for all lexicons. In all cases, the 6-phone was the next best model, and the 4-phone was close behind, implying that n-phone models in general provide an accurate model of words. The syllable-based models performed particularly poorly. Thus, we focus our attention on the 5-phone model in the remainder of the results, treating this as our best guess about the null structure of the lexicon.

<sup>7</sup>Other smoothing techniques such as Good Turing or Kneser-Ney cannot be implemented easily as they rely on the number of units for which frequency is equal to one which is not available in every model we tested.

### 3 Results: Overall similarity in the lexicon

We use the 5-phone model to generate simulated null lexicons—ones without any pressure for clumpiness or dispersion other than the 5-phone generating process—and study the position of the real lexicon with respect to the simulated ones. For each language, we generated 30 lexicons with the 5-phone model trained on the entire real lexicon. We additionally constrain the generation to ensure that the distribution of word lengths in each simulated lexicon matches the distribution of word lengths in the real lexicon. On average our best lexicon model generated 52% real words for Dutch, 53% for English, 47% for French, and 41% for German. Note that it is not surprising that the best lexicon model generates *only* about 50% of real words since the smoothing parameter allowed the generation of non-words likely to be attested in the language.

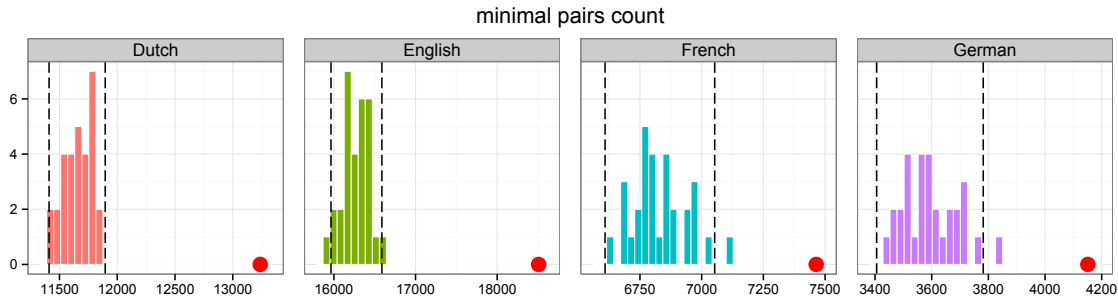
To compare real and simulated lexicons, it is necessary to define a number of test statistics that can be computed on each lexicon to assess how it uses its phonetic space. As in null hypothesis testing, we compute a  $z$ -score using the mean and standard deviation estimated from 30 lexicons generated by our best lexicon model. We then ask whether the real lexicon value falls outside the range of values that could be expected by chance under the null model. The  $p$ -value reflects the probability that the real lexicon value could have arisen by chance under our chosen 5-phone null model.

We present result separately for a number of different measures of wordform similarity.

#### 3.1 Minimal pairs

We first considered the number of minimal pairs present in each lexicon. A minimal pair is a pair of words of the same length for which a single sound differs (e.g., ‘cat’ and ‘rat’). If real lexicons are clumpier than expected by chance, then the real lexicons should have more minimal pairs than their simulated counterparts. If they are more dispersed, the real lexicons will have fewer minimal pairs.

Figure 2 summarizes this hypothesis test, showing how the various simulated lexicons compare to the real lexicons in terms of number of minimal pairs for each language. Each histogram represents a distribution of minimal pair counts broken up by language across the 30 simulated lexicons. The red dot represents the real lexicon value and the dotted lines represent the 95% confidence interval. All histograms fall to the left of the red dot, which suggests that the real lexicon has more minimal pairs than any of the simulated ones in all four languages (all  $ps < .001$ ; see Table 1). This pattern suggests that the real lexicon is clumpier than expected by chance.



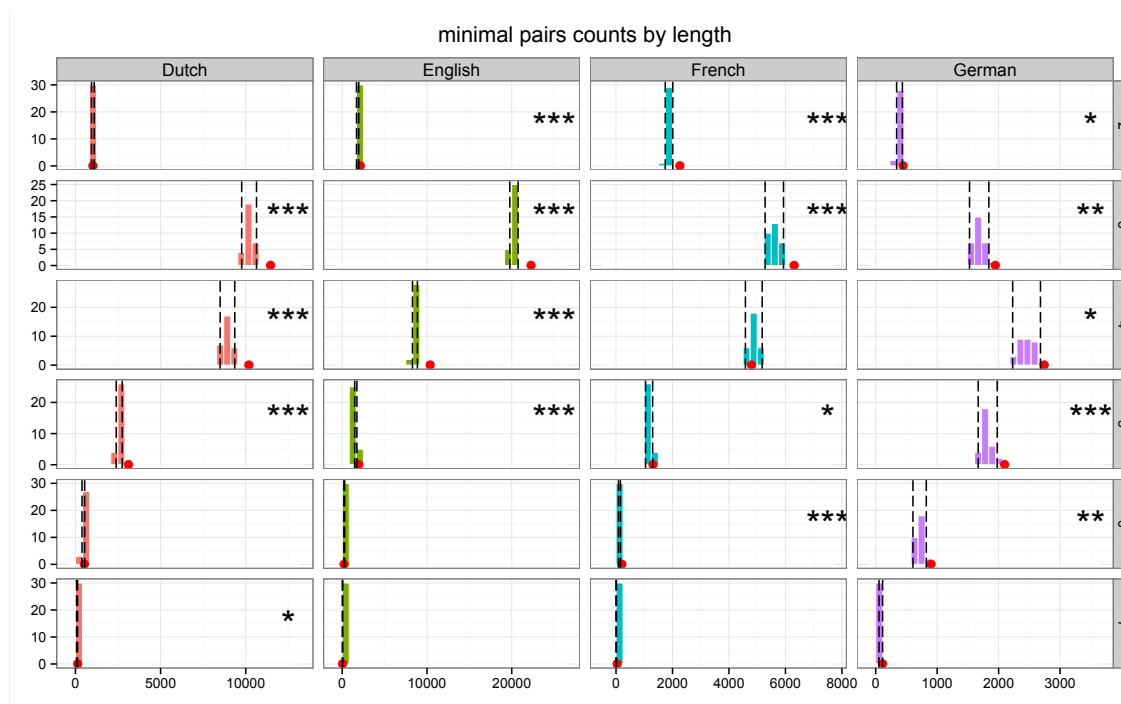
**Figure 2:** Comparison of the total number of minimal pairs for each language (red dot) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. For all four languages, the real lexicon has significantly more minimal pairs than predicted by our baseline.

	Dutch	English	French	German
real	13,237	18,508	7,464	4,151
$\mu$ (simulated)	11,653	16,276	6,830	3,594
$\sigma$ (simulated)	124	159	113	96
$z$	12.77	14.03	5.61	5.80
$p$	<.001	<.001	<.001	<.001

**Table 1:**  $z$ -statistics comparing the total number of minimal pairs in the real lexicon with the chance distribution of mean  $\mu$  and standard deviation  $\sigma$  corresponding to the distribution of minimal pairs counts in the 30 simulated lexicon for each language.

To see whether this effect is driven by words of specific length, we looked at the number of minimal pairs for each length. We concentrated on words of length 2 to 7 which represent more than 90% of all words in each language. As shown in Figure 3, the real lexicon has more minimal pairs than the simulated ones consistently across words of any length. For all languages, the effect is larger for words of smaller length (length 3 to 4; 30 to 50% of all words in each language) where most minimal pairs are observed. The smaller effect for longer words (especially words of length 7 and above) is likely due to a floor effect since longer words are far less likely to have minimal pairs than short words. Note that, for words of length 2, we see a somewhat degenerate case since there are relatively few possible 2-phoneme words, yet for at least 3 languages it appears that there are more minimal pairs of length 2 than what would be expected by chance. This is explained by the smoothing parameter of the model that allows the generation of unseen sequences of sounds (recall that we smoothed the probability distribution to account for rare sequences of sound that may be unseen in the lexicon of monomorphemes).

As a result the model is not exactly reproducing all the 2-phoneme words of the languages.<sup>8</sup>



**Figure 3:** Comparison of the number of minimal pairs by word length (2-7) for each language (red dots) to the distribution of minimal pairs counts across 30 simulated lexicons (histograms). Vertical black lines represent 95% confidence intervals. One star represents  $p < .05$ , two stars  $p < .01$ , and three stars  $p < .001$ .

### 3.2 Levenshtein distance

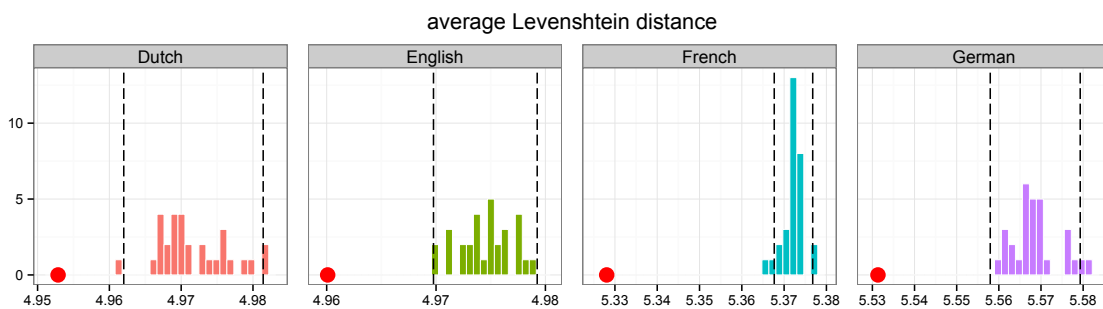
We can evaluate clustering using more global measures by considering the average string edit distance (*Levenshtein distance*) between words (Levenshtein, 1966). The Levenshtein distance between two sound strings is simply the number of insertions, deletions and replacements required to get from one string to another. For instance, the Levenshtein distance between ‘cat’ and ‘cast’ is 1 (insert an ‘s’), and it is 2 between ‘cat’ and ‘bag’ (c → b, t → g). To derive a measure of Levenshtein distance that summarizes the whole lexicon, we compute the *average Levenshtein distance* between words in the lexicon by simply computing the distance between every pair of words in the lexicon and then averaging these distances.<sup>9</sup> If the lexicon is clumpier than expected by chance, words will tend to

<sup>8</sup>Inspection of these 2-phoneme words reveals that most of these words are actual wordforms present in the language (hence attested forms, e.g. "is" in English) but are not counted as distinct monomorphemic lemmas and thus are not included in our real lexicons.

<sup>9</sup> A possible objection to using Levenshtein distances is that there is little apparent difference in phonological confusability between a pair like ‘cats’ and ‘bird’, which has a Levenshtein distance of 4, and a pair like ‘cats’ and ‘pita,’ which has a Levenshtein distance of only 3 but which is arguably even more different since it differs in syllable structure. Ultimately, neither pair is especially confusable: the effects of phonological confusability tail off after 1 or 2 edits.

be more similar to one another and we expect to observe a smaller average Levenshtein distance. In contrast, a larger average Levenshtein distance in the real lexicons relative to the simulated lexicons would suggest that the lexicon is more dispersed than expected by chance.

As shown in Figure 4, the average Levenshtein distance between words is significantly smaller for the real lexicon than in the simulated lexicons for all four languages (see Table 2). The difference is numerically small, but that is to be expected because minimal pairs are statistically unlikely. That is, the edit distance between two words is largely a product of their lengths. For example, on average, the edit distance between two 5-letter words is 5. Nonetheless, the Levenshtein metric provides us with an additional piece of evidence that words in the real lexicons are more similar to each other than what would be expected by chance.



**Figure 4:** Distribution of average Levenshtein distances for each of the 30 simulated lexicons. The red dot represents the real lexicon's value, and the dotted lines are 95% confidence intervals.

	Dutch	English	French	German
real	4.95	4.96	5.32	5.53
$\mu$ (simulated)	4.97	4.97	5.34	5.57
$\sigma$ (simulated)	0.005	0.002	0.002	0.005
$z$	-3.80	-6.0	-6.2	-6.9
$p$	<.001	<.001	<.001	<.001

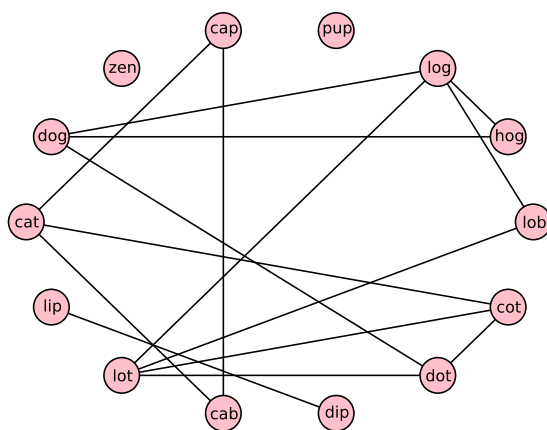
**Table 2:**  $z$ -statistics comparing the average Levenshtein distance in the real lexicon with the chance distribution of mean  $\mu$  and standard deviation  $\sigma$  corresponding to the distribution of average Levenshtein distance in the 30 simulated lexicon for each language.

### 3.3 Network measures

Simply calculating phonological neighbors, however, does not tell us everything about how wordforms are distributed across a lexicon. Perhaps some words have many neighbors while others have few. Or it could be the case that neighbor pairs tend to be more uniformly distributed across the lexicon. To

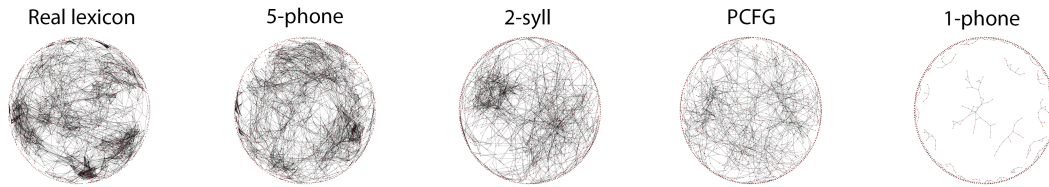


answer these questions, we constructed a phonological neighborhood network as in Arbesman et al. (2010), whereby we built a graph in which each word is a node and any phonological neighbors are connected by an edge, as in the toy example in Figure 5, that shows the situation for a lexicon of 14 words.



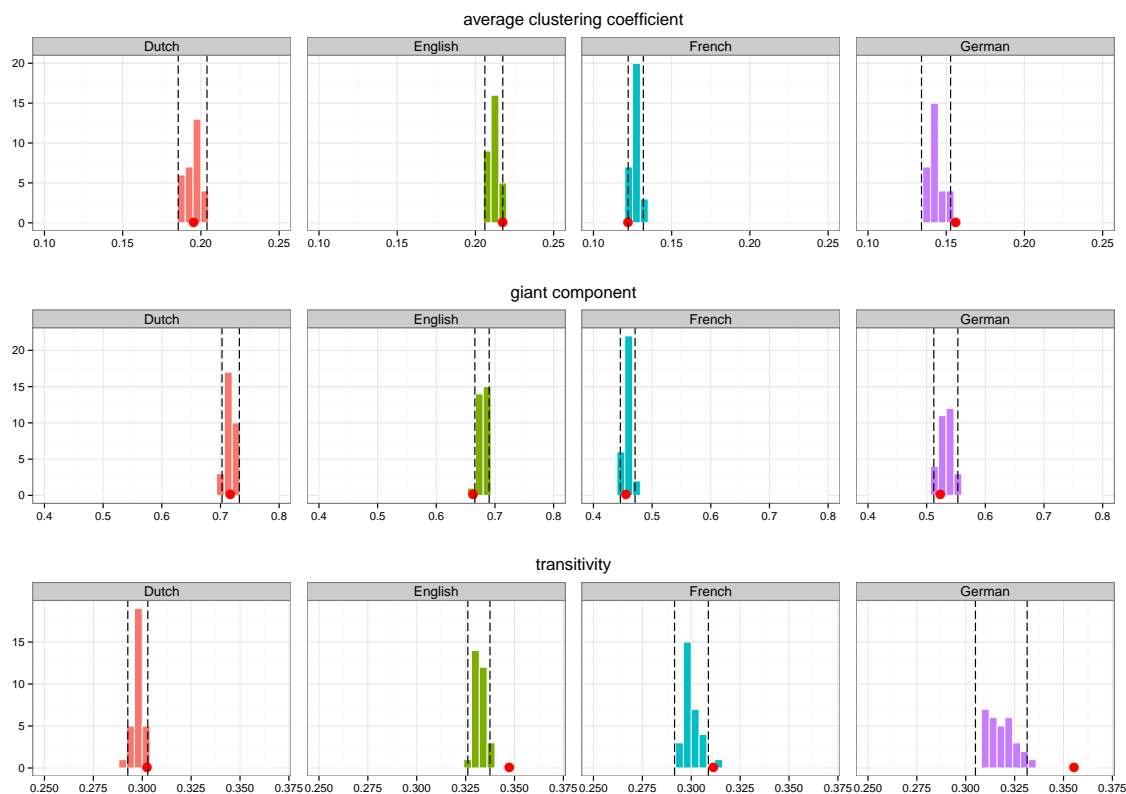
**Figure 5:** Example phonological network. Each word is a node, and any words that are 1 edit apart are connected by an edge.

Figure 6 shows examples of such networks for English 4-phone words, where each word is a node, with an edge drawn between any two words that are phonological neighbors (1 edit away). Words with no or few neighbors tend to be clustered on the outside. (The ring of points around the perimeter of the circle represent the isolates—words with no neighbors.) Words with many neighbors are, in general, plotted more centrally. We compared the shape of lexicons generated by different models to the real lexicon. As can be seen in Figure 6, of all the models, the 5-phone model most closely resembles the real lexicon. Substantially more clustering is observed in the more restrictive generative models: the 5-phone, 2-syll and PCFG models have many more connected neighbors than a 1-phone model. This corresponds to the fact that many more words are possible in the 1-phone model (e.g. ‘ctkw’ is a possible word), than in a more constrained model that respects phonotactics. Therefore the space is largest in the 1-phone model, and the probability of generating a word that is a neighbor of a previously generated word is correspondingly lower. Crucially, however, the real lexicon seems even clumpier overall than the lexicons produced by any of the generative models.



**Figure 6:** Sampling of phonological neighbor network from the different generative models applied on all 4-phone wordforms of the English lexicon. Each point is a word, and any two connected words are phonological neighbors. The simulated lexicons from less constrained generative models are less clustered and have more isolates (words with no neighbors, plotted on the outside ring).

Using techniques from network analysis that have been fruitfully applied to describe social networks and other complex systems (Barabási & Albert, 1999; Wasserman & Faust, 1994; Watts & Strogatz, 1998), we can quantitatively characterize the clustering behavior of the lexicon. We computed the *transitivity*, *average clustering coefficient*, and the percent of nodes in the *giant component*. All three of these measures can be used to evaluate how tightly clustered the words in the lexicon are. A graph's *transitivity* is the ratio of the number of triangles (a set of 3 nodes in which each node in the set is connected to both other nodes in the set) to the number of triads (a set of 3 nodes in which at least two of the nodes are connected). Thus, transitivity in effect asks, given that A is connected to B and B is connected to C, how likely is it that A is also connected to C? The *average clustering coefficient* is a closely related measure that finds the average clustering coefficient across all nodes, where the clustering coefficient of a node is defined as the fraction of possible triangles that *could* go through that node that actually do go through that node. Both values measure the extent to which nodes cluster together. The largest cluster in a network is known as the *giant component*. A network with many isolated nodes will have a relatively small giant component, whereas one in which nodes are tightly clustered will have a large giant component. These measures give us some insight into the internal structure of the lexicon, over and above those obtained by looking at more global measures such as the number of minimal pairs and the average Levenshtein distance. If the real lexicon is clumpier than expected by chance, we predict that, relative to the simulated lexicons, the real lexicons will show higher transitivity, higher average clustering coefficients, and a larger proportion of words in the giant component.



**Figure 7:** Distributions of our best generative model (the histograms) compared to the real lexicon (the red dot) in terms of network measures for lexical networks (where each node is a word and any 2 nodes that are minimal pairs are joined in the network): the percent of nodes in the average clustering coefficient, giant component, and transitivity.

As observed in Figure 7, there is no systematic difference between the real lexicon and the simulated ones regarding the average clustering coefficient measures and the percentage of nodes in the giant component. Yet there is a significant effect of transitivity (see Table 3). The reason that average clustering coefficient shows less of an effect than transitivity is likely that average clustering coefficient is more dependent on low-degree nodes, like the many isolates that exist for longer words in lexical networks (Sporns, 2011). The lack of effect for the giant component measure may simply be because the proportion of words in the giant component is not a particularly robust measure since it can be dramatically shifted by the addition or subtraction of one or two key neighbors. The higher transitivity, however, suggests that in addition to having more overall neighbors in the real lexicons, the neighborhoods themselves are more well-connected than the neighborhoods in simulated lexicons are. That is, if two words A and B are both neighbors of word C, A and B are themselves more likely to be neighbors in the real lexicon than they are in the simulated lexicons.

		Dutch	English	French	German
Average Clustering coefficient	real	0.2	0.22	0.12	0.16
	$\mu$ (simulated)	0.19	0.21	0.13	0.14
	$\sigma$ (simulated)	0.005	0.003	0.002	0.005
	$z$	0.1	2	-2	2.7
	$p$	0.9	.05	.05	<b>&lt;.01</b>
Giant component	real	0.72	0.66	0.46	0.52
	$\mu$ (simulated)	0.72	0.68	0.46	0.53
	$\sigma$ (simulated)	0.008	0.006	0.006	0.01
	$z$	-0.1	-2.4	-0.4	-0.9
	$p$	0.9	<.05	0.7	0.4
Transitivity	real	0.3	0.35	0.31	0.36
	$\mu$ (simulated)	0.3	0.33	0.3	0.32
	$\sigma$ (simulated)	0.003	0.003	0.004	0.007
	$z$	1.8	5.4	2.6	5.5
	$p$	0.07	<b>&lt;.001</b>	<b>&lt;.05</b>	<b>&lt;.001</b>

**Table 3:**  $z$ - statistics comparing various network measure (Average clustering coefficient, proportion of words in the giant component, transitivity) in the real lexicon with the chance distribution of mean  $\mu$  and standard deviation  $\sigma$  corresponding to the distribution of these measures in the 30 simulated lexicon for each language.

### 3.4 Robustness of the results

We chose as our baseline a 5-phone model because it performed best on the cross-validation test. Yet, it is important to note that any pattern of clumpiness or dispersion that we find should occur independently of this specific lexical generation model. To check whether our results were robust across the different measures of wordform similarity, we compared the same measures (minimal pairs count, average Levenshtein distance and network measures) obtained in the 3 best models according to our evaluation (see Figure 1): the 5-phone model, the 6-phone model and the 4-phone model.

## 2 Quantifying word form similarity in the lexicons of natural languages



**Figure 8:** Distributions of a given measure for our best model of word generation (5-phone in dark color), our second best model (6-phone in light color) and our third best model (4-phone in translucent color) compared to the measure in the real lexicons (the red dots) for the four languages and all the measures reviewed so far.

As shown in Figure 8, we find qualitatively similar results with the 3 best models across all the

measures of wordform similarity previously introduced.<sup>10</sup> In general, there were more minimal pairs and lower average Levenshtein distance in the real lexicons than across the three best models. As for the 5-phone model, no conclusive results were obtained for the average clustering coefficient and the giant component measures but the transitivity was higher in the real lexicons than in the three best models of lexicons.

This is evidence that the pattern of clumpiness we found with the 5-phone model is robust across lexical generation models. A pressure for clumpiness is thus visible beyond the particular model of phonotactic probability adopted by the best models produced here.

We also tested whether the German, Dutch, and French infinitival verb endings could be driving clumpiness effects by redoing the analyses above using just root forms (i.e., by removing the infinitival ending from the verbs). One might imagine that, because most verbs end in *-er* in French, for instance, these words have fewer degrees of freedom and thus edit distances will be smaller across the lexicon. In our analysis using just root forms, however, the results were qualitatively the same as when we used lemmas in their infinitive form, likely because the generative models already capture this regularity. That is, our baseline models too have a disproportionate number of words ending in *-er* in French and *-en* in German and Dutch. Because the presence of these infinitival stems does not substantially alter the result, we chose to keep them in the main analysis so as to be consistent with the standard databases we used (CELEX and Lexique).

### 3.5 Interim summary

In general, these measures suggest that the lexicon is clumpy: words tend to be more phonologically similar to each other than would be expected by chance. Word pairs in the real lexicon are more likely to be minimal pairs and more likely to have a small edit distance compared to words in the “chance” simulated lexicons. The chance rate here was determined through *a priori* model comparison of different plausible generating models for words. This technique has allowed us to test for clumpiness vs. dispersion while still respecting the major phonotactic tendencies in each language. It is important to emphasize that these results were computed on monomorphemes, so the results are not an artifact of morphology.

Crucially, the lexicon shows a tendency towards clumpiness above and beyond phonotactics. As we discussed earlier, phonotactic rules themselves can be thought of as a major source of clumpiness in the lexicon, insofar as phonotactics dramatically restricts the space of possible words. Yet, while our best lexical model controls for phonotactic regularity, we still observe clumpiness in the real lexicon compared to this baseline. This suggests additional pressure for clustering beyond just phonotactics.

---

<sup>10</sup>The 3-phone model behaves somewhat differently and in fact shows more clustering than the 5-phone model. But, because its performance on the held-out data set is poor compared to the models shown here, we do not focus on this model.

### 4 Results: Finer-grained patterns of similarity in the lexicon

Across a variety of measures, we found that wordforms tend to be more similar than expected by chance across all languages under study. Yet, while wordform similarity might be explained by a variety of cognitive advantages (see Introduction), it does not necessarily follow that the lexicon is not subject to communicative pressure. A possibility is that the similarity between wordforms may not be uniformly distributed across the real lexicon but may be constrained by other dimensions that maximize their distinctiveness in the course of lexical processing, such as:

1. **phonological distinctiveness:** Not every pair of phonemes is equally confusable. For instance, a minimal pair like ‘cap’ and ‘map’ are unlikely to be confused since /k/ and /m/ are quite distinct. But ‘cap’ and ‘gap’ differ by only the voicing of the first consonant and are thus much more confusable (Miller & Nicely, 1955). From a communicative perspective, this more subtle contrast is potentially much more troublesome for communication and is therefore more likely to be avoided. So even though the number of minimal pairs is higher than expected by chance in natural lexicons, this might not be problematic for communication as long as they are not based on confusable contrasts.
2. **grammatical categories:** Not every pair of words is equally confusable. For instance, nouns (e.g. ‘berry’) are more likely to be confused with other nouns (e.g., ‘cherry’) than words from another grammatical category (e.g., the adverb ‘very’) because they appear in a noun syntactic context which constrains listeners to expect a noun in this position. Therefore, from a communicative point of view, there should be more minimal pairs distributed across syntactic categories than within the same syntactic category to minimize the risk of miscommunication.

In the following we test how the simulated lexicons compare to the real lexicons along these two dimensions.

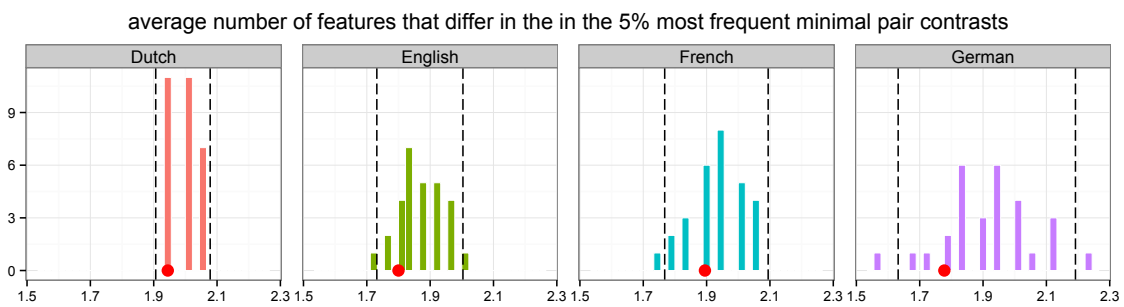
#### 4.1 Wordform distinctiveness in minimal pairs

The accurate recognition of a word depends on the distinctiveness of the phonological contrasts distinguishing words. If lexicons aim to minimize confusability, they should prefer distinctive contrast minimal pairs as opposed to confusable ones. In the case of ‘cap’ and ‘map,’ for instance, one word is unlikely to be confused for the other since the contrast is quite distinctive. But one retains the benefits of being able to re-use the word coda *-ap* in both cases. Thus, it is possible that lexicons can have the learning benefit of having frequent minimal pairs, as long as they are not based on confusable contrasts.

To evaluate this hypothesis, we looked at the 5% most frequent minimal pair contrasts and derived a measure of confusability for these contrasts. Phonemes can be characterized by their phonological features: place of articulation (e.g., labial, dental, palatal), manner of articulation (e.g., stop, fricative, glides) and voice for consonants (voiced, unvoiced); height (close to open), backness (front to back)

and roundness for vowels. For each of the 5% most frequent pairs of contrasts, we calculated the difference in phonological features between each member of the pair. For example the pair /k/ and /m/ has 3 features that differ: place, manner and voicing. The test statistic that we use here is the average number of features that differ in a minimal pair. This measure ranges from 1 (highly confusable) to 3 (highly distinguishable).<sup>11</sup>

Figure 9 shows the average number of features that differ in the 5% most frequent minimal pair contrasts in the real lexicon and across all simulated lexicons for each language. The minimal pairs contrasts in the real lexicon are no more distinguishable in phonetic space than are the minimal pairs in the chance lexicon. This indicates that minimal pairs do not rely on more perceptible contrasts for distinctiveness than what is expected by phonotactics alone.



**Figure 9:** Distributions of the average number of feature difference for the 5% most frequent minimal pair contrasts in the simulated lexicon compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. There is no evidence that these frequent contrasts are more perceptible than expected by chance (all  $ps > .30$ ).

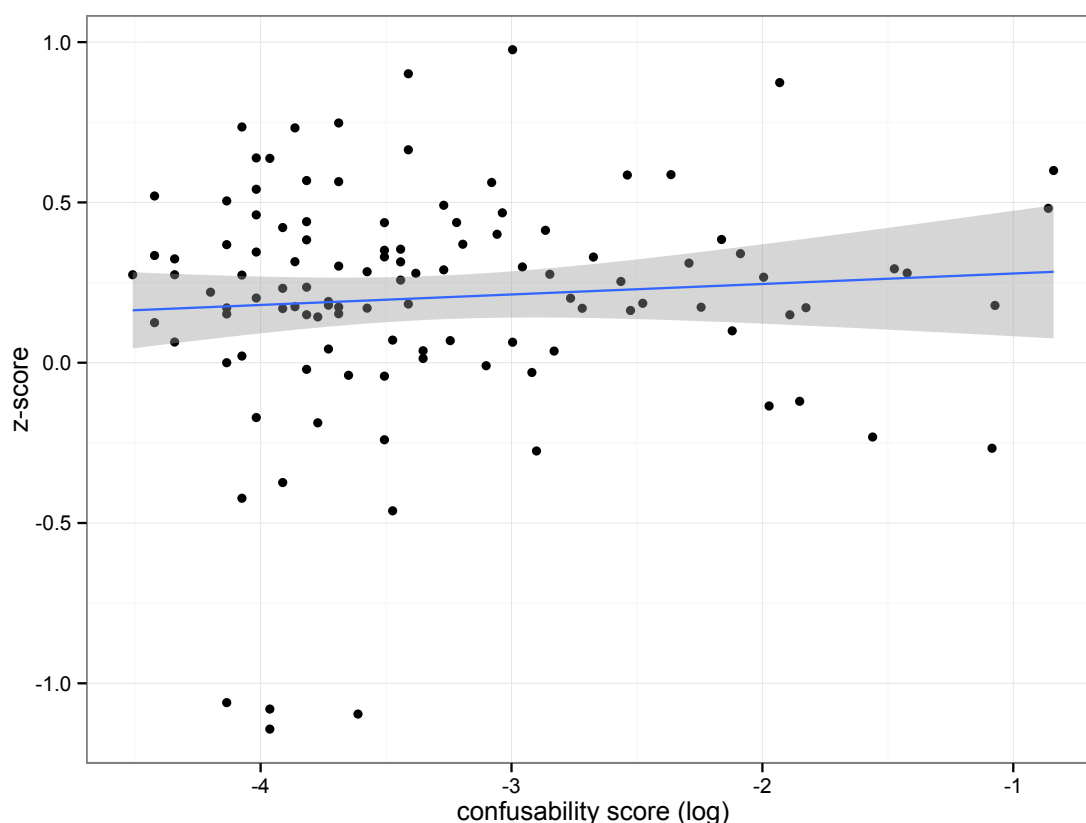
The previous measure showed that frequent minimal pair contrasts are not more perceptible than expected by chance alone. But because we used a coarse measure of confusability (the average number of different phonological features) looking only at the most used contrasts, it could still be the case that a more perceptual and language-specific measure of phoneme confusability—looking at a broader range of possible contrasts—would be a better predictor of clumpiness. If the lexicons prefer minimal pairs to be distinctive then we should observe more minimal pairs having easily perceptible contrasts than those having confusable contrasts. In order to investigate this possibility, we looked at minimal pairs in English for which confusability data between phonemes are readily available (Miller & Nicely, 1955). We computed the distance between the mean number of minimal pairs in our simulated lexicons and the number of minimal pairs in the real lexicon for each of the 120 contrasts present in the Miller and Nicely dataset. The distance is simply the difference between a) the mean number of minimal pairs in the simulated lexicons and b) the number of minimal pairs in the real lexicon, divided by the standard deviation of the value across the 30 simulated lexicons. In effect, this acts as a z-score

<sup>11</sup>For French we added nasalization as a vowel feature. The measure for French vowel contrasts therefore ranged from 1 to 4.



that tell us how far the real lexicon value falls from what we expect under a null model.

Figure 10 shows the  $z$ -score obtained for each phonemic contrast as a function of its confusability (the higher the more confusable). As it can be observed, there is no effect of confusability on the  $z$ -score ( $p > 0.5$ ). That is, there is no evidence that the English lexicon is more clumpy around highly distinctive contrasts than around highly confusable contrasts.



**Figure 10:**  $z$ -score obtained between the mean number of minimal pairs in the real lexicon and in the simulated lexicons for each of the minimal pair contrasts present in the Miller and Nicely’s dataset as a function of their (log) confusability.

Thus, it appears that the clumpiness effect is driven not just by highly distinct sound sequences but is present even when considering highly confusable sounds. This points to a pressure for lexical clumpiness which may work against robust communication.

### 4.2 Wordform similarities within and across grammatical categories

Words do not usually appear in isolation but are embedded in richer linguistic context. A wealth of studies show that adults and children use the context of a sentence to constrain lexical access (Altmann & Kamide, 1999; Borovsky et al., 2012). Hence even if the lexicon is clumpy as a whole, the

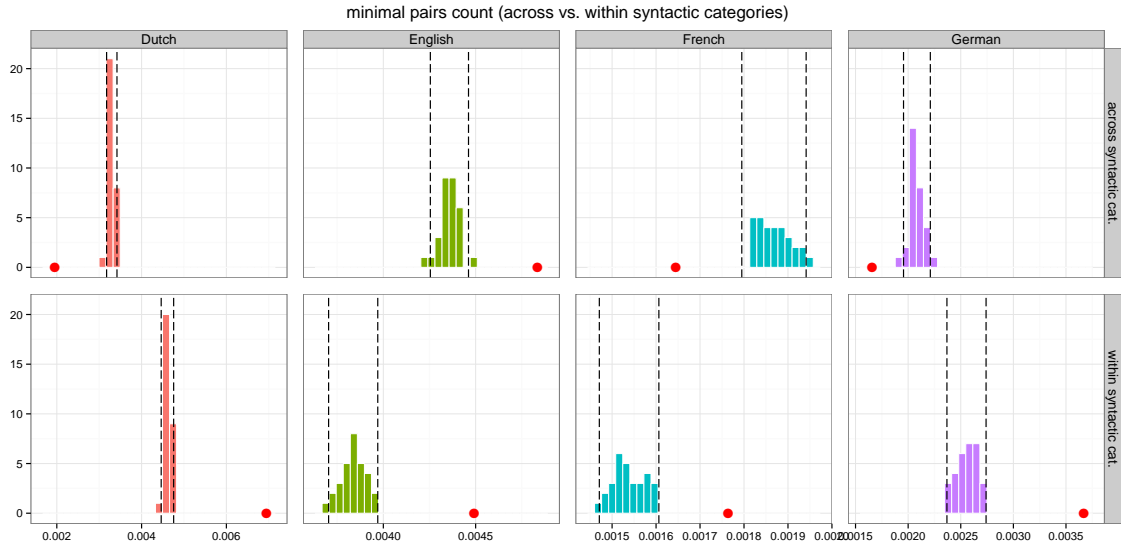
context might be sufficient to disambiguate between two similar wordforms. One obvious contextual disambiguation is the syntactic category of words. For example, consider the sentence “did you see my sock?” The chance that a native English speaker might confuse the word ‘sock’ with ‘lock’ in the context of following ‘my’ might be greater than confusing ‘sock’ with ‘mock’, because ‘lock’ is a noun—which is consistent with the syntactic context—whereas ‘mock’ is a verb, which is inconsistent with the syntactic context. Moreover, because children as young as 18-months have been shown to use function words to recognize and learn the difference between verbs and nouns on-line, these sorts of categorizing effects may be crucial to language acquisition (Cauvet et al., 2014).

As with the lexicon more broadly, there are two possible outcomes that could arise from comparing word forms within as opposed to across syntactic categories. On the one hand, because context is usually enough to distinguish among different parts of speech, confusability of words should be less of a problem across syntactic categories. That is, even though ‘bee’ and ‘see’ are minimal pairs, one is unlikely to misperceive “I was just stung by a *bee*” as “I was just stung by a *see*.” This account predicts more similarity across syntactic categories than within syntactic categories. On the other hand, the effects of learnability and ease of processing may be enhanced by having increased similarity between words of the same part of speech. That is, having nouns that sound like other nouns and verbs that sound like other verbs could convey a processing advantage. Under this account, we would expect more similarity within as opposed to between syntactic category.

For this evaluation, we used the Part Of Speech (POS) tags in CELEX for Dutch, English and German and in Lexique for French to count the number of minimal pairs within the same syntactic categories (e.g., ‘lock’ / ‘sock’) and across different syntactic categories (e.g., ‘mock’ / ‘sock’). For each simulated lexicon, we randomly assigned the syntactic categories of real words of length  $n$  to generated words of length  $n$  and similarly counted the number of minimal pairs appearing within and across the same syntactic categories.<sup>12</sup> Note that for wordforms having several syntactic categories in the real lexicon (homophones, e.g., ‘seam’/‘seem’ which are counted as a single wordform in our lexicons, /sim/), we chose the syntactic category of the most frequent items (e.g., because the most frequent meaning of /sim/ is ‘seem’ it will be categorized as a verb). Because there are more across-category minimal pairs than within-category minimal pairs across languages, we divided the number of minimal pairs appearing across and within categories by the number of across- and within-category word pairs respectively. The final measure is thus the probability of getting a minimal pair, across categories or within categories.

---

<sup>12</sup>This was to ensure that certain categories, such as pronouns, which are reserved for smaller words will not be assigned to longer words.



**Figure 11:** Distributions of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. All 4 languages are significantly more likely to have minimal pairs within categories than would be expected by chance.

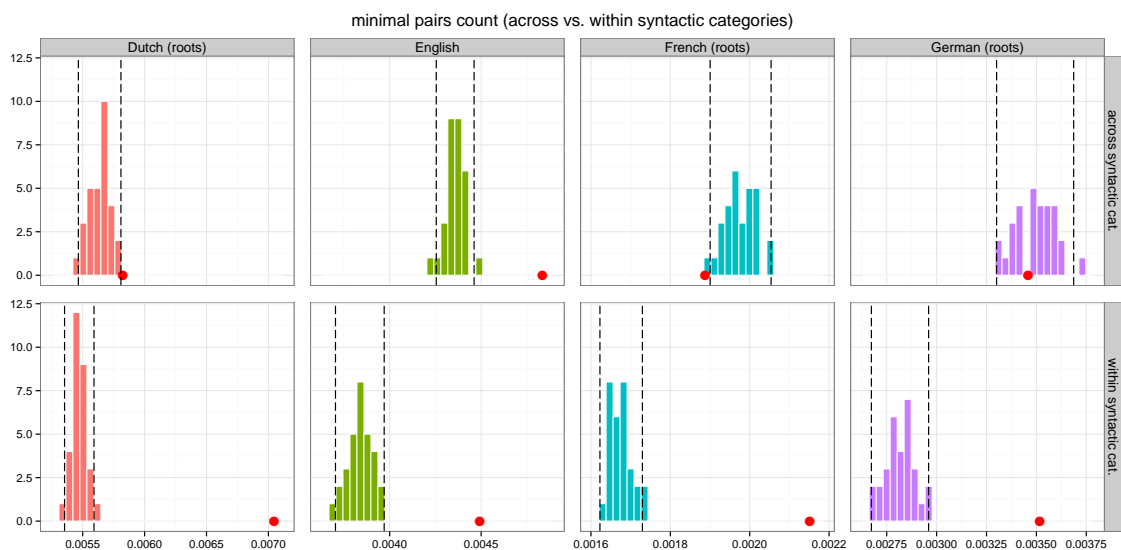
		Dutch	English	French	German
across syntactic categories	real	0.002	0.0048	0.0016	0.0017
	$\mu$ (simulated)	0.0033	0.0044	0.0019	0.0021
	$\sigma$ (simulated)	1e-04	1e-04	1e-05	1e-04
	$z$	-21.5	9	-6	-6.6
	$p$	<.001	<.001	<.001	<.001
	within syntactic categories	real	0.0069	0.0045	0.0018
$\mu$ (simulated)	0.0046	0.0038	0.0015	0.0026	
$\sigma$ (simulated)	1e-04	1e-04	1e-05	1e-04	
$z$	31.2	9.6	6.5	11.7	
$p$	<.001	<.001	<.001	<.001	

**Table 4:**  $z$ - statistics comparing the probability of getting a minimal pair within and across syntactic categories in the real lexicon with the chance distribution of mean  $\mu$  and standard deviation  $\sigma$  corresponding to the distribution of the probability of having a minimal pair in the 30 simulated lexicon for each language. The red  $p$ -values shows a significant effect of clumpiness and the blue ones a significant effect in the opposite direction.

As before, we compare the real lexicon to the simulated lexicons but break the measures down

by similarity within syntactic category (only looking at the similarity of nouns to other nouns, verbs to other verbs, and so on) and between syntactic category (only looking at the similarity of nouns to non-nouns, verbs to non-verbs, etc.). As shown in Figure 11, we found that there are *more* minimal pairs within the same syntactic category in the real lexicons than would be expected by chance for all 4 languages. That is, for within syntactic category analyses, all four languages are clumpier than expected under the null models. For the across-category analysis, the result is less clear. For French, German, Dutch, there are *fewer* minimal pairs across different syntactic categories than would be expected by chance. For English, there are more across-category minimal pairs than expected by chance.

A subsequent post-hoc analysis found that the unclear results for the across-category analysis can in part be explained by the infinitival affixes that appear on French, Dutch, and German verbs. When we remove these verb endings, the across-category differences look roughly like what one expects by chance (see Figure 12). This result is unsurprising since the presence of verb stems like *-er* means that any given verb is less likely to be a neighbor of a noun since most nouns do *not* end in *-er*. The within-category analysis is qualitatively unchanged by focusing on roots (in all cases the real lexicon is clumpier than expected by chance).



**Figure 12:** As in Figure 11, these histograms show the distribution of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon, but without infinitive endings on verbs in Dutch, French and German.

Note that the probability of getting a minimal pair within the same syntactic category is greater than the probability of getting a minimal pair across different syntactic categories for Dutch, French and German but not for English. A possible explanation for this difference is that there is still some verbal morphology present in the lemmas for Dutch, French and German that we could not capture, and this morphology artificially inflates the number of within-category minimal pairs compared to

the number of across-category minimal pairs. For instance, in Dutch, verbs of motion systematically display phonaesthemes (typically a schwa followed by a sonorant) that are not analyzed as suffixes. Another possibility for this difference is that the probability of getting a minimal pair across and within syntactic categories may not be directly comparable because the length distributions for within category words and across categories words are different and may thus drive part of the difference found here. As a result we prefer to concentrate on the comparison of the real lexicon with the simulated lexicons.

### 4.3 Interim summary

To sum up, we did not find evidence that clumpiness is more likely among perceptible than confusable phonological contrasts. That is, it seems that confusable phoneme pairs like ‘m’ and ‘p’ are just as likely to be the basis of minimal pairs as less confusable pairs. One possible explanation for this null result is that even highly confusable phoneme pairs like ‘b’ and ‘p’ are only confusable in certain specific contexts, such as after vowels at the end of words as in ‘cab’ and ‘cap’ (Steriade, 1997). Even then, though, context might be enough to disambiguate the words such that the confusability is not an issue.

We found evidence for more clumpiness within syntactic category than across syntactic categories. This may potentially be the consequence of a more general pattern: words of the same syntactic category may share more phonological properties than with words of different classes (Kelly, 1992). For English words, it is also the case that we see more clustering across categories than expected by chance. But that is not the case for French, German, or Dutch when we control for the presence of infinitival markers. Therefore, at least for these languages, it may even be the case that this syntactic category effect drives the larger clumpiness effect observed across the lexicon. This would be consistent with the findings of Monaghan et al. (2014) and Dautriche et al. (submitted), who show a relationship between semantic and phonological similarity.

## 5 General Discussion

We have shown that lexicons use their degrees of freedom in a systematic and interesting way. While we can still characterize the relationship between wordforms and meanings as arbitrary, structure emerges when one considers the relationships within the space of possible wordforms. Across a wide variety of measures of phonological similarity, the real lexicons of natural languages show significantly more clustering than lexicons produced by the “best” generative model selected by our held-out model comparison procedure.

Because we focused on monomorphemic words, this effect cannot be a result of words sharing prefixes and suffixes. It is also not a product of any structure captured by sound-to-sound transition probabilities such as phonotactic regularities, since our models capture these patterns. This last point is crucial: even though our model took away some clustering effect by capturing sound-to-sound

transition probabilities (compare the density of neighborhood between the network of the 1-phone model to the 5-phone model in Figure 6), there is still some clustering effect that is not explained by frequency distribution of groups of phonemes.

Certainly, one explanation for the clumpiness in the lexicon is shared phonetic properties of semantically related words. Like ‘skirt’ and ‘shirt’, many words in the language share deep etymological roots. Moreover, the presence of sound symbolism in the lexicon is another source of structure in the lexicon not captured by our models. For instance, there is a tendency in English for *gl-* words to be associated with light reflectance as in ‘glimmer’, ‘gleam’ and ‘glisten’ (Bergen, 2004; Bloomfield, 1933). There are additionally cross-linguistic correspondences between form and meaning, such as a tendency for words referring to smallness to contain high vowels (Hinton et al., 2006; Sapir, 1929). Interestingly, recent studies show that phonologically similar words tend to be more semantically similar across measures of wordform similarity over the whole English lexicon (Monaghan et al., 2014) but also in Dutch, French and German. This suggests that clumpiness in the lexicon cannot be attributed to small islands of sound symbolism. Rather, it reveals a fundamental drive for regularity in the lexicon, a drive that conflicts with the pressure for words to be as phonologically distinct as possible.

One possible source of the lexicon’s clumpiness is that speakers preferentially re-use common articulatory sequences. That is, beyond just phonotactics and physical constraints, speakers find it easier to articulate sounds that they already know. Recall our example of the language in which there is only one word for a speaker to learn. She would quickly become an expert. Along those lines, the presence of any given sound sequence in the language makes it more likely that the sequence will be re-used in a new word or a new pronunciation of an existing word. In that sense, the lexicon ‘overfits’: any new word is deeply dependent on the existing words in the lexicon. Note that because our baseline used a lexical generation model, any pressure for re-use must occur over and above the observed statistical trends (e.g., 5-phone sequences) in the language.

Relatedly, lexical clumpiness may be advantageous for word production. While words having many neighbors are challenging for word recognition (Luce, 1986; Luce & Pisoni, 1998), they may be easy words to produce (Gahl et al., 2012). Previous studies suggest that listener-oriented model of speech production— where speakers adjust their speech to ensure intelligibility of words that might otherwise be difficult to understand (as could be words with many neighbors)— are limited by attentional demands and working memory in conversational speech (Arnold, 2008; Lane et al., 2006). However, speakers may produce words with many neighbors faster, because they are easier to access and retrieve (Dell & Gordon, 2003; Gahl et al., 2012). Hence a clumpy lexicon would be beneficial for a speaker-oriented model of speech production associated with rapid lexical access and retrieval.

A clumpy lexicon also may allow for easier compression of lexical knowledge. By having words that share many parts, it may be possible to store words more easily. Though we concentrate here on monomorphemic lemmas, these account only for one third of all the lemmas in the lexicon. The fact that languages reuse words or parts of words in the remaining two thirds of the lemmas shows that re-use of existing phonological material must be important. It may even be the case that, much

as morphology allows the productive combination of word parts into novel words, there exist sound sequences below the level of the morpheme that *also* act as productive units of sound.

The interaction of these cognitive and articulatory constraints with the pressure for communicative efficiency is complex. Despite the fact that one might expect the lexicon to be maximally dispersed for communicative efficiency, these results strongly suggest that the lexicon is not nearly as sparse as it could be—even given various phonetic constraints. Thus, why does communicative efficiency not conflict with clumpiness in the lexicon?

One possibility is that clumpiness does not appear randomly in the lexicon but is organized along dimensions that maximize wordform recoverability. We hypothesized that recoverability could be enhanced if similar wordforms such as minimal pairs were disambiguated by minimally confusable sounds. Our results provide no evidence that the lexicon is less clumpy for confusable sounds than for non-confusable sounds. Relatedly, lexical access might be faster in a lexicon where confusable wordforms span different syntactic categories. Yet we find that, if anything, wordforms are more similar *within* the same syntactic category than what would be expected by chance for all four languages despite the absence of morphology. This is in line with experimental evidence showing that phonological similarity might act as pointer to grammatical categories to facilitate learning (Monaghan et al., 2011).

Another possibility that would explain why communicative efficiency does not conflict with clumpiness in the lexicon is that contextual information outside the word pronunciation is usually enough to disambiguate words. Therefore it simply does not matter whether certain words are closer together in phonetic space than they might otherwise be. Piantadosi et al. (2012) showed that lexical ambiguity, such as dozens of meanings for short words like *run*, does not impede communication and in fact promotes it by allowing the re-use of short words. In a similar way, there may be a communicative advantage from having not just identical words re-used but from re-using words that are merely similar. In all cases, context may be enough to disambiguate the intended meaning and avoid confusion—whether it be confusion between two competing meanings for the same word or confusion between two similar-sounding words.

Likewise, our analysis here concentrated on the phoneme representation of words ignoring the fact that speech contains a lot of fine phonetic details that listeners could use to disambiguate between words. For instance, pairs of homophones such as ‘thyme’/‘time’ in English can be differentiated based on their duration (“gahl2008time”, n.d.), or on their stress pattern (e.g., ‘désert’ and ‘desért’). Kemps et al. (2005) show that English and Dutch listeners are sensitive to fine-grained durational differences between a base word (‘run’) and the base word as it occurs in an inflected or derived word (‘runner’). Being sensitive to these cues may also be useful to disambiguate between words that sound similar such as minimal pairs.

The methodology used here, whereby the real lexicon is compared to a distribution of statistically plausible ‘null’ lexicons, could be generalized to answer other questions about the lexicon and human language more generally. While much previous work has focused on simply measuring statistical properties of natural language, modern computing power makes it possible to simulate thousands

of different languages with different constraints, structures, and biases. By comparing real natural language to a range of simulated possibilities, it is possible to assess which aspects of natural language occur by chance and which exist for a reason.

Of course, we must keep in mind that the present work examines only a small number of European languages. To know whether the effect generalizes would require a larger number of languages, and we undertake exactly such a project in other work (Dautriche et al., submitted). Specifically, we use a corpus of 100+ languages from Wikipedia to show large-scale evidence for a) more frequent words to be more orthographically probable and have more minimal pairs than less frequent words and b) for semantically related words to be more phonetically similar than less related words. While the Wikipedia corpus does not focus on monomorphemes and is therefore less controlled than the results presented here, it suggests that the clumpiness we observe in the lexicons of Dutch, English, German, and French likely generalizes to other languages as well.

In future work, it may be possible to test increasingly sophisticated models of phonotactics using this methodology. Perhaps our models of phonotactics are simply not good enough yet to capture the rich structure of natural language. But the results here suggest that any “null” model that can approximate natural language will need to account for not just the preferred sounds of a language but for the entire space of existing words. That is, the goodness of ‘dax’ as an English word depends not just on an underlying model of English sound structure but on the fact that ‘lax’ and ‘wax’ are words, that ‘bax’ is not, and on countless other properties of the existing lexicon.

Overall, we have shown that lexicons are more richly structured than previously thought. The space of wordforms for Dutch, English, German and French is clumpier than what would be expected by the best chance model by a wide variety of measures: minimal pairs, average Levenshtein distance and several network properties. The strongest evidence comes from minimal pairs, for which the effect size was quite large. From this, we conclude that the clustered nature of the lexicon holds over and above the patterns that are captured by a phonotactic model. Underlying the pressure for dispersion in the lexical system is a deep drive for regularity and re-use beyond standard levels of lexical and morphological analysis.

### **Acknowledgement**

We thank Benoit Crabbé, Emmanuel Dupoux and all members of Tedlab, the audience at AMLaP 2014, and the audience at CUNY 2013 for helpful comments. This work was supported by ANR-10-LABX-0087, ANR-10-IDEX-0001-02, ANR-13-APPR-0012, as well as an NDSEG graduate fellowship and an NSF Doctoral Dissertation Improvement Grant in linguistics to KM and a PhD fellowship from DGA to ID.



### References

- (n.d.).
- Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(01), 9–41.
- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03), 679–685. doi: 10.1142/S021812741002596X
- Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, 23(4), 495–527.
- Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5), 3048–3058.
- Baayen, R. (1991). A stochastic process for word frequency distributions. In *Proceedings of the 29th annual meeting on association for computational linguistics* (pp. 271–278).
- Baayen, R. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.
- Baayen, R., Piepenbrock, R., & van H, R. (1993). The celex lexical data base on cd-rom. *n.s.*
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439), 509–512.
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113, 1001.
- Bergen, B. K. (2004). The psychological reality of phonaestemes. *Language*, 80(2), 290–311. doi: 10.1353/lan.2004.0056
- Bloomfield, L. (1933). *Language*. New York: Henry Holt.
- Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, 112(4), 417–436. doi: 10.1016/j.jecp.2012.01.005
- Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in french 18-month-olds. *Language Learning and Development*, 10(1), 1–18.

- Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, 1, 97-138.
- Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89(3), 183–213. doi: 10.1016/j.jecp.2004.07.004
- Cohen Priva, U. (2008). Using Information Content to Predict Phone Deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics* (p. 90).
- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (submitted). Cross-linguistic effects of semantics on wordform similarity.
- Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 6, 9.
- de Saussure, F. (1916). *Course in general linguistics*. Open Court Publishing Company.
- Flemming, E. (2002). *Auditory representations in phonology*. Routledge.
- Flemming, E. (2004). Contrast and perceptual distinctiveness. In *Phonetically based phonology* (eds. Bruce Hayes, Robert Kirchner, Donca Steriade). Cambridge: Cambridge University Press.
- Gafos, A. I. (2014). *The articulatory basis of locality in phonology*. Routledge.
- Gahl, S. (2015). Lexical competition in vowel articulation revisited: Vowel dispersion in the easy/hard database. *Journal of Phonetics*, 49, 96–116.
- Gahl, S., Yao, Y., & Johnson, K. (2012, May). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806. doi: 10.1016/j.jml.2011.11.006
- Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1216438110
- Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3), 859–896. doi: 10.1007/s11049-012-9169-1
- Graff, P. (2012). *Communicative efficiency in the lexicon* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Hayes, B. (2012). *BLICK - a phonotactic probability calculator*.

- Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3), 379–440. doi: 10.1162/ling.2008.39.3.379
- Hinton, L., Nichols, J., & Ohala, J. J. (2006). *Sound symbolism*. Cambridge University Press.
- Hockett, C. (1960). The origin of language. *Scientific American*, 203(3), 88–96.
- Hockett, C., & Voegelin, C. (1955). *A manual of phonology* (Vol. 21) (No. 4). Waverly Press Baltimore, MD.
- Howes, D. (1968). Zipf's law and miller's random-monkey model. *The American Journal of Psychology*, 81(2), 269–272.
- Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130298–20130298. doi: 10.1098/rstb.2013.0298
- Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition*, 109(1), 54–65. doi: 10.1016/j.cognition.2008.07.015
- Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Bayesian inference for PCFGs via markov chain monte carlo. In *HLT-NAACL* (pp. 139–146).
- Jusczyk, P., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630–645. doi: 10.1006/jmla.1994.1030
- Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99(2), 349–364. doi: 10.1037/0033-295X.99.2.349
- Kelly, M. H., et al. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological review*, 99(2), 349–364.
- Kemps, R. J., Wurm, L. H., Ernestus, M., Schreuder, R., & Baayen, H. (2005). Prosodic cues for morphological complexity in dutch and english. *Language and Cognitive Processes*, 20(1-2), 43–73.
- Lane, L. W., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! speakers' control over leaking private information during language production. *Psychological science*, 17(4), 273–277.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady* (Vol. 10, p. 707).
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

- Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception. Technical Report No. 6.*
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19*(1), 1.
- Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*(1), 133–156. doi: 10.1080/03640210709336987
- Mandelbrot, B. (1958). An informational theory of the statistical structure of language. *Communication theory, 4*86–502.
- Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 59). Cambridge, MA: MIT Press.
- Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology, 3*11–314.
- Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America, 27*(2), 338–352.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General, 140*(3), 325–347. doi: 10.1037/a0022924
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B.*
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516–524.
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science, 16*(1), 24–34.
- Nielsen, A., & Rendall, D. (2012). The source and magnitude of sound-symbolic biases in processing artificial word material and their implications for language learning and transmission. *Language and Cognition, 4*(02), 115–125. doi: 10.1515/langcog-2012-0007
- Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition, 112*(1), 181–186. doi: 10.1016/j.cognition.2009.04.001
- Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*(3), 280–291. doi: 10.1016/j.cognition.2011.10.004

- Piantadosi, S., Tily, H., & Gibson, E. (2013). Information content versus word length in natural language: A reply to ferrer-i-cancho and moscoso del prado martin [arXiv: 1209.1751]. *arXiv preprint arXiv:1307.6726*.
- Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive psychology*, *68*, 33–58.
- Sapir, E. (1929). A study in phonetic symbolism. *Journal of experimental psychology*, *12*(3), 225.
- Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 623–656.
- Sporns, O. (2011). *Networks of the brain*. MIT Press.
- Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1), 413–422.
- Steriade, D. (1997). *Phonetics in phonology: The case of laryngeal neutralization*.
- Steriade, D. (2001). Directional asymmetries in place assimilation: a perceptual account. In *In hume and johnson*.
- Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, *25*(02). doi: 10.1017/S0142716404001109
- Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, *36*(02), 291. doi: 10.1017/S030500090800891X
- Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1175–1192.
- Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken american english. *Behavior Research Methods*, *42*(2), 497–506. doi: 10.3758/BRM.42.2.497
- Vitevitch, M. S. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1-2), 306–311. doi: 10.1006/brln.1999.2116
- Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of memory and language*, *67*(1), 30–44.
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325–329.

- Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4), 491–504.
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.
- Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440–442.
- Younger, D. H. (1967). Recognition and parsing of context-free languages in time  $n^3$ . *Information and control*, 10(2), 189–208.

## Wordform similarity increases with semantic similarity: an analysis of 101 languages

Isabelle Dautriche\*<sup>1</sup>, Kyle Mahowald<sup>2</sup>, Edward Gibson<sup>2</sup>, and Steven T. Piantadosi<sup>3</sup>

<sup>1</sup>Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France

<sup>2</sup>Department of Brain and Cognitive Science, MIT

<sup>3</sup>Department of Brain and Cognitive Sciences, University of Rochester

### Abstract

Although the mapping between form and meaning is often regarded as arbitrary, there are in fact well-known constraints on wordforms which are the result of functional pressures associated with language use and its acquisition. In particular, languages have been shown to encode some meaning distinction in their sound properties that are described to be important for language learning. Here, we investigate the relationship between semantic distance and phonological distance at the large-scale structure of the lexicon. We show evidence in 101 languages from a diverse array of language families that more semantically similar word pairs are also more phonologically similar. We argue that there is a pervasive functional advantage for lexicons to have semantically similar words be phonologically similar as well.

**Keywords:** lexicon, phonetics, semantics, lexical design

---

\*For correspondence, e-mail [isabelle.dautriche@gmail.com](mailto:isabelle.dautriche@gmail.com)

## 1 Introduction

Why do languages have the set of wordforms that they do? Although the mapping between form and meaning is often regarded as arbitrary (de Saussure, 1916; Hockett, 1960), there are in fact well established regularities in lexical systems. The simplest of these involve correlations between word length and frequency (Zipf, 1949) or informativity (Piantadosi et al., 2011). Patterns can also be found in which specific wordforms are in a language, including the presence of clusters of phonological forms (over and above effects of phonotactics or morphology) (Mahowald, Dautriche, Gibson, Christophe, & Piantadosi, *submitted*), observed particularly in high-frequency words (Mahowald, Dautriche, Gibson, & Piantadosi, *submitted*). Deeply semantic regularities are also observed: sound symbolism, in which languages encode some meaning distinction in their sound properties,<sup>1</sup> is one such form-meaning regularity and is present across many languages and cultures (e.g., Bremner et al., 2013; Childs, 1994; Hamano, 1998; Kim, 1977). For instance, adults intuitively pair ‘bouba’ with a picture of a rounded object while they pair ‘kiki’ with a picture of a spiky object (the “bouba-kiki” effect, e.g., Bremner et al. 2013). Relatedly, certain sequences of sounds, called phonesthemes, tend to carry a certain semantic connotation. For instance, there is a tendency in English for *gl-* words to be associated with light reflectance as in ‘glitter’, ‘glimmer’, and ‘glisten’ (Bergen, 2004; Bloomfield, 1933) or words ending with *-ack* and *-ash* associated with abrupt contact (e.g., ‘smack’, ‘smash’, ‘crash’, ‘mash’). Additionally, certain meaning distinctions are present in the phonological form of words more transparently. For instance, semantic features, such as objects vs. actions, that are associated with grammatical distinctions may be marked morphologically (Monaghan & Christiansen, 2008; Pinker, 1984).

Several studies suggest that systematic form-meaning mappings may facilitate word learning (e.g., Imai & Kita, 2014; Monaghan et al., 2011). The idea is that learning similarities among referents (and hence forming semantic categories) may be facilitated if these similarities appear also at the level of the wordform. For instance, it might be easier to learn the association of *lep* and *leb* to CAT and DOG than to CAT and UMBRELLA. This advantage in learning may be an explanation for the observation of sound-symbolism in languages and predicts that phonologically similar words would tend to be more semantically similar. In this spirit, several studies have established that it is easier to learn languages that are compressible (Kemp & Regier, 2012). For instance, in the limit, the easiest language to learn is a language that uses only one word to express all meanings. More generally, it should be easier to learn languages whose words tend to sound similar to each other, as *lep* and *leb*, because there is less phonetic material to learn, remember or produce (Gahl et al., 2012; Stemberger, 2004; Storkel et al., 2006; Storkel & Lee, 2011; Vitevitch & Sommers, 2003).

Yet there may also be a functional disadvantage for form-meaning regularities. Another feature of semantically related words is that they are likely to occur in similar contexts. For instance, weather words like ‘rain’, ‘wind’, and ‘sun’ are all likely to occur in the same discourse contexts—namely

---

<sup>1</sup>Note that this is not specific to spoken languages, sign languages do also map meanings into visual sign (see Strickland et al. *in press*)



when people are talking about the weather. As a result, one might imagine that context makes it more difficult to distinguish between semantically similar words. If someone said, “Weather forecast: \_\_\_ today and tomorrow” the missing word could plausibly have be ‘sun’ or ‘wind’, but it’s unlikely to be ‘boat’ or ‘John’. Therefore, one would also predict that semantically related words should be more distant in phonological space than semantically unrelated words, much like theories positing dispersion of phonemes in vowel space (e.g., Liljencrants & Lindblom, 1972).

In this work, we investigate the relationship between semantic distance and phonological distance. If there is a positive correlation between semantic distance and phonological distance—i.e., more similar wordforms are more semantically similar—then this would imply a pressure for phonological clustering that is tied specifically to meaning. On the other hand, if there is a negative correlation between semantic distance and phonological distance, there would be a pressure for dispersion for words’ meanings to be more distinct relative to phonological distance, likely due to communicative pressures of confusability. Monaghan et al. (2014) previously examined the correlation between semantic distance and phonological distance in English. In this work, the authors found that phonologically similar words tend to be more semantically similar. While this result is telling, the sample of a single language does not indicate if form-meaning regularities in the lexicon are the product of functional pressures that universally apply, or historical accidents of English.

In the present work, the existence of large-scale data sets in a large number of languages makes it possible to investigate semantic and phonological relatedness across human language more generally. We use a dataset of 101 languages extracted from Wikipedia from a diverse array of language families. First, we performed several statistical tests to look at the correlation between semantic similarity (calculated using Latent Semantic Analysis over each Wikipedia corpus) and orthographic similarity: Pearson correlations and a mixed model analysis to ensure that the correlation observed does not depend on a particular language family. Second, we probed the relation between semantic and phonological similarity by using a different measure looking at the interaction of semantic relatedness and the likelihood of finding a minimal pair. Finally, we also used a subset of 4 languages to assess whether the correlation between semantic and phonological similarity still hold in a set of monomorphemic words with phonemic representations. In sum, across all these languages we found that semantically similar words tend to be phonologically similar, providing large-scale, cross-linguistic evidence for phonological clustering of semantically similar words.

## 2 Method

### 101 orthographic lexicons:

We extracted the lexicons of 101 languages from the Wikipedia database (as in Appendix A and Mahowald, Dautriche, Gibson, & Piantadosi (*submitted*)). We define as the lexicon of these language the 5,000 most frequent wordforms in the Wikipedia corpus.<sup>2</sup> Because a proper lemmatizer does not

---

<sup>2</sup>Since we calculated words’ semantic distance for all pairs of words of the same length, this restriction was to limit the

exist for most of these languages, all of the 5,000 most frequent wordforms were included regardless of their morphemic status. In order to minimize the impact of semantic similarity due to morphological regularity (e.g., while comparing ‘cat’ and ‘cats’), we only compared words of the same length (Section 3.3 presents a more rigorous analysis looking at semantic distance in monomorphemic words in a smaller number of languages).

### 3 phonemic lexicons:

To assess whether a correlation between semantic similarity and phonological similarity holds in a set of monomorphemic words with phonemic representations, we also used phonemic lexicons derived from CELEX for Dutch, English and German (Baayen et al., 1995) and Lexique for French (New et al., 2004). The lexicons were restricted to include only monomorphemic lemmas (coded as "M" in CELEX; I.D. (a French native speaker) identified mono-morphemes by hand for French). That is, they contained neither inflectional affixes (like plural *-s*) nor derivational affixes like *-ness*. In order to focus on the most used parts of the lexicon, we selected only words whose frequency in CELEX or in Lexique is strictly greater than 0. Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., ‘bat’) we included this form only once.

All three CELEX dictionaries were transformed so that diphthongs were changed into 2-character strings. In each lexicon, we removed a small set of words containing foreign characters. This resulted in a lexicon of 5459 words for Dutch, 6512 words for English, 4219 words for German and 6782 words for French.

### Variables under consideration:

For each pair of words of the same length in each of the lexicons, we computed the pair’s:

- **Orthographic/Phonological distance:** we used the edit distance, or Levenshtein distance between the two orthographic strings in the case of the 101 orthographic lexicons and phonemic strings in the case of the 4 phonemic lexicons. The smaller the distance, the more similar wordforms are to each other. For example, the words ‘cat’ and ‘car’ are very similar, with an edit distance of 1.
- **Semantic distance:** we used Latent Semantic Analysis (LSA, Landauer & Dumais 1997), a class of distributional semantic models that build on the hypothesis that words’ meanings can be inferred from their context (Harris, 1954). Two words are expected to be semantically similar if their pattern of co-occurrence in some observed text is similar. For example, ‘cat’ and ‘dog’ will be more similar than ‘cat’ and ‘bottle’ because they are more likely to co-occur with the same vocabulary (e.g., animal, domestic, pet, etc.). One advantage of using this technique as a proxy for semantics rather than hand-made lexical taxonomies such as WordNet (Miller, 1995) – which is only extensively developed in English – is that it can be adapted for any language given

---

number of possible calculations for each language.

a sufficiently large corpus. We note however that the results obtained from several measures of word distance using WordNet provide the same results as an LSA model trained on English (see Appendix B).

We applied LSA on Wikipedia for each language using the `Gensim` package (Rehurek & Sojka, 2010) in the `R` programming language (R Core Team, 2013). This model splits the whole Wikipedia corpus into documents consisting of  $n$  lines of text and constructs a word-document matrix where each row  $i$  is a word and each column  $j$ , a document. Each matrix cell  $c_{ij}$  corresponds to the frequency count of word  $i$  in document  $j$ . The matrix is then reduced to a dimension  $d$  corresponding to the number of semantic dimensions of the model using Singular Value Decomposition. The semantic distance between two words is computed as 1 minus the absolute value of the cosine of the angle between the two word vectors in the space of dimension  $d$ . A value close to 0 indicates that two words are close in meaning, whereas values close to 1 indicate that the meanings are not related.

For our purposes we defined a document as a Wikipedia article (number of documents per language corpus: median = 42,989; min = 104 – Buginese; max = 36.6 billion – English) and  $d = 500$  dimensions<sup>3</sup> based on (Fourtassi & Dupoux, 2013; Rehurek & Sojka, 2010). We also discarded words that appear in less than 20 documents and in more than 50% of the documents to account for the fact that very common and very rare terms are weak predictors of semantic content (a procedure commonly used in Machine Learning; Luhn (1958)).

## 3 Results

### 3.1 Large-scale effects of semantics on 101 languages

#### 3.1.1 Pearson correlations analysis

For each language, we computed Pearson correlations between the semantic distance of all pairs of words of the same length (focusing on words of length 3 to 7) and the pairs' orthographic distance. The semantic distance was centered around the mean semantic distance for each length and each language and scaled by the standard deviation for each length and each language. To evaluate the correlation between semantic distance and orthographic distance, we need to compare it to a baseline that reflects the chance correlation between form and meanings in the lexicon. We created such a baseline by randomly permuting the form/meaning mappings for words of a given length, randomly reassigning every word meaning to a word of the same length. For example, the meaning of 'car' could be reassigned to 'cat' and the meaning of 'dog' to 'rat'. Under this permutation, the mapping between form and meaning (unlike in the real lexicon) is entirely arbitrary for words of a given length. For each language, we randomly reassigned meanings 30 times and computed Pearson correlations for

---

<sup>3</sup>For Buginese which was the only language having less documents than 500 (the number of dimensions), we took  $d = 20$  based on (Fourtassi & Dupoux, 2013).

each word length. We then asked whether the correlation between wordform distance and semantic distance of the real lexicons falls outside the range of correlation values that could be expected by chance, where chance means random form-meaning assignments.

Figure 1 summarizes this hypothesis test for 4-letter words across the 101 languages. Each bar represents the Pearson correlation score for a given language, and each color represents a language family. We observe that a) all correlations are positive but one (Buginese<sup>4</sup>); b) most of the correlations are significantly positive (in 74/101 languages; dark colors) meaning that the correlation between semantic distance and orthographic distance is more positive than what would be expected by chance alone.

As in standard null hypothesis testing, we compute a  $z$ -score using the mean and standard deviation of correlations scores estimated from these 30 meanings rearrangements. The  $p$ -value reflects the probability that the real lexicon correlations could have arisen by chance. As can be seen in Table 1, we found that the great majority of languages display a significant positive correlation between semantic distance and orthographic distance for all lengths. Yet, even though the correlation is highly significant, one needs to observe that this is a tiny effect explaining only a very small amount of the variance ( $r < 0.05$ ).

word length	mean correlation	proportion showing positive correlation	proportion showing significant correlation
3 letters	0.049	1	0.72
4 letters	0.041	0.99	0.74
5 letters	0.040	0.99	0.73
6 letters	0.040	1	0.94
7 letters	0.047	0.91	0.71

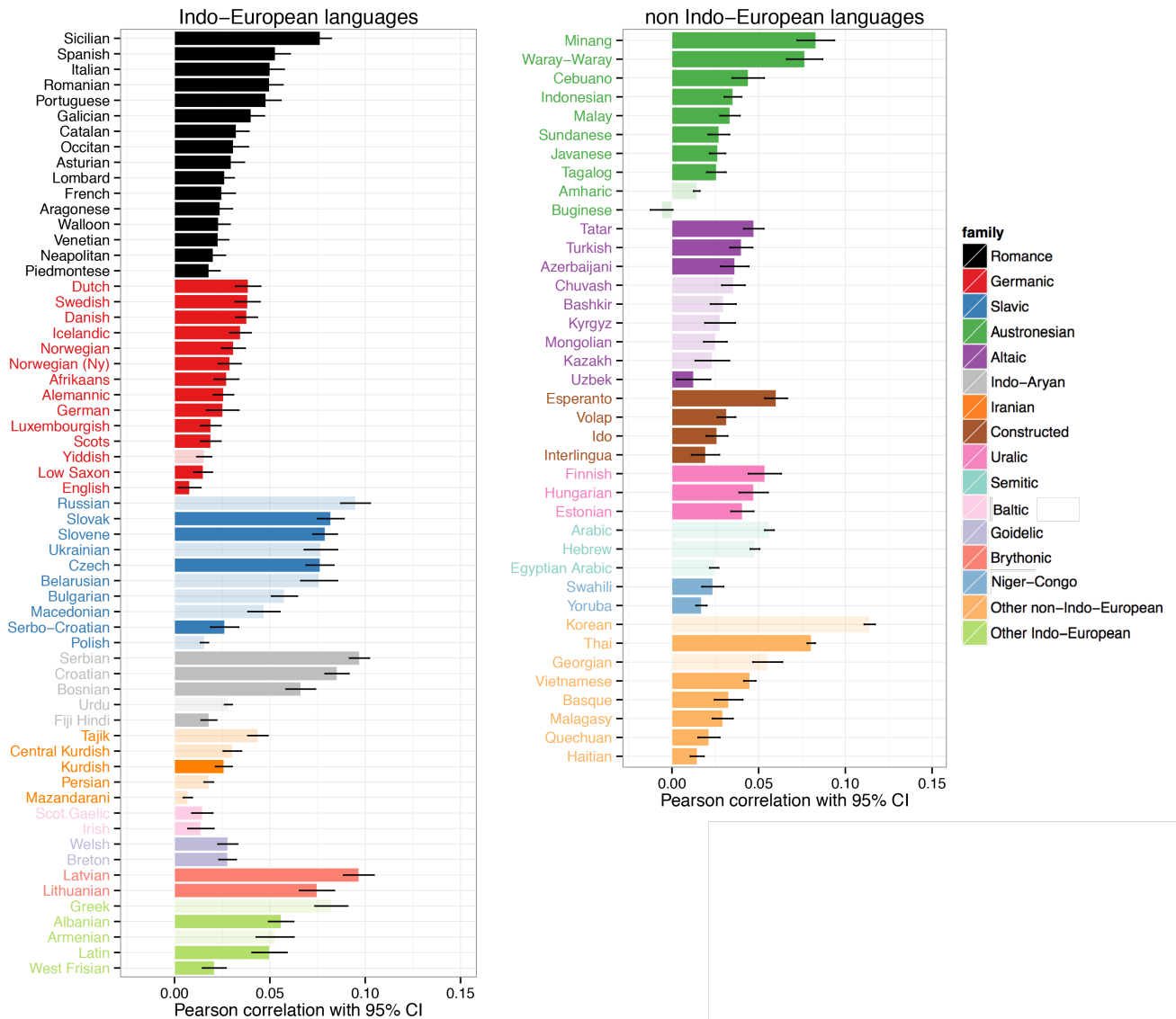
**Table 1:** For each length: (a) the mean Pearson correlation across languages for the relationship between semantic and orthographic distance; (b) the proportion of languages that show a positive correlation between semantic distance and orthographic distance, and (c) the proportion of languages for which this relationship is significantly different from chance at  $p < .05$ , chance being the correlation obtained during 30 random form-meaning reassignments.

### 3.1.2 Mixed effect analysis

To ensure that the observed effect does not depend on a particular language family, we ran a mixed effect regression predicting scaled semantic distance for each pair of words from the Levenshtein distance between the words of the pair. We used a maximal random effect structure with random intercepts for each language, language sub-family, and language family and slopes for Levenshtein distance for each of those random intercepts. Because of the large number of data points, we fit each

<sup>4</sup>Recall that Buginese was our smaller corpora. Inspection of the words of that corpus revealed that, in addition, most of the nouns were names of places (on average 60% from a random samples of 100 words).

## 2 Quantifying word form similarity in the lexicons of natural languages



**Figure 1:** Pearson correlation between semantic distance ( $1 - \text{cosine}$ ) and orthographic distance (Levenshtein distance) for each language for word of length 4. Languages are grouped per language family for Indo-European languages (left plot) and non Indo-European languages (right plot). Dark colors are used for significant Pearson correlations ( $p < .05$ ) and light colors for non-significant correlations.

length separately (words of length 3 through length 7). We compared the full model to an identical model without a fixed effect for the number of minimal pairs using a likelihood ratio test.

Table 2 shows the coefficient estimates for an effect of Levenshtein distance on semantic distance. For every word length, the coefficient for Levenshtein distance is significantly positive meaning that increased semantic distance comes with increased Levenshtein distance beyond effects of language family or sub-family.

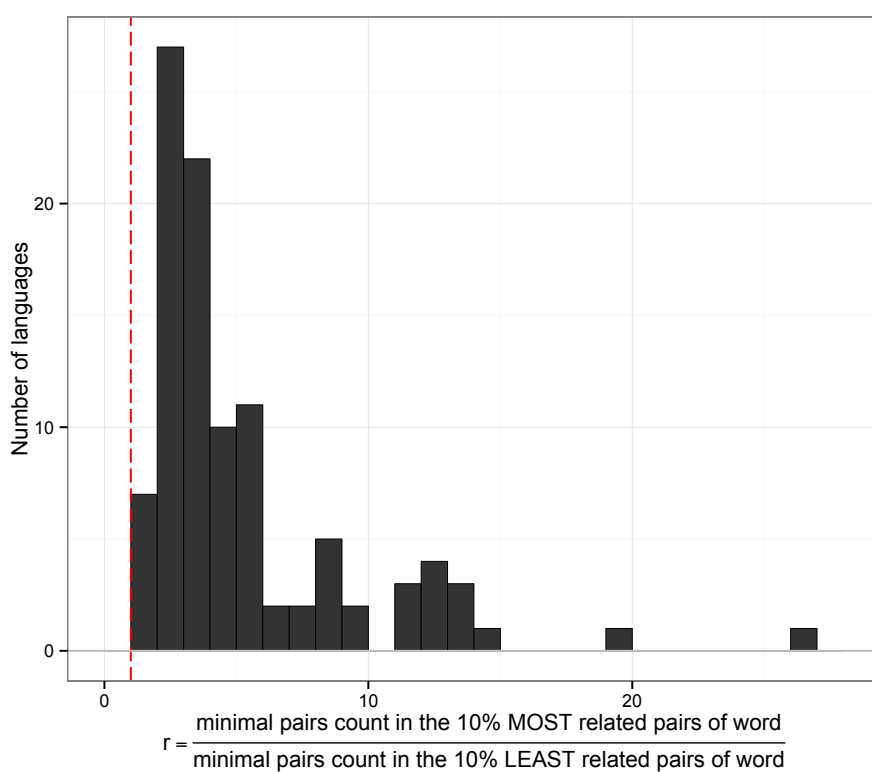
word length	Levenshtein distance
3 letters	.11 ***
4 letters	.07 ***
5 letters	.06 ***
6 letters	.04 ***
7 letters	.04 ***

**Table 2:** Summary of the full models including random intercepts and slopes for language, sub-family, and family for Levenshtein distance for each word length. Three asterisks means that by a likelihood test, the predictor significantly improves model fit at  $p < .001$ .

### 3.2 Likelihood of finding a minimal pair in 101 languages

In addition we looked at the interaction of semantic relatedness and the likelihood of finding a minimal pair. For each language, we compared the number of minimal pairs in the top 10% of semantically related words pairs  $n_{top}$ , and in the bottom 10% of semantically related words pairs,  $n_{bottom}$ , by looking at the ratio  $\frac{n_{top}}{n_{bottom}}$ . A ratio below 1 means that there are more minimal pairs in semantically unrelated words than in related words, while a ratio greater than 1 means that there are more minimal pairs among semantically related words than unrelated words. Figure 2 shows the histogram of the distribution of ratio  $\frac{n_{top}}{n_{bottom}}$  across all languages. As we can observe, in all 101 languages, minimal pairs are on average 3.52 (median of the distribution) more likely to appear in the top 10% semantically related words than in the least 10% related words.<sup>5</sup>

<sup>5</sup>Note that we obtain qualitatively the same results by looking at the 25% most related and the 25% least related words or other percentages.

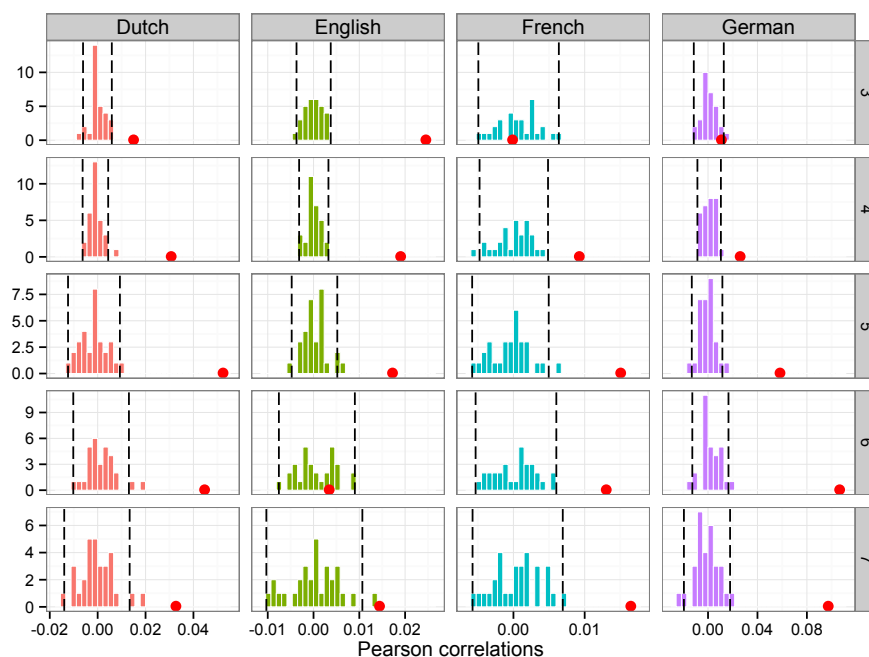


**Figure 2:** Distribution of the ratio of the number of minimal pairs in the 10% most related words compared to the number of minimal pairs in the 10% least related words in a given lexicon, across all the languages. A ratio below 1 means that there are more minimal pairs in semantically unrelated words than in related words, while a ratio greater than 1 means that there are more minimal pairs among semantically related words than unrelated words.

### 3.3 Generalizing form-meaning regularity to monomorphemic words

One obvious explanation for the positive correlation between semantic and orthographic distances is the presence of morphological regularity among the 101 lexicons we studied here. Even though we studied words of the same length to limit this effect, there is certainly some morphological regularity remaining (e.g., ‘capitalist’ / ‘capitalism’). To separate the correlation between phonological and semantic distance due to morphemic regularity from the correlation we are interested in, we restricted our analysis to four languages, Dutch, English, French and German, for which mono-morphemic codes are readily available.

For the monomorphemes of Dutch, English, French and German, we computed Pearson correlations between semantic distance and phonological distance for each word length and compared it to the correlations obtained after 30 random form-meaning reassignments. As shown in Figure 3, the correlations obtained in the real lexicons for each word length (the red dot) tend to be significantly more positive than the correlations obtained in 30 random configurations of form-meaning pairings (the histograms) (see also Table 3).



**Figure 3:** Pearson correlations between semantic distance and phonological distance for word of length 3 to 7 (in rows) for Dutch, English, French and German. Each histogram shows a distribution of correlations obtained after 30 random form-meaning assignments (chance level). The red dots are the correlations found in the real lexicon for that particular length. The dotted lines represent the 95% interval. The red dots tend to be to the right of the histograms.



Word length		Dutch	English	French	German
3	real	0.015	0.025	0	0.011
	$r$ (simulated)	0	0	0.001	0.001
	$\sigma$ (simulated)	0.003	0.002	0.003	0.006
	$z$	4.9	12.9	-0.3	1.7
	$p$	<b>&lt;.001</b>	<b>&lt;.001</b>	0.769	0.099
4	real	0.031	0.019	0.009	0.026
	$r$ (simulated)	-0.001	0	0	0.001
	$\sigma$ (simulated)	0.003	0.002	0.002	0.005
	$z$	11.6	11.6	3.8	5.2
	$p$	<b>&lt;.001</b>	<b>&lt;.001</b>	<b>&lt;.001</b>	<b>&lt;.001</b>
5	real	0.052	0.017	0.015	0.058
	$r$ (simulated)	-0.002	0	0	-0.001
	$\sigma$ (simulated)	0.006	0.003	0.003	0.006
	$z$	9.8	6.7	5.6	9.3
	$p$	<b>&lt;.001</b>	<b>&lt;.001</b>	<b>&lt;.001</b>	<b>&lt;.001</b>
6	real	0.045	0.003	0.013	0.107
	$r$ (simulated)	0.001	0.001	0	0.002
	$\sigma$ (simulated)	0.006	0.004	0.003	0.007
	$z$	7.3	0.6	4.4	14
	$p$	<b>&lt;.001</b>	0.525	<b>&lt;.001</b>	<b>&lt;.001</b>
7	real	<b>&lt;.05</b>	<b>&lt;.05</b>	<b>&lt;.05</b>	0.097
	$r$ (simulated)	0	0	0.001	-0.001
	$\sigma$ (simulated)	0.007	0.005	0.003	0.01
	$z$	4.7	2.7	4.9	10.3
	$p$	<b>&lt;.001</b>	<b>&lt;.01</b>	<b>&lt;.001</b>	<b>&lt;.001</b>

**Table 3:**  $z$ -statistics comparing the Pearson correlations ( $r$ ) between semantic distance (1 - cosine) and orthographic distance (Levenshtein distance) for each word length (2 to 7 phones) and each language with the chance distribution of mean  $\mu$  and standard deviation  $\sigma$  corresponding to the distribution of Pearson correlations obtained in 30 random form-meaning mappings for each word length and each language (see Figure 3).

Overall semantic distance is positively correlated with phonological distance ( $r = 0.04$ ) significantly more than what would be expected by chance ( $p < .001$  across all lengths and all languages). Thus we replicate the pattern observed among the 101 lexicons: similar wordforms tend to be more semantically similar than distinct wordforms. This is not the result of morphological similarity here since we looked only at monomorphemes in these four languages.

## 4 General Discussion

We have shown that across 101 languages, similar sounding words tend also to be more semantically similar above and beyond what could be expected by chance (an extension of Monaghan et al. (2014)

in English). In order to remove the contribution of morphology from this correlation, we conducted the same analysis on the set of monomorphemic lemmas of a restricted number of languages and found exactly the same pattern of results. This suggests that the pattern of clumpiness in the lexicon may be in part explained by form-meaning regularities, over and beyond morphological regularity, across a large range of typologically different languages.

What could be the reasons of form-meaning regularity in the lexicon? One possibility is that form-meaning regularity is due to etymology. Etymology is an important source of regularity in form-meaning mappings: certain words are historically related or derived from other words in the lexicon (even when the lexicon is restricted to morphologically simple words). For example, ‘skirt’ and ‘shirt’ are historically the Old Norse and Old English form of the same word, whose meanings have since diverged. Similarly, the presence of local sound-symbolism (e.g., the phonesthemes *gl-* in English) may drive the correlation. Yet, previous work showed that neither etymological roots nor small clusters of sound symbolic words were sufficient to account for the pattern of systematicity observed across the English lexicon (Monaghan et al., 2014). Though this needs to be confirmed for the languages under study here, this suggests a global pattern of form-meaning systematicity across the whole lexicon over and above etymological roots.

Another possibility is that form-meaning regularity is carried by the grammatical category of the words. Even though we looked at monomorphemes, words from the same grammatical category share phonological features (Cassidy & Kelly, 1991; Kelly, 1992), such that nouns sound more similar to other nouns and verbs to other verbs (see also Mahowald, Dautriche, Gibson, Christophe, & Piantadosi (*submitted*)), and are overall more semantically closer to words of the same grammatical category (e.g., verbs are more likely to map onto actions and nouns onto objects). Such systematic form-meaning mappings may be helpful during language learning to cue grammatical categories (Monaghan et al., 2011) and may be one of the outcomes of language transmission and evolution (Kirby et al., 2008) such that the optimal structure of the vocabulary may be one that incorporates form-meaning regularities at the large scale of the lexicon.

Still, the prevalence of wordform similarity in the lexicon conflicts in theory with communicative efficiency. Imagine a language that displays an extreme pattern form-meaning regularity where similar and frequent concepts such as CAT and DOG will be associated with similar wordforms such as ‘feb’ and ‘fep’ respectively. These words will be easily confused since their forms differ only from one phoneme and their meanings are similar. Nevertheless, we observed a correlation between semantic similarity and phonological distance. Perhaps, then, semantically similar words are not as confusable as one might suspect. Indeed, context is typically sufficient to disambiguate between meanings, since adult speakers use many cues when processing spoken sentences (e.g. prior linguistic context Altmann & Kamide (1999); visual information Tanenhaus et al. (1995); speaker Creel et al. (2008)). As a result, finer-grained contextual information may be sufficient most of the time for adults’ listeners to distinguish between phonologically similar words.

To our knowledge, with 101 languages in the sample, this is the largest cross-linguistic analysis showing a correlation between semantic similarity and phonological similarity among monomor-

phemic words showing evidence of systematicity in form-meaning mappings beyond morphological regularity (at least for Dutch, English, French and German). Ultimately, the results here suggest a functional advantage to having lexicons in which there is a positive correlation between phonetic and semantic similarity.

### References

- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*.
- Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80(2), 290–311.
- Bloomfield, L. (1933). *Language*. New York: Henry Holt.
- Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). “bouba” and “kiki” in namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. *Cognition*, 126(2), 165–172.
- Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30(3), 348–369.
- Childs, G. T. (1994). African ideophones. *Sound symbolism*, 178–204.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106(2), 633–664.
- de Saussure, F. (1916). *Course in general linguistics*. Open Court Publishing Company.
- Fourtassi, A., & Dupoux, E. (2013). A corpus-based evaluation method for distributional semantic models. *ACL 2013*, 165.
- Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806.
- Hamano, S. (1998). *The sound-symbolic system of japanese*. ERIC.
- Harris, Z. S. (1954). Distributional structure. *Word*.
- Hockett, C. (1960). The origin of speech. *Scientific American*, 203, 88–96.
- Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130298.
- Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99(2), 349–364.
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084), 1049–1054.

- Kim, K.-O. (1977). Sound symbolism in Korean. *Journal of Linguistics*, 13(01), 67–75.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686.
- Landauer, T., & 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(2), 211.
- Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862.
- Lin, D. (1998). An information-theoretic definition of similarity. In *ICML* (Vol. 98, pp. 296–304).
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2), 159–165.
- Mahowald, K., Dautriche, I., Gibson, E., Christophe, A., & Piantadosi, S. (submitted). Lexical clustering in efficient language design.
- Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. (submitted). Cross-linguistic effects of frequency on wordform similarity.
- Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. *Corpora in language acquisition research: History, methods, perspectives*, 139–164.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3), 325–347.
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language. *Philosophical Transactions of the Royal Society B*.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3), 516–524.
- Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.
- Rehurek, R., & Sojka, P. (2010). *Software framework for topic modelling with large corpora*.
- Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, 90(1), 413–422.
- Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49(6), 1175–1192.

- Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2), 191–211.
- Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (*in press*). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*.
- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632.
- Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4), 491–504.
- Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

## A Appendix: Dataset of 101 lexicons from Wikipedia

We started with lexicons of 115 languages from their Wikipedia databases (<https://dumps.wikimedia.org>). We then excluded languages for which a spot-check for non-native (usually English) words in the top 100 most frequent words in the lexicon between 3 and 7 characters revealed more than 80% of words were not native. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the 3-7 letter words in Chinese Wikipedia are often English. However, we included languages like Korean in which words generally consist of several characters. After these exclusions, 101 languages remained.<sup>6</sup> We analyzed the data both with and without these exclusions, and the exclusions do not significantly affect the overall direction or magnitude of the results. The languages analyzed included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, there are 12 language families represented as well as a Creole and 4 constructed languages (Esperanto, Interlingua, Ido, Volap) that have some speakers. (The analysis is qualitatively the same after excluding constructed languages.) The languages analyzed are shown in Tables 4 and 5.

To get a sense of how clean these Wikipedia lexicons are, we randomly sampled 10 languages for which we then inspected the 100 most frequent words and an additional 100 random words to look for intrusion of English words, HTML characters, or other undesirable properties.

For the top 100 words in the lexicons of the 10 sampled languages, we found at most 3 erroneous words. For the same languages, we also inspected a randomly selected 100 words and found that the mean number of apparently non-intrusive words was 93.5 (with a range from 85 to 99). The most common intrusion in these languages was English words.

**West Germanic:** Afrikaans, German, English, Luxembourgish, Low Saxon, Dutch, Scots, Yiddish, Alemannic; **Goidelic:** Irish, Scottish Gaelic; **Brythonic:** Breton, Welsh; **Hellenic:** Greek; **South Slavic:** Bulgarian, Macedonian, Serbo-Croatian, Slovene; **Albanian:** Albanian; **Iranian:** Central Kurdish, Persian, Kurdish, Mazandarani, Tajik; **Romance:** Aragonese, Asturian, Catalan, Spanish, French, Galician, Italian, Lombard, Neapolitan, Occitan, Piedmontese, Portuguese, Romanian, Sicilian, Venetian, Walloon; **West Slavic:** Czech, Polish, Slovak; **Armenian:** Armenian; **Italic:** Latin; **North Germanic:** Danish, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmal), Swedish; **Baltic:** Lithuanian, Latvian; **Indo-Aryan:** Fiji Hindi, Marathi, Urdu, Bosnian, Croatian, Punjabi, Serbian; **East Slavic:** Belarusian, Russian, Ukrainian; **Frisian:** West Frisian

**Table 4:** Table of Indo-European languages used, language families in bold.

<sup>6</sup>We excluded: Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, Kannada

**Austronesian:** Minang, Amharic, Indonesian, Malay, Sundanese, Cebuano, Tagalog, Waray-Waray, Buginese, Javanese; **Altaic:** Mongolian, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Turkish, Tatar, Uzbek; **creole:** Haitian; **Austroasiatic:** Vietnamese; **Kartvelian:** Georgian; **Niger-Congo:** Swahili, Yoruba; **Vasonic:** Basque; **Afro-Asiatic:** Malagasy; **Quechuan:** Quechua; **Semitic:** Arabic, Egyptian Arabic, Hebrew; **Korean:** Korean; **Uralic:** Estonian, Finnish, Hungarian; **Tai:** Thai; **constructed:** Esperanto, Interlingua, Ido, Volap

**Table 5:** Table of non-Indo-European languages used, language families in bold.

## B Appendix: Comparison between LSA and Wordnet

We additionally compared the Pearson correlations between semantic distance and phonemic distance across different measures of semantic distance: (a) 1 minus the cosine distance between co-occurrence vectors obtained by training a LSA model on the English Wikipedia and (b) several measures relying on WordNet structure to produce a score to quantify the distance between two concepts. Table 6 shows such a comparison for the 3702 nouns of the English phonemic lexicon using the Wordnet *path* measure (the minimum path length between two concepts in the WordNet network) and WordNet *lin* information content measure (Lin, 1998). Overall all semantic distance measures show the same qualitative pattern for every word length: there seems to be a positive correlation between semantic similarity and phonological distance in the English lexicon showing that semantically similar nouns tend also to be phonologically similar.

word length	LSA (cosine)	wordnet (path)	wordnet (lin)
3 letters	.021 ***	.018 ***	.012 ***
4 letters	.013 ***	.013***	.014 ***
5 letters	.011 ***	.002 *	.022 ***
6 letters	.004 ***	.011 *	.037 *
7 letters	0.01 **	.015 **	.017 *

**Table 6:** Comparison of Pearson correlations coefficients for each word length using different semantic similarity distances.

## 2.3 Summary and Discussion

Across languages and language families, I find that the structure of word form similarity in the lexicon is not arbitrary but is the result of functional pressures, such that 1) there is more phonological similarity in the lexicon than expected by chance (**section 2.1**), and 2) semantically related pairs of words are also orthographically related (**section 2.2**).

### Null lexicons

The methodological contribution of this work is a new way of assessing linguistic structure through the creation of random baselines that provide a null hypothesis for how the linguistic structure should be in the absence of communicative and cognitive pressures. This could be used to assess the distribution of other linguistic phenomena.

Note however that the chance level depends entirely on how we define it. Here our question required us to model the phonotactics of the language sufficiently well in order to create null lexicons that make plausible assumptions about the true generative process of words (contra a random typing model, Howes, 1968). Yet this is not to say that a *n-gram* model on phones is the right way to think about words and certainly some better models of phonotactics already exists (BLICK, Hayes, 2012) but their adaptation to other languages than English is not easy.<sup>10</sup>

In addition, the chance level I defined is language-specific, that is our best non-word generative model was trained on a single language to generate non-words of that language. One could imagine that the chance level should be more general, i.e., trained across all possible languages, since our primary question concerns the presence of general, thus non language-specific, cognitive pressures on the set of word forms. Yet, this may be unrealistic in practice. Since languages have different phonotactic constraints that may influence the space of possible words for that language, comparing measures of word form similarity in a global random baseline to each individual languages may not be representative of the constraints, in the absence of functional pressures, for that language.

### Limitations

One limitation of this work is the use of orthographic corpora in **section 2.2**. Orthography is often taken as an approximation for phonology, yet it would be useful to dispose of

<sup>10</sup>Note that at least for English, generating random baselines using non-words that are phonotactically "good" according to BLICK does not change the pattern of results obtained in **section 2.1**.



lexicons that are transcribed phonemically and are morphologically tagged. However since 101 languages were used in this study, this, to some extent, takes away this concern.

Another perhaps more important limitation regarding the scope of this dissertation is the focus on word form similarity alone. Indeed, we are still left with the question of how much homophony is present in languages? The methodology developed here, however, cannot give a satisfying answer to that question. Recall that to compare word form similarity in a given language to chance level, I focused on monomorphemes; but homophones are not only limited to the set of monomorphemes: for instance in French "porte" means both the noun *door* and the verb *to carry* conjugated in the present tense singular. Thus, homophony may arise in morphologically complex words (e.g., *to carry<sub>pres.sing</sub>*) that do not appear in the list of monomorphemes. One could generate null lexicons for the set of word forms in a given language, encompassing thus all possible forms. Yet, this would go beyond the assumptions of our current generative model (a *n-gram* on phones), as it is not equipped with a mechanism to generate morphologically complex words.

### Conclusions

In sum, across a large range of measures, we showed that there is more phonological similarity in the lexicon than expected by chance and that these words tend to be correlated with greater semantic similarity. This suggests that there is a pressure for the lexicon to be more clumpy, that is, to be more compressible. As we argued, such a pressure may be beneficial not only for speakers, as it minimizes articulatory effort and relieves memory load (as less sound sequences are used), but also for learners as it may help them to learn some aspects of their language (as word form regularity may be helpful in segmenting words from speech and more systematic form-meaning mappings help category formation). However, such properties may be detrimental for some other aspects of learning as it suggests that children may have a hard time differentiating these forms to attribute them meanings. One important question is thus: how do children manage to learn such a lexicon?

## 3 Learning confusable and ambiguous words

Carey (1978b) has described word learning as starting with a process where children "*flag new word!*" upon hearing a phonological sequence with no current lexical entry". Indeed, one feature of novel words is that they are often composed of unfamiliar sequences of sounds. However, a new word can be phonologically similar or even identical to a word that already exists in the child's lexicon and yet, be associated with a novel meaning. For instance, the child may already know the word "sheep" but needs to be able to identify that "ship", a minimally different word form, is a different word despite the phonetic variability of the speech signal. Similar-sounding words present thus learners with a challenging case where they need to find the right balance between phonological tolerance, to recognize known words, and phonological sensitivity, to be able to learn these new words.

Not only must children be able to identify novel *word forms* in the signal to consider them as candidate lexical entries, they also must be able to identify novel *meanings* even when the word form is identical to a form they already know, as in the case of homophones. For instance the child may already know that "bat" means bat-animals and be confronted with a sentence such as "aluminum bats are much easier to swing when compared to wooden bats". How does the child determine that "bat" is used here to refer to a baseball-bat and not an animal-bat? Homophony thus presents learners with a unique word learning situation where they cannot rely on the signal alone to determine whether a phonological form is a candidate for a novel entry in the lexicon as a new word.

In this chapter, I investigate whether 18- to 20-month-old toddlers take into account other factors than phonology when determining what counts as a new word. Specifically, in **section 3.1**, I test whether toddlers take into account the syntactic context to determine whether a novel phonological neighbor of a word they know could be interpreted as a novel word (learning "tog" when "dog" is already in their lexicon). In **section 3.2**, I take these results further, and look whether children's ability to learn homophones depends on the syntactic or semantic context they are presented in. Finally in **section 3.3**, I evaluate whether similar-sounding words and homophones that exhibit properties that make them learnable by children are more represented in the lexicon of natural languages than similar-sounding words and homophones that are harder to learn.



Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

## Cognition

journal homepage: [www.elsevier.com/locate/COGNIT](http://www.elsevier.com/locate/COGNIT)

## Learning novel phonological neighbors: Syntactic category matters

Isabelle Dautriche<sup>a,b,\*</sup>, Daniel Swingley<sup>c</sup>, Anne Christophe<sup>a,b</sup><sup>a</sup>Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Études Cognitives (École Normale Supérieure – PSL Research University), Paris, France<sup>b</sup>Maternité Port-Royal, AP-HP, Faculté de Médecine Paris Descartes, France<sup>c</sup>Department of Psychology and Institute for Research in Cognitive Science, University of Pennsylvania, United States

## ARTICLE INFO

## Article history:

Received 14 December 2014

Revised 3 June 2015

Accepted 4 June 2015

## Keywords:

Word learning

Lexical access

Phonetic sensitivity

Language acquisition

## ABSTRACT

Novel words (like *tog*) that sound like well-known words (*dog*) are hard for toddlers to learn, even though children can hear the difference between them (Swingley & Aslin, 2002, 2007). One possibility is that phonological competition alone is the problem. Another is that a broader set of probabilistic considerations is responsible: toddlers may resist considering *tog* as a novel object label because its neighbor *dog* is also an object. In three experiments, French 18-month-olds were taught novel words whose word forms were phonologically similar to familiar nouns (noun-neighbors), to familiar verbs (verb-neighbors) or to nothing (no-neighbors). Toddlers successfully learned the no-neighbors and verb-neighbors but failed to learn the noun-neighbors, although both novel neighbors had a familiar phonological neighbor in the toddlers' lexicon. We conclude that when creating a novel lexical entry, toddlers' evaluation of similarity in the lexicon is multidimensional, incorporating both phonological and semantic or syntactic features.

© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

Many of the words young children hear are not yet in their vocabulary. As a result, in everyday conversation toddlers must often decide whether a given word-form corresponds to a word they already know, or to a word to be learned. In principle, children could accomplish this by checking to see if each utterance can be parsed entirely into a sequence of familiar words. If it cannot, perhaps the unidentified portions correspond to new words.

The problem, of course, is to define what counts as an instance of a familiar word and what does not. Different instances of a given word do not all sound the same. Talkers have different voices and varying accents (e.g., Labov, 1966); words sound different depending on the phonetic context they appear in (e.g., Holst & Nolan, 1995), and speakers routinely blend sounds together or omit completely entire sounds and even whole syllables of words (e.g., Ernestus & Warner, 2011; Johnson, 2004). Such phenomena are present in the speech parents direct to their children (e.g., Bard & Anderson, 1983). Drawing the boundary between the set of acceptable instances of a word, and the instances that cannot correspond to that word, is complex.

Traditionally, it is said to be the role of the language's phonology to define the set of phonetic differences that distinguish words,

to resolve these ambiguities. If words are represented as phonological descriptions adequate for maintaining contrast, and heard utterances are converted into phonological descriptions during speech comprehension, a simple comparison procedure should be adequate for identifying new words. If a word-form in the utterance fails to line up with any word-forms in the lexicon, this means that a new word has been heard.

This might not work for children, for several reasons. Children's skills of phonetic categorization are inferior to adults' and undergo substantial refinement well into the school years, despite the rapid progress toward language-specific perception made in infancy (e.g., Hazan & Barrett, 2000; Kuhl, 2004). In many cases children may not successfully characterize utterances in phonological terms. And even when they can, it is not clear that children understand that phonological distinctions are meant to signal lexical distinctions. Although children recognize words more easily when the words are spoken with their canonical pronunciations than when spoken with deviant pronunciations (e.g., Swingley, 2009), this does not imply that the mispronunciations are interpreted as novel words (e.g., White & Morgan, 2008). Toddlers do resist interpreting some discriminable, but not phonological, differences as contrastive (Dietrich, Swingley, & Werker, 2007; Quam & Swingley, 2010), which suggests some sophistication in relating speech and the lexicon. But being wary of interpreting a non-phonological distinction as if it could distinguish words does not imply the inverse skill of readily interpreting phonological distinctions as contrastive.

\* Corresponding author at: Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure, 29 rue d'Ulm, P.J., 75005 Paris, France.

E-mail address: [isabelle.dautriche@gmail.com](mailto:isabelle.dautriche@gmail.com) (I. Dautriche).

One study tested whether toddlers could use a single-feature phonological distinction to assign a novel meaning to a word-form that sounded similar to a very familiar one (Swingley & Aslin, 2007). 19-month-olds were shown a novel object, which was repeatedly named using clear (hyperarticulated) speech. In some cases the novel name given was similar to a familiar word (e.g., *tog*, similar to *dog*), and in some cases it was not (e.g., *shang*, not similar to any words children knew). Children were tested using a fixation procedure in which pictures of two novel objects were presented on a screen, and one of the pictures was labeled using its novel name (e.g., “Look at the {*tog*, *shang*”). Fixation to the named picture was used to index learning of the word. In two experiments, children were able to learn words that sounded very different from the other words in their vocabularies (like *shang*), but children did not learn the phonologically similar words (like *tog*). For some of the items tested, children of the same age had previously shown discrimination of the nonce label and its familiar counterpart, so perceptual discrimination *per se* was apparently not at issue (e.g., Swingley & Aslin, 2002).

Why might this be? One possibility is that phonological competition *alone* is the problem. The lexical entry of *dog* might be activated by the phonologically neighboring form *tog*, interfering with children’s considering the possibility that a new word was being offered. This explanation of the experimental results is consistent with a view that children first adopt a phonological criterion of similarity, which apparently requires a greater difference than the single phonological feature tested in the experiment, and proceed accordingly.

Another possibility is that a broader set of probabilistic considerations is responsible. Not only is *tog* phonologically similar to a well-entrenched word, but it is also syntactically and semantically similar: both *tog* and *dog* are *nouns* referring to *objects*. Considering that the 18-month-old lexicon is relatively sparse in both phonology and semantics (Swingley & Aslin, 2007; but see Coady & Aslin, 2003 for older children) the appearance of a novel word that is both phonologically similar to, and somewhat semantically close to, a familiar word, might seem implausible to children, leading them to suppose that the novel word might in fact be a rather dubious instance of the familiar word.

Adults too may, in some conditions, fail to interpret a one-feature phonological change as lexically meaningful (e.g., White, Yee, Blumstein, & Morgan, 2013). Under conditions in which the speech signal and the referential context are less clear (conditions which prevail quite generally in human communication), adults can interpret phonologically novel word forms as instances of known words (e.g., Cole, Jakimik, & Cooper, 1978). For example, upon hearing “this singer has a beautiful *foice*”, listeners are more likely to misperceive *foice* as an instance of *voice*. In such a case, adults find it plausible that the word *voice* has been uttered since both the syntactic and the semantic context constrained their lexical search toward singing-related nouns. Although /f/ and /v/ are lexically contrastive in English, the difference in voicing value may plausibly be interpreted as noise rather than indicating the presence of a new word in this particular context. In arriving at an analysis of spoken sentences, adults use a diverse array of sources of information: the physical context (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995); the prior linguistic context (e.g., Altmann & Kamide, 1999); pragmatic expectations supported by the discourse (e.g., Nieuwland & Van Berkum, 2006); and idiosyncrasies of the speaker (e.g., Creel, Aslin, & Tanenhaus, 2008). In a sense, all of these are needed while interpreting speech because speakers are aware that listeners have this information at their disposal, and frequently provide only just enough phonetic information to allow the listener to resolve the intended meaning given the context (e.g., Hawkins, 2003).

These findings with adults highlight the importance of factors other than phonology in interpreting speech. Yet it is open to question whether toddlers identify words primarily using phonological criteria, or whether, like adults, they take into consideration a broader range of probabilities in judging the likelihood that a phonological distinction implies a novel word. In support of the latter, here we present evidence that toddlers evaluate other factors than phonological features, such as syntactic or semantic features, when evaluating the possibility that a novel sequence of sounds is a new word.

We started from Swingley and Aslin (2007)’s result that toddlers failed to learn new object labels that sounded similar to familiar object labels. In three experiments, French 18-month-olds were taught object labels that were phonological neighbors of a familiar *noun* (a noun-neighbor, as *tog* was, for *dog*), neighbors of a familiar *verb* (a verb-neighbor, like teaching *kiv*, a neighbor of *give*) or no-neighbors (such as *shang*). The noun-neighbor and the verb-neighbor were both phonologically similar to a familiar word in children’s lexicon. But only the noun-neighbor was also semantically and syntactically similar to its neighbor; the verb-neighbor was not. If children take into account semantic or syntactic likelihoods when interpreting novel neighbors, verb-neighbors should be perceived as sufficiently distinct from any word in the lexicon to be easily assigned a novel object meaning – just like no-neighbor words – whereas noun-neighbors are expected to suffer from the competition with the familiar noun and be hard to learn. In contrast, if children fail to learn both noun-neighbors and verb-neighbors, this would indicate that children stake everything on phonological similarity in deciding whether a word-form is a new word.

## 2. Experiment 1

Experiment 1 sought to replicate Swingley and Aslin (2007)’s results showing that phonological neighbors of a familiar noun (noun-neighbors) are hard for toddlers to learn. We taught French 18-month-olds two novel object labels: a noun-neighbor (e.g., “ganard,” a neighbor of “canard” *duck*) and a no-neighbor (e.g., “torba”). Word learning was then evaluated using a language-guided looking method (Fernald, Zangl, Portillo, & Marchman, 2008; Swingley, 2011). Children were presented with the two novel objects and heard sentences that named one of the pictures (e.g., “il est où le ganard?” *where is the ganard?*). An above-chance proportion of looks toward the target picture after word onset was taken as evidence that the word had been learned.

### 2.1. Method

#### 2.1.1. Participants

Sixteen French 18-month-olds participated in the study, ranging in age from 17;19 (months; days) to 18;23, with a mean of 18;13 ( $SD = 0;8$ ; 7 girls). An additional 8 children were not included in the sample because they refused to wear the sticker necessary for eye-tracking ( $n = 3$ ), fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 3$ ), no increase in average proportion of looks toward the target during familiar-word trials ( $n = 1$ )<sup>1</sup> and hearing problems reported by the parents ( $n = 1$ ). The attrition rate was somewhat higher than

<sup>1</sup> Following previous pilot experiments, before commencing testing here we decided on an exclusion criterion of rejecting children who looked at the target on average less than 55% of the time (from word onset until the end of the trial) over the 8 familiar-word trials. Individual time courses were inspected to be sure to not reject children who only quickly looked toward the target instead of having a sustained look; there were no such cases. This criterion was applied blind to condition performance.

expected, which we tentatively attribute to the children's having just participated in a separate study involving 10 min of active searching for toys.

### 2.1.2. Apparatus, procedure and design

Each child was taught two words: one novel word whose phonological form was similar to a noun they know (noun-neighbor) and one that had no phonological neighbor in their lexicon (no-neighbor). Before coming to the lab, parents filled out a questionnaire of vocabulary including all the neighbors of the test words. This was to ensure that children would be taught a novel word that neighbored a noun they already knew. Toddlers sat on their parent's lap about 70 cm away from a television screen. Their eye movements were recorded by an Eyelink 1000 eye-tracker. We used a 5-point calibration procedure. Once the calibration was judged acceptable by the experimenter, the experiment began.

The experiment was composed of two phases: a teaching phase and a testing phase. During the teaching phase, children were presented with a first introductory video and 4 teaching videos, 2 for each novel word. Which of two objects the noun-neighbor referred to was counterbalanced across toddlers. The order of presentation of the teaching video was interleaved between the two words and counterbalanced across toddlers. After presentation of the teaching videos, the test phase started as soon as children looked at a fixation cross.

The test phase was composed of 16 trials: 8 trials with familiar words and 8 test trials with novel words, 4 per novel word. Each trial started with the simultaneous presentation of two pictures on the right and left sides of the screen. Two seconds later, the audio stimuli started: (“Regarde le [target], tu le vois le [target]?” Look at the [target], Do you see the [target]?). The trial ended 3.5 s after the first target word onset. Trials were separated by a 1 s pause. No immediately consecutive trials presented the same pictures or words. Target and distractor pictures appeared the same number of times on the right and the left side of the screen. Target side did not repeat more than two times on consecutive trials.

The whole experiment lasted about 5 min.

### 2.1.3. Materials

**Novel words.** All novel words were bisyllabic and started with a stop consonant. We used the *Lexique* database (New, Pallier, Brysbaert, & Ferrand, 2004) to identify 4 novel words whose phonological form was similar to a common noun (noun-neighbors) and 4 novel words that had no phonological neighbor in children's lexicon (no-neighbor).

The 4 noun-neighbors differed from their real-noun counterparts by inversion of the voicing value of the initial consonant. The four words were *pallon*, *ganard*, *gochon*, and *pateau* (/palɔ̃/, /ganɑ̃/, /goʒɔ̃/, /pato/), neighbors of *ballon*, *canard*, *cochon*, and *bateau* (ball, duck, pig, boat) which are all likely to be known by children of that age according to CDI reports from previous studies. Frequency counts of the familiar nouns in a corpus of child directed speech (Lyon corpus, Demuth & Tremblay, 2008) were as follows (frequencies were calculated on the phonological forms of these words thus conflating the singular and the plural of the nouns): *ballon*, 201; *canard*, 179; *cochon*, 180; *bateau*, 105. Parents were also asked to report any other neighbors likely to be known by their children. Both *ganard* and *gochon* had no other phonological neighbors than the familiar noun competitor that we chose, but children knew one or two other familiar nouns close to *pallon* (*salon*, living-room) and *pateau* (*gateau*, cake; *rateau*, rattle). Thus, *ganard* and *gochon* had a phonological neighborhood density of 1 in children's lexicon; *pallon* had a phonological neighborhood density of 2 and *pateau*, on average, of 2.25. All noun-neighbors had no

other neighbors in another syntactic category likely to be known by children.

The no-neighbors were generated from an *n*-phone model trained on the *Lexique* database with the constraint that they should be phonologically similar to less than 2 low frequency words in the French lexicon. Four phonotactically legal bisyllabic no-neighbor words were chosen: *prolin*, *barlié*, *torba*, *lagui* (/pʁolɛ̃/, /barljé/, /tɔ̃rba/, /lagi/). To ensure that children would not learn the no-neighbors better than the noun-neighbors simply because these words were phonotactically easier (Graf Estes & Bowen, 2013; Storkel, 2001), we ensured that the sound-to-sound probabilities were on average higher for the noun-neighbors than for the no-neighbors (cumulative bigram log-probability  $\log P = -7.07$  for no-neighbors;  $\log P = -5.59$  for noun-neighbors; this was calculated using a *n*-phone model on the set of word types in the French lexicon, taken from the *Lexique* database; New et al., 2004).

Noun-neighbors and no-neighbors were yoked in pairs, such that each child would learn one of 4 pairs of words: (*prolin*, *gochon*), (*barlié*, *pateau*), (*torba*, *pallon*), (*lagui*, *ganard*). Children were all presented with a noun-neighbor for which they had a phonological neighbor in their lexicon according to parental report.

**Novel objects.** The novel objects were two unfamiliar animals. One resembled a pink white-spotted octopus with an oversized head. The other looked like a rat with bunny ears and a trunk (see Fig. 1). At the end of the experiment, parents were asked whether their child was familiar with either animal; all parents said no.

**Teaching videos.** Word teaching was done on a television screen. A first introductory video showed a speaker (the last author) playing with a car (*une voiture*) and labeling it several times in a short story. This video was intended to familiarize children with the procedure, showing them that the speaker would talk about the object she manipulates. The teaching phase included four short videos of about 30 s each. In each video, the same speaker talked about the novel object she was playing with and labeled it 5 times using one of the novel words. The noun-neighbor word was used in two videos, and the no-neighbor word in the other two. In total, toddlers heard each novel word 10 times.

**Testing stimuli.** The pictures were photographs of objects on a light gray background. For familiar trials, we chose 8 objects that children of that age are likely to know: *voiture*, *banane*, *poussette*, *chaussure*, *chien*, *poisson*, *cuillère*, *maison* (car, banana, baby-stroller, shoe, dog, fish, spoon, house). Pictures were yoked in pairs (e.g., the banana always appeared with the car). For test trials, the pictures of the two novel animals were always presented together (as in Fig. 1).

The audio stimuli consisted of the sentences “Regarde le [target], tu le vois le [target]?” (Look at the [target], Do you see the [target]?) or “il est où, le [target]? Regarde le [target].” (Where is the [target]?)

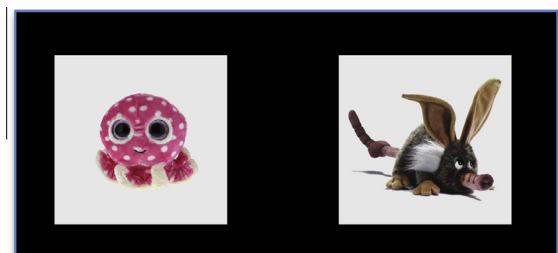


Fig. 1. Novel objects used during the experiments.



Look at the [target]!) where [target] was the target word and was pronounced two times in a given trial. All sentences were recorded by the last author (the same speaker as in the videos). The average duration of the novel words was 610 ms for the noun-neighbors and 598 ms for the no-neighbors.

#### 2.1.4. Measurement and analysis

Gaze position on each trial was recorded via an eye-tracker with a 2 ms sampling rate. We inspected the time course of eye movements from the onset of the first occurrence of the target word (“Look at the [target].”) until the end of the trial. One recurrent problem when analyzing continuous time series is the choice of a window of analysis. Time can be made categorical by choosing a series of consecutive time windows tailor-made to the data and then performing separate analysis on each time window; or, more frequently in the infant literature, by imposing a single, large window to maximize the chances of observing a change in eye fixations. While the first option leads to a problem of multiple comparisons, the second often conflates response time and accuracy, thus resulting in a loss of information (Swingley, 2011) and both options are subject to biases inherent to window selection. Here, in order to test whether toddlers had learned each novel word we conducted a cluster-based permutation analysis (Maris & Oostenveld, 2007) to find a time window where we observed a significant increase in looks toward the target picture. This type of analysis, originally developed for EEG data, is free of time-window biases, preserves the information available in the time series and is able to cope well with multiple comparisons.<sup>2</sup>

The cluster-based analysis works as follows: at each time point we conducted a one-tailed<sup>3</sup> *t*-test on fixations to the target compared to chance (0.5). All fixation proportions were transformed via the arcsin square function to fit better the assumptions of the *t*-test. The means and variances were computed over subjects within conditions. Adjacent time points with a significant effect ( $t > 2$ ;  $p < .05$ ) were grouped together into a cluster. Each cluster was assigned a single numerical value measuring its size, and defined as the sum of all the *t*-values within the cluster (intuitively, a cluster is larger if it contains time-points for which the two conditions are very significantly different, and if it spans a longer time-window). To obtain the probability of observing a cluster of that size by chance, we conducted 1000 simulations where conditions (novel label, chance) were randomly assigned for each trial. For each simulation, we computed the size of the biggest cluster identified with the same procedure that was used for the real data (sum of all the *t*-values within a cluster of significant *t*-values). Clusters in the children's data were taken as significant if the probability of observing a cluster of the same size or bigger in the randomized data was smaller than 5% (that is, if a cluster that big was observed in less than 50 cases over 1000), corresponding to a *p*-value of 0.05.

It is important to note that the criterion for including a time bin in a cluster ( $t > 2$  in our study) is independent of the process which assesses cluster significance, so it does not affect the likelihood of a false positive. Yet, it does have an influence on the size of the time window that one can find. If the threshold is low then the time window will be wider. However the same low threshold will be applied to the randomized data as well, such that the chance of

getting a bigger cluster will also increase under the null hypothesis, thus maintaining the rate of false positive under 0.05.

In addition, to test whether there was a significant difference between conditions (whether children found the noun-neighbor harder to learn than the no-neighbor), we conducted an additional cluster-based permutation analysis in which clusters were formed on the basis of paired two-tailed *t*-tests comparing the looking proportions between conditions at each time point.

Thus in total, we conducted three cluster-based analyses: one for each word condition (no-neighbor; noun-neighbor) comparing the average proportion of looks toward the target picture for each test word to 50%, and one comparing the looking proportions between conditions.

#### 2.2. Results

Fig. 2 shows the average proportion of looks toward the target picture for familiar and test words (noun-neighbor and no-neighbor) from the onset of the first target word (*Regarde le [target], tu le vois le [target]? Look at the [target], do you see the [target]?*) until the end of the trial.

Children showed recognition of the no-neighbor (green curve in Fig. 2) but not the noun-neighbor (red curve in Fig. 2). The cluster-based permutation analyses revealed that they fixated the correct picture above chance when asked to look at the no-neighbor (2178–2568 ms time-window, green-shaded area in Fig. 2;  $p < .05$ ) but stayed around chance level in the case of the noun-neighbor (no significant time window found by the cluster-based permutation analysis). The difference between the recognition of the no-neighbor and the noun-neighbor was significant in the time window ranging from 2044 to 2852 ms ( $p < .01$ , gray-shaded area in Fig. 2). Thus, children learned the no-neighbor but not the noun-neighbor ( $p = 0.19$ ). We also observed that the recognition of the no-neighbor occurred with a delay of about 900 ms compared to the recognition of familiar words (gray curve in Fig. 2), a finding we will return to later in discussing subsequent experiments.

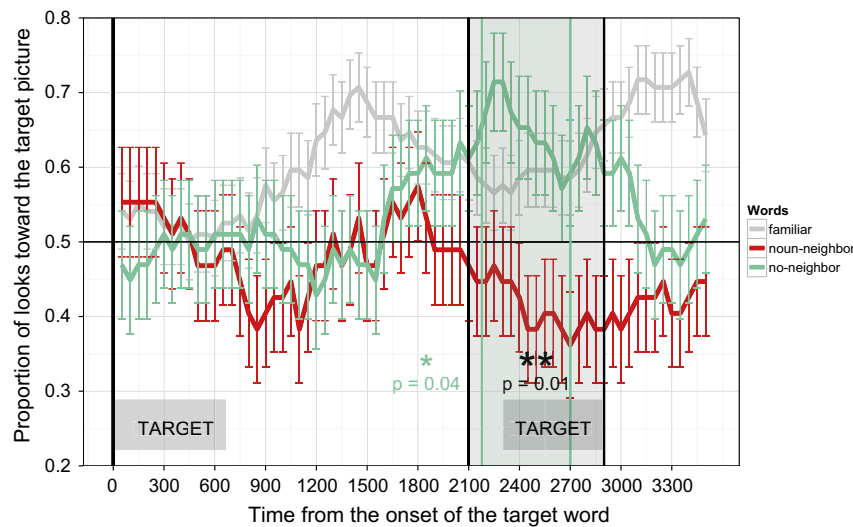
#### 2.3. Discussion

After a brief but intensive exposure to a pair of novel words, French 18-month-olds performed better when tested on a novel no-neighbor (e.g., “torba”) than on a novel noun-neighbor (e.g., “ganard,” a novel neighbor of “canard,” *duck*) in a word recognition task. When presented with the two novel objects on the screen and hearing a sentence labeling one of them, children correctly recognized the no-neighbor but failed to recognize the noun-neighbor. This may be surprising given that children of that age can infer the meaning of a word via mutual exclusivity (e.g., Markman & Wachtel, 1988). That is, if they learnt the no-neighbor then they should be able to infer the meaning of the noun-neighbor by process of elimination during the test phase. Yet several studies have shown that mutual exclusivity effects seem to disappear when children are confronted with a novel word that is phonologically similar to a familiar word (e.g., Merriman, Marazita, & Jarvis, 1995). When children heard “Regarde le pallon!” *look at the pallon*, they may start looking for a “ballon”, the phonological competitor, and go back and forth between the two images to find the closest match.

This result replicates Swingley and Aslin's (2007) findings with English and Dutch 18-month-olds, showing that children of that age find it hard to learn a phonological neighbor of a familiar noun. We will now test toddlers' ability to learn a novel neighbor of a familiar verb.

<sup>2</sup> Following a reviewer's suggestion, we also compared the results of this method with more traditional methods such as the salience-corrected fixations (see Swingley, 2011, for a discussion of this measure) and the more recent growth curve analysis method (Barr, 2008). Both methods of analysis lead to the same conclusion and are available upon request to the first author.

<sup>3</sup> Note that we used one-tailed *t*-tests because our hypothesis was directional as we expected a higher-than-chance looking proportion when the word was recognized, yet using two-tailed *t*-test did not change the pattern of results. In particular none of the clusters of fixations below chance level passed the permutation test.



**Fig. 2.** Proportion of looks toward the target picture from the onset of the target word (Regarde le [target], tu le vois le [target]? *Look at the [target], do you see the [target]?*) for the noun-neighbor (red), the no-neighbor (dark green) and the familiar words (gray). Toddlers performed significantly better on the no-neighbor than on the noun-neighbor: they successfully learned the no-neighbor (green shaded time window) as shown by an increase of looks toward the correct picture, but failed to learn the noun-neighbor, staying at chance level. The gray-shaded time window corresponds to the region where toddlers were more likely to look at the target picture when asked for the no-neighbor than when asked for the noun neighbor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

### 3. Experiment 2

In Experiment 1, 18-month-olds failed to learn a noun-neighbor (e.g., “ganard”) showing that they are sensitive to its phonological similarity to a known word in their lexicon (“canard” *duck*). In Experiment 2, we build on this failure to investigate whether toddlers are able to appreciate other factors than phonological features when deciding whether a given word-form corresponds to a novel word or is an instance of an already-known word. In the same task, we taught French 18-month-olds two novel object labels: one with a phonological neighbor from a different syntactic category (verb-neighbor; e.g., “barti” neighbor of “parti” *gone*) and one with no neighbors (no-neighbor e.g., “torba”).

Following Experiment 1, we expected children to learn the no-neighbor. If toddlers were to fail to learn the verb-neighbor, this would be evidence that phonological similarity to a known word is sufficient to prevent toddlers from considering the verb-neighbor as a novel word. On the contrary, if toddlers were to succeed in learning the verb-neighbor – just as they learn the no-neighbor word – this would indicate that children take into account not only phonological likelihood but also syntactic and/or semantic likelihood when deciding whether a given word-form denotes a novel word or not.

#### 3.1. Method

##### 3.1.1. Participants

Sixteen French 18-month-olds, ranging from 17;18 to 19;4 with a mean of 18;8, ( $SD = 0;15$ ; 8 girls) took part in this experiment. Twelve additional children were replaced because of refusal to wear the sticker necessary for eye-tracking ( $n = 2$ ), fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 6$ ), experimenter error ( $n = 2$ ), no increase in the average proportion of looks toward the target during familiar trials ( $n = 2$ ).

##### 3.1.2. Apparatus, procedure and design

Similar to Experiment 1 except that this time children were taught two words with a phonological neighbor: one novel word

whose phonological form was similar to a verb they know (verb-neighbor) and one whose phonological form was not familiar to any word they know (no-neighbor).

##### 3.1.3. Materials

Similar to Experiment 1 except for the set of novel words used in the teaching phase.

**Novel words.** We chose 4 novel words whose phonological forms were similar to a common verb (verb-neighbors) and 4 novel words that had no neighbors in toddlers’ lexicons (no-neighbors). Words were selected following the same procedure as in Experiment 1.

The 4 no-neighbors were the same as in Experiment 1. The 4 verb-neighbors were chosen following the same criteria as for the noun-neighbors: they were all bisyllabic words starting with a stop consonant and differing from a common verb in the voicing of that initial consonant: *barti*, *dombé*, *gassé*, *tonné* ( $/b\alpha r\alpha ti/$ ,  $/d\ddot{o}b\epsilon/$ ,  $/g\alpha s\epsilon/$ ,  $/t\ddot{o}n\epsilon/$ ), being neighbors of *parti*, *tombé*, *cassé*, *donné* (*gone*, *fallen*, *broken*, *given*) and having no other neighbors known to children, according to parental report.<sup>4</sup> The verb-neighbors were modeled on the past participle forms of the verbs. This form was chosen because it is very common (the most frequent morphological form for 3 of the 4 verbs; Demuth & Tremblay, 2008) and because it is bisyllabic. Frequency counts of the familiar nouns in a corpus of child directed speech (Lyon corpus, Demuth & Tremblay, 2008) were as follows: *parti*, 112; *tombé*, 411; *cassé*, 263; *donné*, 252. These counts were calculated on the phonological form of the neighbor in parental input and thus included the past participle form of the verb as well as the infinitive form (except for *parti* whose infinitive form is not homophonic to the past participle).

The average duration of the novel words in the test sentences was 620 ms for the verb-neighbors and 598 ms for the no-neighbors.

<sup>4</sup> The word-form *parti* is a homophone and can be used as a noun meaning “political party” or “part”. None of the children we tested knew these meanings, based on parental report.



## 3.1.4. Measure and analysis

Similar to Experiment 1.

## 3.2. Results

Eye movement results were analyzed as in Experiment 1. As shown in Fig. 3, toddlers started to look more toward the target picture for both the verb-neighbor (blue curve) and the no-neighbor (green curve) soon after the end of the target word. The cluster-based permutation analyses found a significant time-window where the proportion of looks to the target was significantly above chance for the verb-neighbor condition (1092–1746 ms, blue-shaded time-window;  $p < .01$ ) as well as for the no-neighbor condition (950–1254 ms, green-shaded time-window;  $p < .05$ ). There was no significant difference between conditions (verb-neighbor, no-neighbor) suggesting that one word was not recognized better than the other (no time window found).

## 3.3. Discussion

Toddlers successfully learned a verb-neighbor in Experiment 2 and failed to learn a noun-neighbor in Experiment 1, although both words had a familiar phonological neighbor in toddlers' lexicon. Performance on the verb-neighbor and the no-neighbor were not different (in Experiment 2), suggesting that the phonological resemblance to a familiar word in their lexicon did not impact their understanding of the verb-neighbor as a novel word, compared to the no-neighbor. Here, toddlers were not overwhelmed by the phonological similarity to a known word, presumably because the likelihood that the novel noun, “un barti”, would be considered as a plausible variant of the familiar verb, “parti” gone, is low. This suggests that toddlers integrate semantic and/or syntactic likelihood in the process of creating a novel lexical entry.

Contrary to Experiment 1, where the recognition of the no-neighbor started about 1500 ms after word onset, in Experiment 2 there was no delay in the recognition of the novel words: toddlers recognized the verb-neighbor and the no-neighbor at about the same time as they recognized the familiar words (roughly 600 ms after word onset). The crucial difference between Experiment 1 and Experiment 2 is the presence of a novel word that is difficult to learn, the noun-neighbor. One possibility is

thus that the presence of the noun-neighbor in the test trials hindered the recognition of the no-neighbor in Experiment 1 (cf. Swingley & Aslin, 2007). Recall that during the test trials, toddlers were presented with the two novel objects and asked to select the noun-neighbor half of the time, and the no-neighbor the other half. This may have confused toddlers, if the link between the noun-neighbor and the novel object was difficult to make for them. If the presence of the noun-neighbor is a major reason why we observe a delay in Experiment 1, then we might expect that the recognition of any novel word, including the verb-neighbor, should be slowed down when taught together with a noun-neighbor.

## 4. Experiment 3

In Experiment 3, we seek to directly compare toddlers' performance for learning a noun-neighbor versus learning a verb-neighbor in a within-subjects design. Using the same experimental materials and basic design from the prior experiments, here we taught children two novel object labels: one noun-neighbor as in Experiment 1 and one verb-neighbor as in Experiment 2. Following Experiment 1 and Experiment 2, we expected toddlers to succeed in learning the verb-neighbor, and to fail to learn the noun-neighbor.

## 4.1. Method

## 4.1.1. Participants

Sixteen French 18-month-olds were tested (ranging from 17;26 to 18;29 with a mean of 18;8,  $SD = 9$ , 7 girls). An additional 8 children were not included in the final sample because of refusal to wear the sticker necessary for eye-tracking ( $n = 3$ ), fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 2$ ), no increase in average proportion of looks toward the target during familiar trials ( $n = 2$ ), or no knowledge of the phonological neighbors ( $n = 1$ ).

## 4.1.2. Apparatus, procedure and design

Similar to Experiment 1 except that this time children were taught two words with a phonological neighbor: one novel word whose phonological form was similar to a verb they know

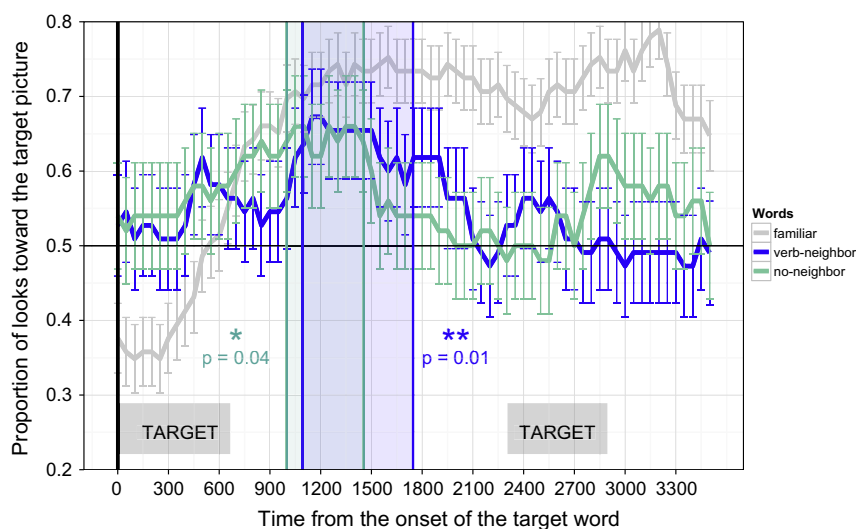


Fig. 3. Proportion of looks toward the target picture from the onset of the target word (Regarde le [target], tu le vois le [target]? Look at the [target], do you see the [target]?) for the verb-neighbor (blue), the no-neighbor (green) and the familiar words (gray). Toddlers successfully learned both the verb-neighbor (blue-shaded time window) and the no-neighbor (green-shaded time window). There was no significant difference between the verb-neighbor and no-neighbor conditions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(verb-neighbor) and one whose phonological form was similar to a noun they know (noun-neighbor).

#### 4.1.3. Materials

The materials were similar to those of Experiment 1 except for the set of novel words used in the teaching phase.

**Novel words.** We used the 4 noun-neighbors used in Experiment 1 (novel words whose phonological form was similar to a noun) and the 4 verb-neighbors from Experiment 2 (novel words whose phonological form was similar to a verb). All 4 verb-neighbors had only one phonological neighbor known to the children, the two noun-neighbors, *ganard* and *gochon* had exactly one phonological neighbor and the two other noun-neighbors, *pallon* and *pateau* had on average 1.75 phonological neighbors in children's lexicon. The noun-neighbors' cumulative bigram log-probability was slightly lower than the one of the verb-neighbors ( $\log P = -5.59$  for noun-neighbors;  $\log P = -6.21$  for verb-neighbors). Verbs were, on average, 56% more frequent than the nouns, based on counts from the Lyon corpus of French child-directed speech (Demuth & Tremblay, 2008).

Verb-neighbors and noun-neighbors were yoked in pairs, such that each child would learn one of 4 pairs of words: (*pallon*, *gassé*), (*ganard*, *tonné*), (*gochon*, *barti*), (*pateau*, *dombé*). Children were taught a verb-neighbor and a noun-neighbor for which they knew the phonological neighbors, according to parental report.

The average duration of the novel words in the test sentences was 620 ms for the verb-neighbors and 610 ms for the noun-neighbors.

#### 4.1.4. Measure and analysis

Similar to Experiment 1.

#### 4.2. Results

As can be seen in Fig. 4, we replicated the pattern of results observed in Experiment 1 and 2. Toddlers successfully learned the verb-neighbor: they looked toward the correct picture at above-chance rates for the verb-neighbor (1660–2930 ms, blue-shaded time-window;  $p < .01$ ) but resisted learning the noun-neighbor, showing no recognition of the novel word (no

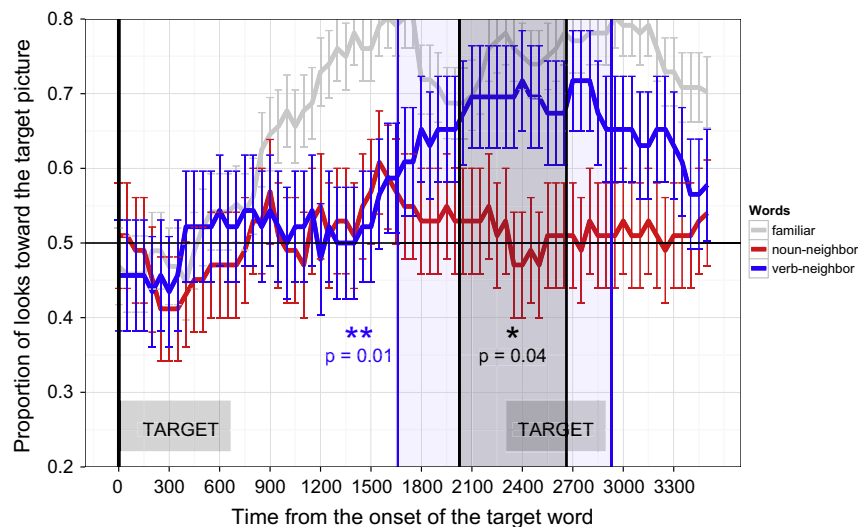
significant time-window found;  $p = 0.80$ ). As a result, toddlers recognized the verb-neighbor significantly better than the noun neighbor (2024–2662 ms, gray-shaded time-window;  $p < .05$ ), showing that toddlers' processing of these phonological neighbors is significantly different, depending on the syntactic category of the neighboring word.

Note that toddlers recognized the verb-neighbor at about 1500 ms after target word onset, a delay comparable with the time course of recognition of the no-neighbor in Experiment 1. This suggests that the presence of the noun neighbor in these two experiments slowed down the recognition of the other novel word.

#### 4.3. Discussion

Children failed to learn an object label when it was a phonological neighbor of a noun they knew (as in Experiment 1) but succeeded when it was a phonological neighbor of a verb they knew (as in Experiment 2). Experiment 3 replicated this phenomenon within children, ruling out variation among children as a possible explanation of the difference between the results of Experiments 1 and 2. The failure to learn a noun-neighbor cannot be attributed to phonological competition alone, because both the noun-neighbor and the verb-neighbor had a frequent phonological neighbor in the children's lexicon. The most likely explanation, then, is that children take into account semantic and/or syntactic likelihood when interpreting a novel word.

An unexpected observation was that toddlers were slowed down in their recognition of newly-taught words which were tested at the same time as noun-neighbors: no-neighbors in Experiment 1, and verb-neighbors in Experiment 3. Given that both no-neighbors and verb-neighbors were observed to be recognized quickly in Experiment 2 (in the absence of the noun-neighbor), this suggests that the presence of the object the noun-neighbor referred to was sufficient to delay recognition of the other object in Experiments 1 and 3. To our knowledge, no prior study has reported a delay in novel word recognition while learning phonological noun-neighbors, though Swingley and Aslin (2007) did find that performance on familiar nouns was affected by children's (unsuccessful) exposure to novel noun-neighbors. The confusion triggered by noun-neighbors is consistent with our interpretation in terms of toddlers estimating



**Fig. 4.** Proportion of looks toward the target picture from the onset of the target word (Regarde le [target], tu le vois le [target]? Look at the [target], do you see the [target]?) for the verb-neighbor (blue), the noun-neighbor (red) and the familiar words (gray). Toddlers recognized the verb-neighbor significantly above chance level (blue-shaded time window) but failed to recognize the noun-neighbor. As a result, their performance was significantly better for the verb-neighbor than for the noun-neighbor within the time region shaded in gray. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the likelihood that a given word-form corresponds to a novel lexical entry or an existing one: When the odds are against the creation of a novel lexical entry – because the phonological and semantic/syntactic similarity appear too great to be coincidental – children attempt to reconcile the deviant phonological form with the existing lexical entry. This process might involve trying to figure out in what circumstances the observed phonological change may be licensed by the native language, as well as what extension of the word's meaning may encompass both the familiar referent and the novel referent. Whatever the exact nature of this process, it may contribute to the confusion that is observed.

### 5. General discussion

A simple rule for determining whether a spoken word corresponds to a word in the lexicon is to compare the phonological form of this word to the phonological forms in the lexicon. When the forms match, the word's identity is known; if the heard form matches no known words, it may be a candidate for entry into the lexicon as a new word. Part of the function of a phonological system is to ensure that this matching process can work, at least when the signal itself is carefully produced and clearly heard.

Young children apparently have some difficulty in operating with their developing phonology in this way. For example, they have some trouble learning similar-sounding words (e.g., *Stager & Werker, 1997*). *Swingley and Aslin's (2007)* finding, in which the proximity of a novel word like *tog* to the familiar word *dog* made the novel word hard to learn, showed that lexical activation processes, whose function is to account for the sounds in an utterance in terms of the correct string of words (and possibly in spite of mispronunciation or misperception), can conflict with word learning processes. In that study, the novel object intended as the referent for the novel word (e.g., *tog* had no real resemblance to its familiar counterpart (*dog*). Similarly in the present experiments, the plush pink octopus object labeled here as *pateau* is not a reasonable member of the category of boats (*bateaux*). Such considerations suggest that children's difficulty might be purely a matter of phonological proximity: *tog* (or *pateau*) is simply too close to a familiar word to permit ready detection as a novel form.

The present work shows that this is not the case. In fact similarity of the novel referent to its competitor's denotation matters. The similarity of a *pateau* (plush octopus) to a *bateau* (boat) is not great, but it is greater than the similarity of the same octopus to the meaning of *parti* or *cassé* (gone, broken) because it shares the same syntactic category (nouns) and the same broad semantic category (toys).

With the present data we cannot determine whether it is the syntactic difference that is most important, or the semantic one, or both; but to get some purchase on this question we looked for potential item differences within the experiments. Recall that two of our familiar nouns were animals (*canard*, duck, and *cochon*, pig) and two were artifacts (*ballon*, ball, and *bateau*, boat). Given how fundamental the distinction between animals and artifacts is, even to infants (e.g., *Setoh, Wu, Baillargeon, & Gelman, 2013*), our items *ganard* and *gochon* (as plush animals) were probably semantically much closer to their competitors *canard* and *cochon* than *pateau* or *pallon* (as plush animals) were to *bateau* or *ballon*. If semantic similarity were the main driving force behind our results, we would expect children to learn *pateau* and *pallon* more easily than *ganard* and *gochon*. Indeed, inspection of the results revealed a nonsignificant trend toward better performance on *pateau* and *pallon* than *ganard* and *gochon* for participants in Experiment 1 and 3. It could be interesting to vary this similarity systematically with more items; here, there are confounding features of the words, such as uneven phonotactic probability and neighborhood density, that

make interpreting this trend difficult. At any rate, although the present result cannot tell us whether the semantic or the syntactic difference between the neighbors and their familiar competitors play a role, it is clear that 18-month-olds take more into consideration than the sounds alone.

Might some unmeasured difference between our noun and verb neighbors be responsible for our effects? For example, frequent words are generally recognized more readily than infrequent words (e.g., *Solomon & Postman, 1952*). Could it be that the interfering effect of noun neighbors derives from stronger activation of those words, and not from the direct consequences of semantic or syntactic distance from the novel objects? This might happen if the phonological form of nouns were more strongly established in children's lexicons than the phonological form of verbs: indeed, verbs in French occur in more varied phonological forms than nouns, due to morphology. However, as we reported, the exact phonological forms of the verbs we used were, on average, 56% more frequent than the nouns in parental input, and as frequent as the nouns in children's production, based on counts from the Lyon corpus of French child-directed speech (*Demuth & Tremblay, 2008*), and parents in each experiment reported that their children knew the neighboring words.

Another possibility is that independently of frequency, the meanings of the noun neighbors were better entrenched in children's lexicons than the meanings of the verb neighbors, leading to greater interference. Yet if more entrenched representations lead to greater interference, we would also expect that verb-neighbors should lead to more interference than a word with no neighbor. That's not what we observe: children learnt verb-neighbors just as well as no-neighbors in Experiment 2. So while we cannot dismiss the possibility that more entrenched representations of the familiar nouns over the familiar verbs plays a role in children's interpretation of novel neighbors, our data provide little support for this hypothesis. Thus our main point would still hold, namely that semantic or syntactic similarity plays a role in children's interpretation of novel neighbors.

Thus, we propose that young toddlers' evaluation of similarity in the lexicon in the context of word learning is multidimensional, incorporating both phonological and semantic and/or syntactic features. The plausibility of a syntactic contribution to the results is supported by prior studies showing that children, like adults, use the sentence context to build on-line expectations about the syntactic category of an upcoming word (e.g., *Bernal, Dehaene-Lambertz, Millotte, & Christophe, 2010*). Toddlers as young as 14 to 18 months expect a noun to follow a determiner and expect a verb after a personal pronoun (*Cauvet et al., 2014; He & Lidz, 2014; Kedar, Casasola, & Lust, 2006; Shi & Melancon, 2010; Zangl & Fernald, 2007*). For instance, *Cauvet et al. (2014)* showed that French 18-month-old toddlers trained to turn their head for a known target noun ("la balle" *the ball*), responded more often to the word "balle" when it appeared in a noun context ("une balle" *a ball*) than when it appeared (incorrectly) in a verb context ("on balle" *they ball*). In fact, in that last case, they did not turn their head more often than for control sentences which did not contain the target word at all. In our study, when children were processing the syntactic context of our sentences, they should have expected a noun at the point where the verb-neighbor, *parti*, was heard. Since *parti* occurred in a context where the familiar verb *parti* was not expected, one possibility is that children did not access the familiar verb *parti* at all, and therefore that they did not even notice the similarity with a word present in their lexicon. Another possibility is that children may have accessed *parti* despite the nominal context because the integration of contextual cues is limited by toddlers' developing executive function abilities (e.g., *Khanna & Boland, 2010*, but see *Rabagliati, Pylykänen, & Marcus, 2013*) Yet the presence of additional cues provided by the learning

situation (i.e., repetition of the verb-neighbor in a noun context, presence of a novel object and contingent gaze cues from the speaker whenever the verb-neighbor was used) may render the possibility that the verb-neighbor *barti* is a novel word a more plausible alternative than for the noun-neighbor.

Children process words in context, just as adults do. In particular, the linguistic context plays a prominent role in constraining lexical access and thus in estimating the likelihood that a novel phonological word-form is a novel lexical entry rather than a variant of a known word. Manipulating the linguistic context by placing a verb-neighbor in a noun syntactic frame indicated to children that a new meaning was appropriate for the novel word-form. We would expect the same result to be found using other syntactic frames (e.g., pronouns) or by doing the symmetric manipulation (i.e., presenting a noun-neighbor in a verb syntactic frame). This suggests also that the linguistic context may play a role in learning several meanings for perfectly identical word-forms (homophones; see also Casenhiser, 2005), a possibility that we are currently exploring.

Learning neighbors of familiar words is difficult for toddlers, but as we showed, this difficulty disappears when the novel words appear in contexts that are sufficiently different from their known neighbors (either syntactically or semantically or both). If learnability influences language changes, then this constraint on early lexical acquisition might have a long-lasting impact on the overall structure of the lexicon. Do lexicons avoid similar-sounding words? And when similar-sounding words do occur, are they preferentially distributed across syntactic or semantic categories to improve their learnability (and their recoverability)? Recent studies observed that not only do mature lexicons contain many similar-sounding words, perhaps even more than would be expected by chance (Dautriche et al., 2014), but there is also a tendency for phonologically similar words to be more semantically similar than phonologically distinct words (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Dautriche et al., 2014). In sum, lexicons appear to favor similar-sounding words which are semantically related.

At first sight this might appear at odds with the present study, yet there are two ways to resolve this apparent inconsistency. First, a rich literature suggests that similar-sounding words display a range of advantages for language use: they are easier to remember, produce and process for adults (e.g., Vitevitch, 2002; Vitevitch, Chan, & Roodenrys, 2012; Vitevitch & Stamer, 2006) and preschoolers (e.g., Storkel & Lee, 2011; Storkel & Morrisette, 2002). Also, greater systematicity of form-to-meaning mappings could facilitate the grouping of words into categories (Padraic Monaghan, Christiansen, & Fitneva, 2011). Overall, the processing benefits for similar-sounding words might outweigh an initial learning disadvantage. Second, an early disadvantage for learning similar-sounding words may not actively exert a selective pressure for words that are more phonologically dissimilar because children may eventually manage to learn neighbors through repeated exposure. Thus, instead of being reflected in the static organization of the lexicon, the constraint we uncovered may be reflected in the dynamics of early lexical growth: early in children's lexical development, novel words may preferably be added whenever they can be easily distinguished from already existing words along at least one dimension (phonological, syntactic, and/or semantic). Previous work looking at the growth of the lexicon focused on how either phonological similarity or semantic similarity influences word learning, but not on potential interactions between several dimensions (Carlson, Sonderegger, & Bane, 2014; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Steyvers & Tenenbaum, 2005; but see Regier et al., 2001).

In sum, our work shows that 18-month-old children process words in context, using multiple sources of information.

Phonological similarity alone does not serve as a kind of filter that collapses phonological neighbors in advance of meaningful analysis. Rather, 18-month-olds appear to evaluate simultaneously the phonological, syntactic and/or semantic likelihood of this sequence of sounds being a new word.

### Acknowledgements

This work was funded by grants from the Région Ile-de-France, Fondation de France, LabEx IEC (ANR-10-LABX-0087), IdEx PSL (ANR-10-IDEX-0001-02), the ANR 'Apprentissages' (ANR-13-APPR-0012) to AC, NIH grant R01-HD049681 to DS, and a PhD fellowship from DGA to ID.

### References

- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3), 247–264.
- Bard, E. G., & Anderson, A. H. (1983). The unintelligibility of speech to children. *Journal of Child Language*, 10(02), 265–292.
- Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4), 457–474.
- Bernal, S., Dehaene-Lambertz, G., Millotte, S., & Christophe, A. (2010). Two-year-olds compute syntactic structure on-line. *Developmental Science*, 13(1), 69–76.
- Carlson, M. T., Sonderegger, M., & Bane, M. (2014). How children explore the phonological network in child-directed speech: A survival analysis of children's first word productions. *Journal of Memory and Language*, 75, 159–180.
- Casenhiser, D. M. (2005). Children's resistance to homonymy: An experimental study of pseudohomonyms. *Journal of Child Language*, 32(2), 319–343.
- Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, 10(1), 1–18.
- Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, 30(2), 441–469.
- Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, 64(1), 44–56.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106(2), 633–664.
- Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi S. T. (2014). Lexical clustering in efficient language design. *Oral presentation in AMLaP*. Edinburgh, Scotland.
- Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, 35(1), 99.
- Dietrich, C., Swingle, D., & Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences*, 104(41), 16027–16031.
- Ernestus, M., & Warner, N. (2011). An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(3), 253–260.
- Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*, 113–132.
- Graf Estes, K., & Bowen, S. (2013). Learning about sounds contributes to learning about words: Effects of prosody and phonotactics on infant word learning. *Journal of Experimental Child Psychology*, 114(3), 405–417.
- Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31(3), 373–405.
- Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, 28(4), 377–396.
- He, A. X., & Lidz, J. (2014). Development of the verb-event link between 14 and 18 months. *Paper presentation accepted for the 39th Boston University Conference on Language Development (BUCLD)*. Boston, MA: Boston University.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, 20(6), 729–739.
- Holst, T., & Nolan, F. (1995). The influence of syntactic structure on [s] to [ʃ] assimilation. *Papers in Laboratory Phonology, IV*, 315–333.
- Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29–54).
- Kedar, Y., Casasola, M., & Lust, B. (2006). Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Development*, 77(2), 325–338.
- Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, 63(1), 160–193.
- Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843.
- Labov, W. (1966). *The social stratification of English in New York city*. Cambridge University Press.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190.



### 3 Learning confusable and ambiguous words

- Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2), 121–157.
- Merriman, W. E., Marazita, J., & Jarvis, L. (1995). Children's disposition to map new words onto new referents. *Beyond Names for Things: Young Children's Acquisition of Verbs*, 147–183.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3), 325–347.
- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130299–20130299.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3), 516–524.
- Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7), 1098–1111.
- Quam, C., & Swingle, D. (2010). Phonological knowledge guides 2-year-olds' and adults' interpretation of salient pitch contours in word learning. *Journal of Memory and Language*, 62(2), 135–150.
- Rabagliati, H., Pytkäinen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, 49(6), 1076–1089.
- Regier, T., Corrigan, B., Cabasaan, R., Woodward, A., Gasser, M., & Smith, L. (2001). The emergence of words. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 815–820).
- Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, 110(40), 15937–15942.
- Shi, R., & Melancon, A. (2010). Syntactic categorization in French-learning infants. *Infancy*, 15(5), 517–533.
- Solomon, R. L., & Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, 43(3), 195.
- Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640), 381–382.
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41–78.
- Storkel, H. L. (2001). Learning new words phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research*, 44(6), 1321–1337.
- Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2), 191–211.
- Storkel, H. L., & Morrisette, M. L. (2002). The Lexicon and Phonology: Interactions in Language Acquisition. *Language, Speech, and Hearing Services in Schools*, 33(1), 24–37.
- Swingle, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536), 3617–3632.
- Swingle, D. (2011). The looking-while-listening procedure. *Research Methods in Child Language: A Practical Guide*, 29–42.
- Swingle, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, 13(5), 480–484.
- Swingle, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, 54(2), 99.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.
- Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 735–747.
- Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, 67(1), 30–44.
- Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, 21(6), 760–770.
- White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59(1), 114–132.
- White, K. S., Yee, E., Blumstein, S. E., & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, 68(4), 362–378.
- Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, 3(3), 199–231.

## 3.2 Learning homophones: syntactic and semantic contexts matter

As I showed, children find it difficult to learn a phonological neighbor of a word they know when the known word and the novel word are also close syntactically (and/or semantically). Yet, this difficulty is reduced when phonological neighbors are presented in a different *context*: When a phonological neighbor of a verb is presented in a noun context, there is no interference from the known verb. Thus, the syntactic context helps children to distinguish between two minimally different words and to accept that they possess two different meanings (section 3.1). Here, I investigate whether in the absence of any phonological cue to distinguish between the meanings of a pair of homophones, children use contextual cues to identify when a given word form is likely to instantiate a new meaning. In other words, can children learn a second meaning for a word form they already know?

Previous studies looking at homophone comprehension in preschoolers suggest that 5-year-olds find it difficult to use the semantic context of the sentence to derive the meaning of a homophone (Beveridge & Marsh, 1991; Campbell & Bowe, 1977). When presented with the less common meaning of a pair of homophones (e.g., the "wing" of a castle) in a disambiguating context, children failed to use the semantic information to interpret the word form and instead accessed its primary meaning in more than 80% of the cases (e.g., the "wing" of a bird) (Campbell & Bowe, 1977). While this may suggest at first that contextual cues might not be very relevant to acquire the meanings of homophones, since children appear not to use them, there are several important limitations to these studies. Importantly, children's lexical knowledge of the less common meaning of a pair of homophones was not assessed prior to test. In fact, when tested on two meanings they know, 4-year-olds had no problem to correctly interpret homophones when the context selected either the primary meaning or the subordinate meaning of the word form (Rabagliati, Pylkkänen, & Marcus, 2013).

Most relevantly, several developmental studies investigated the acquisition of homophony where children were taught a second meaning for a word they know (e.g., "door" would label an unfamiliar object) (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997). In these studies, preschoolers listened to stories where the known word was used to label a novel referent in relation with pictures they were presented with. For instance, they could see a picture with a tree and the novel referent on the top of it and hear "Way up in the tree, Tommy saw a *door*" (the word was used once or twice in a sentence across these 3 studies) and asked to point to the referent of the word in a set of picture (including or not the known referent of the word, i.e., a door). Mazzocco (1997) showed that toddlers and preschoolers found it easier to learn a novel meaning for a novel word form than a second meaning for a known word form. In this task where they were asked to map a known word (e.g. "door") to a referent that already had a label (e.g. a clown), children had to

learn not only that this word form have several meanings, but also to learn that it is a synonym for the known referent (the clown). Their difficulty with homophones may thus just reflect their difficulties in learning synonyms (children are biased to assume that word extensions are mutually exclusive, the *mutual exclusivity bias*, see Markman & Wachtel, 1988). Yet, even when the known referent was replaced by an unfamiliar referent (e.g., a tapir), children still found it more difficult to learn a second meaning for a known word than to learn a meaning for a novel word (Casenhiser, 2005; Doherty, 2004).

All in all, these studies suggest that children find it difficult to learn homophones when (1) they are used in a context that does not bring sufficient evidence that a new meaning is appropriate and when (2) the learning situation is not sufficiently ecological. Crucially, previous studies did not manipulate whether the learning situation they proposed to children could plausibly lead them to conclude that an additional meaning was likely for that known word. They also relied on rather poor learning instances, where the word was used only a limited number of time (once or twice) in stories that failed to provide sufficient evidence to constrain a potential novel meaning for the word. Indeed it may not be sufficient for children to know that "door" refers to an unfamiliar animal (e.g., a tapir), as the referent of the word is still ambiguous (e.g., Quine, 1960). They may need additional evidence about what properties are associated with the new meaning of "door" (e.g., living in the jungle, eating berries) to narrow down its meaning.

Here I propose that children's ability to learn homophones crucially depends on the context they are presented in. That is, homophones may be easier to learn when the two meanings are made sufficiently *distant* by the context in which they are used. In 4 experiments, I manipulate different sources of information that may help children to identify when a novel meaning for a known word is appropriate:

- **Syntactic and semantic distance:** The previous results with neighbors (see section 3.1) suggest that using a known word with a different syntactic category may increase the likelihood of adding a new meaning for this word. In **Experiment 1**, toddlers will be taught an animal label that is homophonous with a known verb (e.g., "an eat"). If children take into account semantic and/or syntactic likelihoods when identifying whether a given word form would instantiate a novel meaning, they should be able to learn verb-homophones (taught as animal labels).
- **Semantic distance:** Experiment 1 cannot determine whether it is the syntactic distance or the semantic one that is most important when learning homophones. In **Experiment 2 and 3**, children will be taught animal labels that are homophonous with a known artifact (semantically distinct) or with a known animal (semantically close). If a semantic distinction between meanings of a pair of homophones is sufficient to learn them, children should have no problem learning artifact-homophones (taught as animal labels).

- **Syntactic distance:** Similarly, the syntactic distance alone may be sufficient to learn homophones. In **Experiment 4**, children will be taught animal labels that are homophonous with a known animal (e.g., "un<sub>masculine</sub> chat", *a cat*) but associated with a different gender (e.g., "une<sub>feminine</sub> chat"). Since grammatical gender is not a semantic property in French, it can thus be used to test the effect of syntactic distance independently from semantic distance.
- **Neighborhood density:** Besides syntax and semantics, another important factor in language learning and processing is the phonological context of the word in the lexicon. Dense neighborhoods slow down lexical retrieval in adults compared to sparse neighborhoods (Vitevitch & Luce, 1998), hence it may be easier to learn a secondary meaning for a word form with high phonological neighborhood density than for a word form with low phonological neighborhood density simply because the primary meaning of the word would be less accessible. In **Experiment 5**, children will be taught an animal label that is homophonous with the label of a known artifact which has no neighbor in children's lexicon. If neighborhood density has an influence, such sparse artifact-homophones should be more difficult to learn than dense artifact-homophones.

In these experiments, I will test 20-month-olds, thus younger children than in previous studies (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997), for several reasons: (1) toddlers of this age already use the different sources of information that I propose to investigate here during lexical processing (noun vs. verb, e.g., Cauvet et al., 2014; semantic relations, e.g., Arias-Trejo & Plunkett, 2013; gender cues, e.g., Van Heugten & Christophe, in press; neighborhood density, e.g., Newman, Samuelson, & Gupta, 2008) and (2) they may already have acquired a certain number of homophone pairs (de Carvalho et al., 2014) suggesting that learning lexical ambiguities should be possible at this young age. In these experiments, toddlers will be taught homophones in the word learning task used in section 3.1, which has the advantage of presenting novel words in a richer context than in previous studies on preschoolers.

### 3.2.1 Experiment 1 - Manipulating the syntactic and semantic distance

Experiment 1 investigated whether the syntactic and/or semantic context would be sufficient to signal that a novel meaning for a known word was appropriate. Most relevantly, Casenhiser (2005) found that 4-year-olds find it easier to learn an additional meaning for a known word when used in a disambiguating linguistic context. As in section 3.1, I manipulate the syntactic and/or the semantic context by placing a known *verb* in a *noun* frame to label a novel animal (**a verb-homophone**; e.g., "an eat"). In such a case, both meanings are used in distinct syntactic contexts ("to eat" is a verb while "an eat" is a



noun) and they are also semantically distinct ("to eat" refers to an action and "an eat" refers to an animal). If toddlers evaluate the syntactic and/or semantic features of words when identifying novel words, verb-homophones should be perceived as sufficiently distinct from the known verb to be assigned a novel meaning.

#### Method

**Participants.** Thirty-two French 20-month-olds, ranging from 19;1 to 20;9 with a mean age of 20;3, (SD = 0;5; 13 boys) took part in this experiment. Five additional children were replaced because of fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 3$ ), experimenter error ( $n = 1$ ), born at less than 37 weeks' gestation ( $n = 1$ ).

**Apparatus, procedure and design.** Similar to the Experiments of section 3.1 except that this time toddlers were taught one novel word that was homophonous with a verb (a verb-homophone) and a novel word that did not resemble any word in the children's lexicons (a non-homophone).

**Material.** Similar to the Experiments of section 3.1 except for the set of novel words used in the teaching phase.

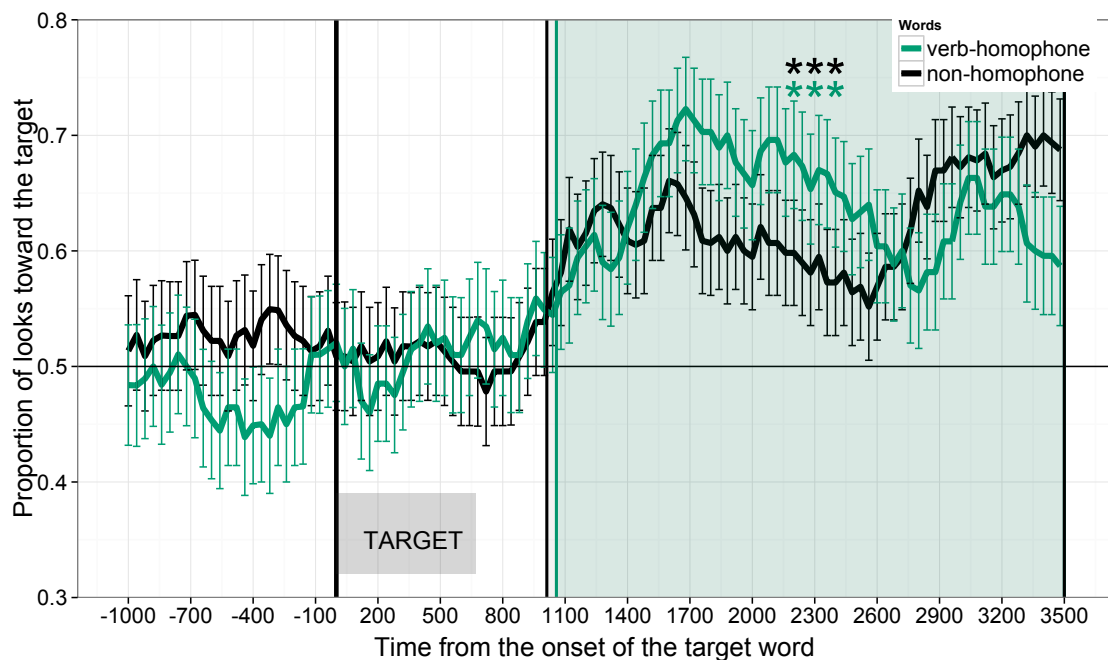
*Novel words.* I chose 4 phonological word forms of verbs that toddlers of that age are likely to know (according to the CDI of previous studies) and the 4 novel word forms used in Experiments of section 3.1.

The 4 non-homophones were all bisyllabic: "prolin", "barlier", "torba", "lagui" (/pʁɔlɛ̃/, /barljé, /tɔʁba/, /lagi/) and had no phonological neighbors with words that toddlers knew. The 4 verb-homophones were also all bisyllabic words: "manger", "tomber", "casser", "cacher" (/mãʒe/, /tõbe/, /kase/, /kaʃe/) meaning: *eat, fall, break* and *hide*. The verb-homophones were the infinitive and the past participle forms of the known verbs. These forms were used because they were bisyllabic and they are very common (the most frequent morphological form for all 4 verbs; Demuth & Tremblay, 2008). The average duration of the novel words in the test sentences was 640ms for the verb-homophones and 622ms for the non-homophones.

**Measure and Analysis.** Similar to the Experiments of section 3.1.

## Results

Figure 3.1 shows the proportion of looks towards the target picture from the beginning of the first target word ("Regarde le [target]! tu le vois le [target]" *Look at the [target]!* *Do you see the [target]?*) until the end of the trial.



**Figure 3.1:** Proportion of looks towards the target picture from the beginning of the first target word (do you see the [target]?) for the verb-homophone (in green) and for the non-homophone (in gray). Toddlers successfully learnt both novel words as they significantly increased their look towards the correct picture after target word onset.

Toddlers recognized both the verb-homophone (green curve) and the non-homophone (black curve). The cluster-based permutation analysis (Maris & Oostenveld, 2007) identified a significant time-window where toddlers looked significantly above chance (0.5) for the verb-homophone (from 1000ms after word onset,  $p < .001$ ; green shaded area in Figure 3.1) and for the non-homophone (from 1000ms after word onset,  $p < .001$ ; gray shaded area in Figure 3.1). There was no significant difference between conditions, suggesting that toddlers learnt both words equally well.

## Discussion

Toddlers successfully learnt the verb-homophone just as well as the non-homophone. This may be surprising given that several studies have shown that even preschoolers find it

difficult to associate a novel meaning to a known word (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997). Yet contrary to previous studies, toddlers may not have been disturbed by the re-use of a known verb, as the syntactic context, a noun frame, decreased the likelihood that the speaker could possibly be using the known meaning of the verb (see also Casenhiser, 2005). This suggests that children can use the syntactic context to identify whether a novel meaning is appropriate for a word, even when the word is already associated to a meaning.

One possibility is that toddlers did not even notice that a familiar verb was used in this experiment. Indeed, toddlers use the syntactic context to constrain lexical access: For instance they expect a noun after a determiner and a verb after a pronoun (e.g., Cauvet et al., 2014; Shi & Melancon, 2010; Zangl & Fernald, 2007). Thus, when presented with a noun phrase such as "C'est un manger!" (*this an eat!*), toddlers may simply not have accessed the familiar verb and this made it very easy to link an additional meaning to the word, just as easy as for the non-homophone.

Yet another possibility is that toddlers initially noticed that the verb they knew was used in an incorrect frame. Indeed, two-year-old children display an early left-lateralized brain response when an expected noun is incorrectly replaced by a verb (e.g., "je prends la **mange**" *I take the eat*, Bernal et al., 2010). Relevantly, ERP studies looking at adults' processing of lexical ambiguities in reading tasks found that the syntactic context alone was insufficient to constrain lexical access in the case of noun/verb homographs (e.g., the park/to park) as evidenced by a frontal negativity compared to unambiguous words (e.g., Lee & Federmeier, 2012). Yet additional semantic constraints on the meaning of the words eliminated the frontal negativity (e.g., Lee & Federmeier, 2009). Such a frontal component has been suggested to reflect the recruitment of frontally mediated meaning selection mechanisms needed to disambiguate noun-verb homographs in the absence of constraining semantics (e.g., Novick, Trueswell, & Thompson-Schill, 2010) suggesting that the syntactic context alone may be insufficient to suppress totally the inappropriate meaning of the word. Yet, in our experiment, because the verb is never used to convey its initial meaning (thus the verb meaning is never pre-activated) but repetitively used in a noun frame with a visual support (i.e., the novel referent), social support (i.e., the speaker looking contingently to the referent each time she uses the verb-homophone) and supplemented with information about its novel meaning (e.g., "Un manger, ça a des grandes oreilles" *Eats have big ears*), this may have increased the likelihood that the known meaning of the verb was inappropriate in that context and supported the identification of an additional meaning for the known verb form.

At any rate, the present results show that children have no problem learning noun-verb homophones in a supportive context. This has important consequences for theories of syntactic development looking at the acquisition of syntactic categories. One common assumption is that children may start grouping words into categories by observing their

distributional context: For instance, all words  $X$  appearing in the context [you  $X$  the] are likely to be verbs. To obtain adult-like categories (nouns, verbs, etc), learners would need to merge these context-based categories, possibly by grouping together the categories that share a portion of their categorized words (Mintz, 2003). For instance, [you  $X$  the] and [he  $X$  the] would be grouped together since words such as "drink", "eat", "give" would appear in both contexts. Yet words that appear in multiple categories should make distributional cues to category less effective. Learners could conflate distributions and create a single category that contains both nouns and verbs. In fact, noun/verb homophones have been cited as evidence against the logical possibility of learning grammatical categories from their distributions (Pinker, Lebeaux, & Frost, 1987). To get around this problem, two solutions have been proposed.

The first solution suggests that children may be sensitive to a (small) perceptual distinction between nouns and verbs such that they would be able to maintain two phonologically distinct representations of the same word form (e.g., park-NOUN and park-VERB Conwell, 2015; Conwell & Morgan, 2012). This is supported by several sources of evidence showing that there are prosodic differences between noun and verb homophones, such that noun tokens are longer than verb tokens (Conwell & Morgan, 2012; Shi & Moisan, 2008).<sup>11</sup> Following this hypothesis, learners would use this sensitivity to tackle the ambiguity problem, such that words that appear in more than one lexical category should not pose a problem for children.

The second solution suggests that children could start building syntactic categories by grouping words together according to their semantic category as soon as they start to know the meaning of basic words (from 9 months of age, Bergelson & Swingley, 2012, 2013) (Brusini, 2012; Gutman, Dautriche, Crabbé, & Christophe, 2014). For example, they could start grouping together "toy", "car", and "spoon" because they all refer to concrete objects, and "drink", "eat", and "give" because they all refer to actions. The knowledge of a few content words may allow learners to discover the distributional context in which they appear: They could notice that nouns/objects appear most of the time preceded by determiners and verbs/actions by pronouns or auxiliary, and subsequently use this information to categorize new words. Following this hypothesis, children rely on the context to attribute a category (or a meaning) to novel words.

Certainly, these two solutions are not mutually exclusive. What I showed here is that at 20 months of age, children are not confused by noun/verb homophones because they

<sup>11</sup>This is, however, not surprising given that nouns may be more likely than verbs to appear at the end of a prosodic phrase as in [the kids' *brush*] [is on the table] (where brackets indicate prosodic phrase boundaries) and thus is typically lengthened compared to verbs that are more likely to appear phrase-internally as in [the kids] [*brush* their teeth] (Delais-Roussarie, 1995). Thus, these prosodic differences may not be part of the phonological form of the word but depend on the position of the word in the sentence: Both nouns and verbs homophones appear to have the same duration when they are situated in the same position in the sentence (Sorensen, Cooper, & Paccia, 1978).

already process words in context, suggesting that noun/verb homophones may not be a major problem for syntactic development. Interestingly, similarly to toddlers in this study, several computational (Bayesian) models of syntactic category acquisition are able to deal well with noun-verb ambiguity (e.g., Goldwater & Griffiths, 2007; Parisien, Fazly, & Stevenson, 2008). Yet these models find it difficult to identify the syntactic category of a noun when there is a strong lexical bias in favor of the verb of the homophone pair (Parisien et al., 2008) while toddlers, as I just showed, have no problem in such cases.

Importantly, here, I concentrated on learning noun/verb *homophones*, yet one interesting following question is how does that relate to the acquisition of noun/verb *polysemes* that share the same form and have related meanings (e.g., a kiss/to kiss). I showed that children are sensitive to the syntactic context in which a known word is used to decide whether the word may instantiate a new meaning. Yet the existence of polysemes suggests that noun/verb distributional cues do not necessarily trigger the formation of a novel *distinct* lexical entry. Indeed noun/verb polysemes share not only a common phonological form but also the same lexical representational base (e.g., Caramazza & Grober, 1976). Oshima-Takane, Barner, Elsabbagh, & Guerriero (2001) propose that children can learn the cross-categorical use of these words not only by paying attention to the distributional cues of the noun/verb pair but also, by using the semantic information of the words, noting for instance that the same word could be used for an artifact and its function (e.g., brush) (see also Lippeveld & Oshima-Takane, 2014). Thus paying attention to the semantic relation between the meanings of a word may help children to differentiate between noun/verb homophones and noun/verb polysemes.

To conclude, I started by noting that learning homophones may be easier when their meanings are made sufficiently distinct by the context in which they are used. When a *verb* form referring to an *action* was used as a *noun* to label a novel *object*, children had no problem to learn it. Yet the present result cannot tell whether the syntactic distance (verb/noun) or the semantic distance (action/object) mattered here. In Experiment 2, I investigate the effect of semantic distance alone on learning noun-noun homophones.

#### 3.2.2 Experiment 2 & 3 - Manipulating the semantic distance

Learning homophones may be easier when the semantic distance between their meanings is large. For example, "bat" is likely to be unambiguous in a context where one speaks about sport, as we do not expect the bat-animal meaning in this context. Intuitively, homophones seem to map onto clearly distinct meanings (e.g., animal-bat/baseball-bat, flour/flower, mussel/muscle, etc.) suggesting an advantage for homophones that are semantically distinct over semantically close (I come back to this idea in section 3.3).

In Experiment 2, toddlers were taught either a novel animal noun that was homophonous

with a noun referring to an artifact (**an artifact-homophone**, syntactically identical, but semantically distant; e.g., "un pot", *a potty*) or a novel animal noun that was homophonous with a noun referring to an animal (**an animal-homophone**, both syntactically and semantically close; e.g., "un chat", *a cat*). Contrary to Experiment 1, a non-homophone was not taught, so as to minimize task difficulty, since learning a second meaning for a known noun may be more challenging than learning a second meaning for a known verb. Yet because children were taught a single word, I also controlled that a potential learning effect would not be due to a preference for looking at the only labeled animal, namely toddlers might look more towards the labeled animal during the test phase not because they associated the novel label to this animal but because this animal is the only one that has been named during the teaching phase. Therefore, during the test phase toddlers were tested both on the homophone (the artifact-homophone condition or the animal-homophone condition), and on an untaught non-homophone (the non-homophone condition). If toddlers in the non-homophone condition looked more towards the unlabeled animal (or at least behaved differently from the homophone conditions; the *mutual exclusivity effect*), this would suggest that toddlers formed a form-meaning association between the artifact-homophone, or the animal-homophone, and its animal referent.

## Experiment 2

### Method

**Participants.** Thirty-two French 20-month-olds took part in this experiment, sixteen learnt an artifact-homophone (range = 19;2 months to 20;8 months, mean = 19;8, SD = 0;6, 8 boys) and sixteen learnt an animal-homophone (range = 19;4 months to 20;9 months, mean = 20;2, SD = 0;4, 6 boys). Sixteen additional children were replaced because of technical problem ( $n = 7$ )<sup>12</sup>, fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 6$ ), refusal to wear the sticker necessary for eye-tracking ( $n = 2$ ) and not knowing any of the test words according to their parents' report ( $n = 1$ ).

**Apparatus, procedure and design.** Similar to Experiment 1 except that this time toddlers were taught a single novel word that was homophonous with a known noun referring to an artifact (an artifact-homophone) or a known noun referring to an animal (an animal-homophone). The other plush animal was still presented, for a video of the same duration, with the same movements and story line, but without any label (only pronouns were used – "do you see this one? It has big ears..."). During the test phase, toddlers had

<sup>12</sup>For Experiments 2 and 3, the position of the eye-tracker relative to the screen was improperly centered resulting in a loss of data when children were looking towards one side of the screen.

4 test trials on the homophone label (the artifact- or the animal-homophone) and 4 test trials on the non-homophone label (the non-homophone condition).

**Material.** Similar to Experiment 1 except for the set of novel words used in the teaching phase.

*Novel words.* I chose 4 phonological word forms of nouns labeling artifacts that toddlers of that age are likely to know, 4 phonological word forms of nouns labeling animals that toddlers of that age are likely to know (according to the CDI of previous studies) and 4 novel non-homophones.

The 4 artifact-homophones were all monosyllabic words: "verre", "pot", "pull", "bain" (/vɛʁ/, /po/, /pyl/, /bɛ̃/) meaning: *glass, potty, sweater* and *bath*. On average, these words had a phonological neighborhood density of 3.8 (irrespective of the syntactic category of the word) and an average frequency count of 152 in a corpus of child directed speech (the Lyon corpus, Demuth & Tremblay, 2008).

The 4 animal-homophones were also all monosyllabic words: "chat", "loup", "poule", "mouche" (/ʃa/, /lu/, /pul/, /muʃ/) meaning: *cat, wolf, hen* and *fly*. These words had an average phonological neighborhood density of 3.1 (irrespective of the syntactic category of the word) and a frequency count of 157 in a corpus of child directed speech (the Lyon corpus, Demuth & Tremblay, 2008).

The 4 novel non-homophones were identical to those used in Experiment 1.

The average duration of the novel words in the test sentences was 455ms for the artifact-homophones, 517ms for the animal-homophones and 622ms for the non-homophones.

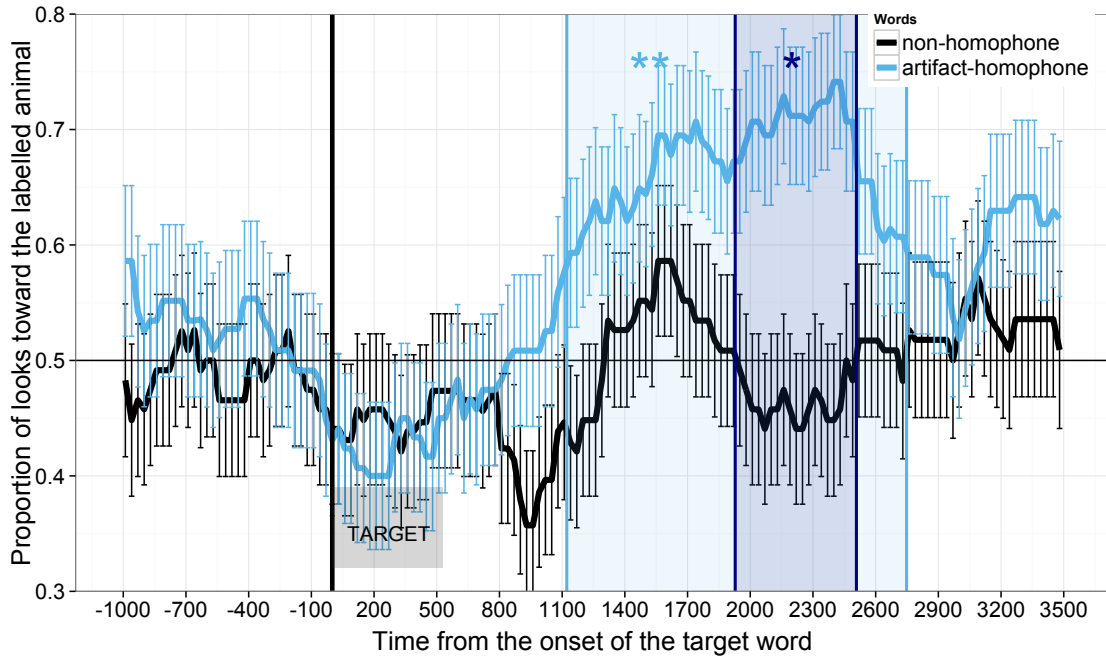
**Measure and Analysis.** Similar to Experiment 1, except that this time the dependent variable was the proportion of looks towards the plush animal labeled during the learning phase (in order to compare any potential difference in behavior between the test trials on the homophone label and the mutual exclusivity trials on the non-homophone label). The cluster analysis was this time run from -1500ms before target word onset until the end of the trial.

## Results

### Artifact-homophone group

Figure 3.2 shows the proportion of looks towards the referent of the artifact-homophone (the labelled animal during the learning phase) from -1000ms before the beginning of the

first target word ("Regarde le [target]! tu le vois le [target]" *Look at the [target]! Do you see the [target]*) until the end of the trial.



**Figure 3.2:** Proportion of looks towards the artifact-homophone referent from the beginning of the first target word (do you see the [target]?) for the artifact-homophone (in blue) and for the non-homophone (in black). Toddlers successfully learnt the artifact-homophone as they significantly increased their look towards the correct picture after target onset, and this behavior was significantly different from the non-homophone condition, in which children tended to switch back to the unnamed plush animal (dark blue shaded area).

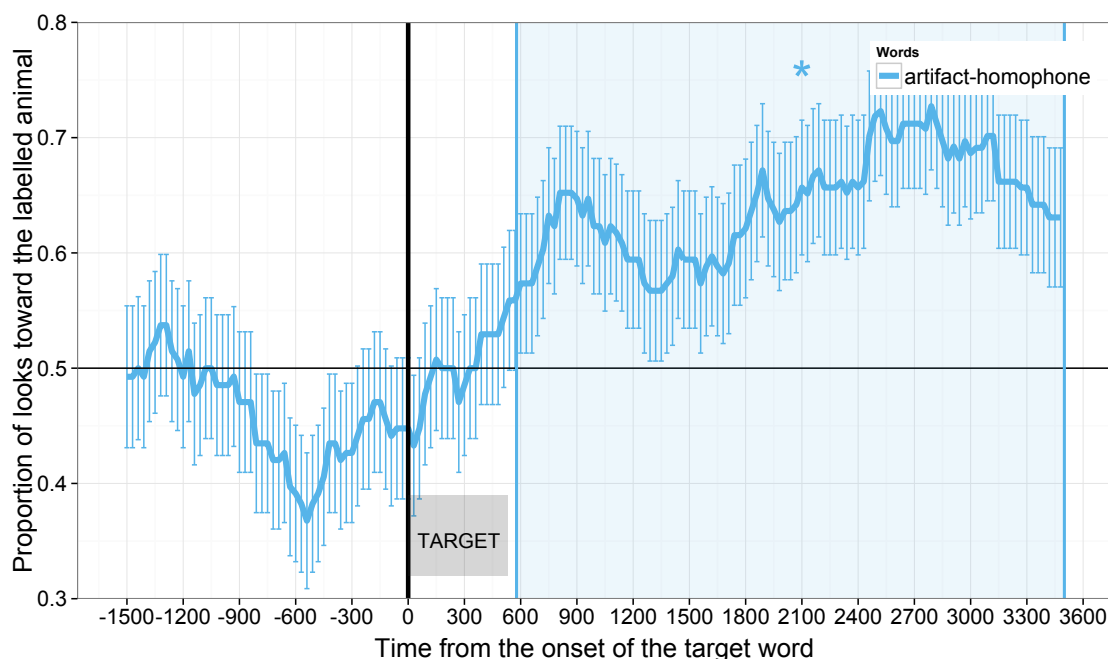
Toddlers looked significantly above chance towards the target in the artifact-homophone condition (blue curve; from 1122ms after target word onset;  $p < .01$ ; light blue shaded area in Figure 3.2) and did not show any preference for any of the objects in the non-homophone condition ( $p > 0.3$ ). Crucially there was a significant difference in performance between the artifact-homophone condition and the non-homophone condition (from 1928 ms to 2508ms after target word onset; dark blue shaded area;  $p < .05$ ). This difference in performance ensures that the increase in looking towards the artifact-homophone was not due to a preference for looking at the only labeled animal.<sup>13</sup>

<sup>13</sup>Another group of toddlers tested on the non-homophone condition only showed a significant mutual exclusivity effect: They looked more at the unnamed animal. One possibility is that we do not observe a mutual exclusivity effect here because the task was more complex, since toddlers were tested on two words (the artifact-homophone and the non-homophone) instead of one (the non-homophone).



### 3 Learning confusable and ambiguous words

This effect was replicated with another group of 16 toddlers that was tested only on the animal-homophone during the test phase.

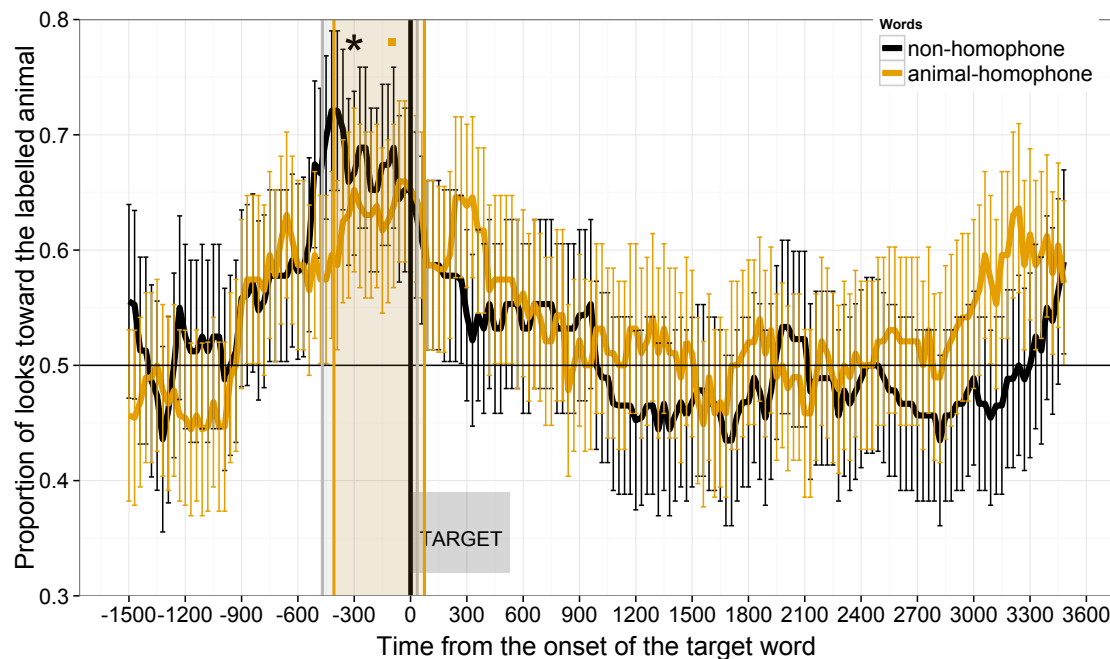


**Figure 3.3:** Replication of the artifact-homophone group except that this time toddlers were tested on the artifact-homophone label only. Toddlers looked significantly above chance towards the target in the artifact-homophone condition (blue curve; from 578ms after target word onset;  $p < .05$ ; blue shaded area)

Toddlers successfully learnt the artifact-homophone. This suggests that they have no problem learning a second meaning for a known word when this additional meaning is semantically distinct from the original meaning of the word. Yet, if semantic distinction is really what matters, toddlers should fail to learn a second meaning for a known word when the additional meaning is semantically *close* to the known meaning of the word (the animal-homophone group).

#### Animal-homophone group

Figure 3.4 shows the proportion of looks towards the labelled referent from the beginning of the first target word ("Regarde le **[target]**! tu le vois le [target]" *Look at the [target]!* *Do you see the [target]*) until the end of the trial.



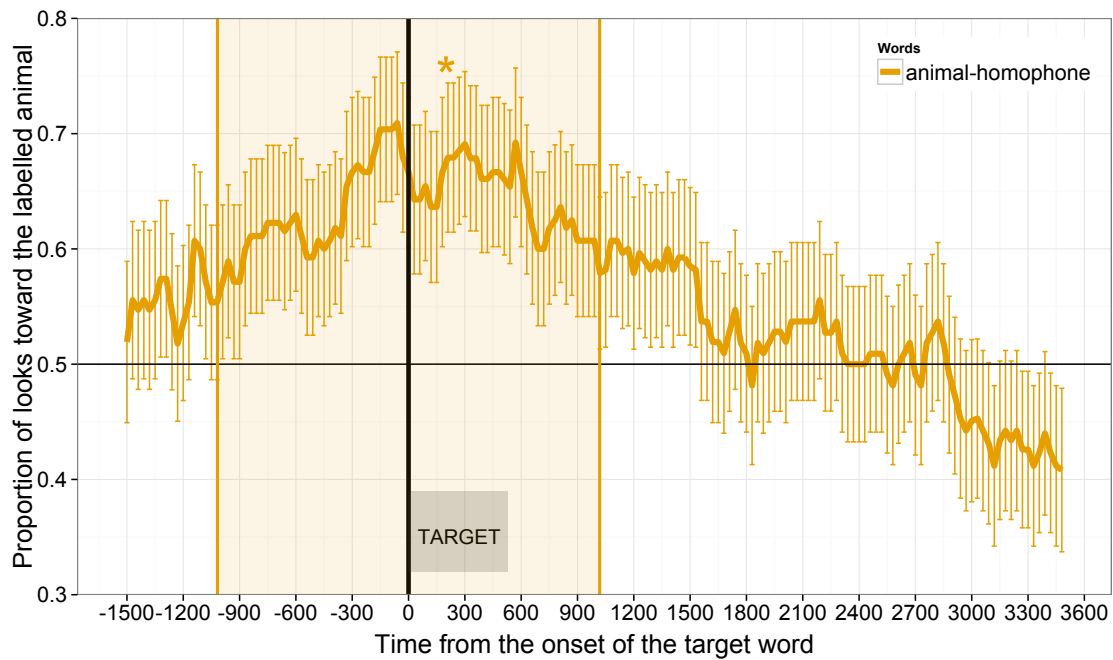
**Figure 3.4:** Proportion of looks towards the animal-homophone referent from the beginning of the first target word (do you see the [target]?) for the animal-homophone (in gold) and for the non-homophone (in black).

Crucially, toddlers behaved differently when they learnt an animal-homophone: In both the animal-homophone condition (gold curve) and the non-homophone condition (black curve) they looked at the animal-homophone referent at a rate above chance **before** the onset of the target word (from about -400ms to 70ms around target word onset for both conditions;  $p_s < .05$ ). There was no difference between conditions. Thus, not only did toddlers in the animal-homophone condition fail to show any recognition of the animal-homophone label, they also failed to apply mutual exclusivity when tested on a non-homophone, suggesting that they did not learn to associate the animal-homophone word to the correct plush animal.

There was a significant difference between the artifact-homophone condition and the animal-homophone condition (from 1924ms to 2584ms after target word onset,  $p < .04$ ). This suggests that toddlers are confused when the two meanings of a pair of homophones are semantically close.

### 3 Learning confusable and ambiguous words

This effect was replicated with another group of 16 toddlers that was tested only on the animal-homophone during the test phase.



**Figure 3.5:** Replication of the animal-homophone group except that this time toddlers were tested on the animal-homophone label only. Similarly they looked at the target picture significantly above chance yet again **before** target word onset (from -1018ms to 1000ms around target word onset;  $p < .05$ ).

In sum, toddlers did not show recognition of the animal-homophone, instead they looked at the target animal before word onset as if they were surprised by the possibility of such a form-meaning association.

### Discussion

Toddlers had no problem learning an artifact-homophone but failed to display any learning of the animal-homophone. This suggests that toddlers treated these labels differently and this critically affected their identification of what counts as a novel lexical entry.

One possibility is that they found it easier to learn homophones that are semantically distinct over homophones that are semantically close. When the speaker used an artifact label to name a novel animal, the difference between the normal usage of the word and this novel situation is so great that it looks unlikely that the speaker could use the label to refer to the original meaning. However, when the original meaning (an animal) is close enough to the novel meaning (another animal), as in the case of the animal-homophone, it

may be more difficult for toddlers to differentiate between a less prototypical member of the original meaning of the label and a novel meaning instance.

Another possibility is that some unmeasured difference between the set of artifact labels and the set of animal labels was responsible for the observed effect. While both sets of words were matched for frequency, neighborhood density in toddlers' lexicon and phonotactic probability, toddlers may have a better lexical representation for the animal labels than the artifact labels used in this study, leading to greater interference (e.g., McKenna & Parry, 1994; Setoh, Wu, Baillargeon, & Gelman, 2013, for some evidence that animals may have a special status). Thus, toddlers may find it more difficult to learn a second meaning for an animal-label than for an artifact-label not because the semantic distance between the two meanings is greater for the artifact-homophones but because toddlers may have greater difficulty in suppressing the primary meaning of the animal-homophones. The next experiment disentangles between these two possibilities.

#### Experiment 3

Experiment 3 was similar to Experiment 2, except that this time toddlers were taught that the animal-homophone labels used in Experiment 2 could label a novel **artifact** (e.g., "un chat", *a cat*, was used to label a novel music instrument). Thus, the set of animal-homophone labels was identical to Experiment 2 but, crucially, the semantic distance between the two meanings of the label increased. If semantic distance between meanings of a pair of homophones is a major reason why learning an artifact-homophone is easier than learning an animal-homophone in Experiment 2, then toddlers should have no problem learning an animal-homophone when the novel meaning is sufficiently distant semantically from the animal category. On the contrary if better lexical representations for the animal-labels used in Experiment 2 led to the observed effect, then toddlers should still fail to learn an animal-homophone even though its second meaning is semantically distinct from its original meaning.

#### Method

**Participants.** Fourteen<sup>14</sup> French 20-month-olds took part in this experiment. 14 additional children were replaced because of technical problem ( $n = 10$ ; see footnote 12), refusal to wear the sticker necessary for eye-tracking ( $n = 2$ ), because the parent took out the earphones during the experiment ( $n = 1$ ) and because the child did not know any of the target words ( $n = 1$ ).

---

<sup>14</sup>This experiment is on-going.

**Apparatus, procedure and design.** Similar to the animal-homophone group in Experiment 2.

**Material.** Similar to Experiment 2 except for the set of novel objects used during the teaching phase.

*Novel objects.* The novel objects were two unfamiliar artifacts. One was a music instrument composed of 8 spinning colored bells and the other was a colored spinning top (see Figure 3.6).



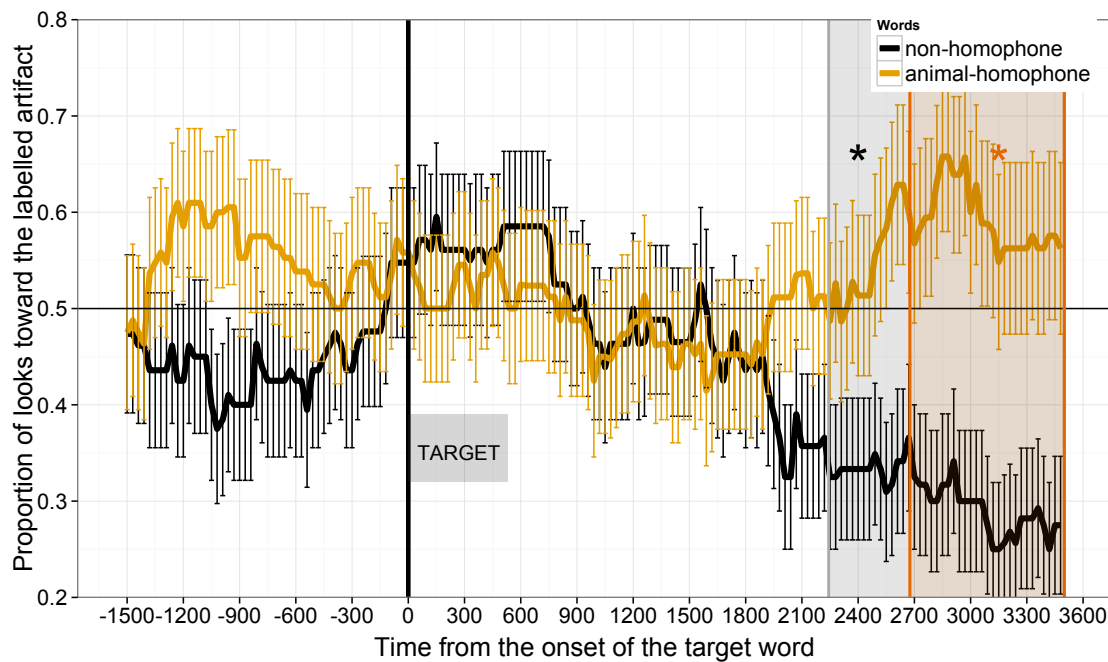
**Figure 3.6:** Novel objects used in Experiment 3.

**Measure and Analysis.** Similar to Experiment 2.

### Results

Figure 3.7 shows the proportion of looks towards the labelled referent from the beginning of the first target word ("Regarde le **[target]**! tu le vois le [target]" *Look at the [target]!* *Do you see the [target]*) until the end of the trial.

Importantly the animal-homophone label did not trigger a "surprisal effect" as in Experiment 2: When the animal-homophone is used to label an artifact, toddlers increased their looks towards the correct artifact referent (though this preference was not significant;  $p = .3$ ). This time, they also correctly performed mutually exclusivity as they looked significantly below chance to the animal-homophone referent (grey shaded area, from 2242 ms until the end of the trial;  $p < .05$ ).



**Figure 3.7:** Proportion of looks towards the animal-homophone referent (an artifact) from the beginning of the first target word (do you see the [target]?) for the animal-homophone (in gold) and for the non-homophone (in black).

Crucially there was a significant difference in performance between the animal-homophone condition and the non-homophone condition (from 2676 ms to 3500 after target word onset, dark orange shaded area,  $p < .05$ ). This indicates that toddlers treated these two words differently, very much like the artifact-homophone group in Experiment 2 (where artifact-homophones labelled novel animals; there was no significant difference between Experiment 3 and the artifact-homophone group in Experiment 2 for both the homophone and the non-homophone conditions  $p > .1$ ) and importantly, differently from the animal-homophone group (where animal-homophones labelled animals; although there were no significant difference between Experiment 3 and the animal-homophone group in Experiment 2 for the homophone and the non-homophone conditions,  $p > .3$ ).

## Discussion

Toddlers were able to learn an artifact-homophone when it labeled a novel animal but failed to learn an animal-homophone labeling the same animal (Experiment 2). However when the animal-homophone was used to label a novel artifact, toddlers seemed to be able to learn it (Experiment 3). Thus, the results suggest that toddlers have no problem learning a second meaning for a known word if this second meaning is semantically distinct from the first known meaning.

However, while toddlers' performance in Experiment 3 did not differ from the artifact-homophone group in Experiment 2, it seems that toddlers were not as successful to learn an animal-homophone as an artifact-homophone even when the semantic distance between the novel and the primary meaning is kept the same. There may be two possibilities to account for this observation: First, there may be a small effect of animal labels vs. artifact labels: It could be the case that lexical representations for animals are more entrenched than the lexical representations of the artifacts chosen in this set of experiments and thus interfere more with the creation of a novel lexical entry; Second, the stories of Experiment 3 and Experiment 2 are different and involve different sets of objects. Thus it may be the case that the stories used in Experiment 3 are more complicated for toddlers than the stories of Experiment 2. In particular because they involve more complicated vocabulary (e.g., for artifacts: "button", "playing music", "bell", "spinning" vs. for animals: "ear", "nose", "legs", "jumping") that may be less known to toddlers of that age. Future work will control for that possibility by teaching toddlers non-homophones for these novel artifacts to check whether these stories generally make word learning more challenging.

In sum, in Experiment 1 showed that toddlers had no problem to learn a second meaning for a known word when this second meaning was syntactically and semantically distinct from the known meaning. Experiment 2 showed that a semantic distinction between the meanings of a pair of homophones was sufficient to learn them. One question that remains is whether a syntactic distinction between the members of a homophone pair may be sufficient to learn two meanings for the same word form. I investigate this question in Experiment 4.

#### 3.2.3 Experiment 4 - Manipulating the syntactic distance using gender

Experiment 4 investigated whether solely increasing the syntactic distance between the meanings of a pair of homophones may facilitate the acquisition of these meanings. To isolate the effect of syntax from semantic, I focused on grammatical gender. Crucially, grammatical gender is a lexical property, as opposed to a semantic property in languages where gender categories are not clearly defined in semantic terms (as it is the case in French). In gender-marking languages, the gender of the noun determines the form of associated determiners and adjectives. In French, feminine nouns are preceded by a gender-marked definite article "la" or indefinite "une" and masculine nouns by the gender-marked definite article "le" or the indefinite "un" when used in their singular form. Such gender cues have been shown to constrain lexical access in adults (e.g., Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Spinelli & Alario, 2002) but also in young children (E. K. Johnson, 2005; Lew-Williams & Fernald, 2007; Van Heugten & Shi, 2009). Interestingly for the current study, adults use such gender-marked context to selectively access the meaning of homophones (Spinelli & Alario, 2002). For instance, in French, /sel/ means both *saddle*

(feminine) and *salt* (masculine) and each meaning is accessed independently when preceded by a gender-marked article. Gender could thus be used to distinguish between different meanings of the same phonological form, by preventing the activation of lexical candidates that do not belong to the same gender category.

I explored whether a context marked for grammatical gender can help toddlers to identify a second meaning for a known word when the original and the second meanings are associated with different genders. In Experiment 4, toddlers were taught that a novel animal label was homophonous with an animal noun they already knew (as in Experiment 2) but this time in a different gender-marked context (**a gender-homophone**, semantically identical to the first meaning but syntactically different; e.g., "une chat", *a cat<sub>feminine</sub>*, normally masculine in French). If a gender-marking context is sufficient to identify an additional meaning for a known word, then toddlers should recover from their failure to learn an animal-homophone (Experiment 2) and correctly learn it when presented in a different gender-context.

### Method

**Participants.** Sixteen French 20-month-olds took part in this experiment (range = 19;2 months to 21 months, mean = 20;2, SD = 0;5, 7 boys). Two additional children were replaced because of fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data.

**Apparatus, procedure and design.** Similar to Experiment 2 except that this time there was no non-homophone condition (as toddlers' behavior was consistent between the same experiment with and without this condition, compare Figure 3.2 and 3.3 as well as Figure 3.4 and 3.5).

**Material.** Similar to Experiment 2 expect for the gender of the labels.

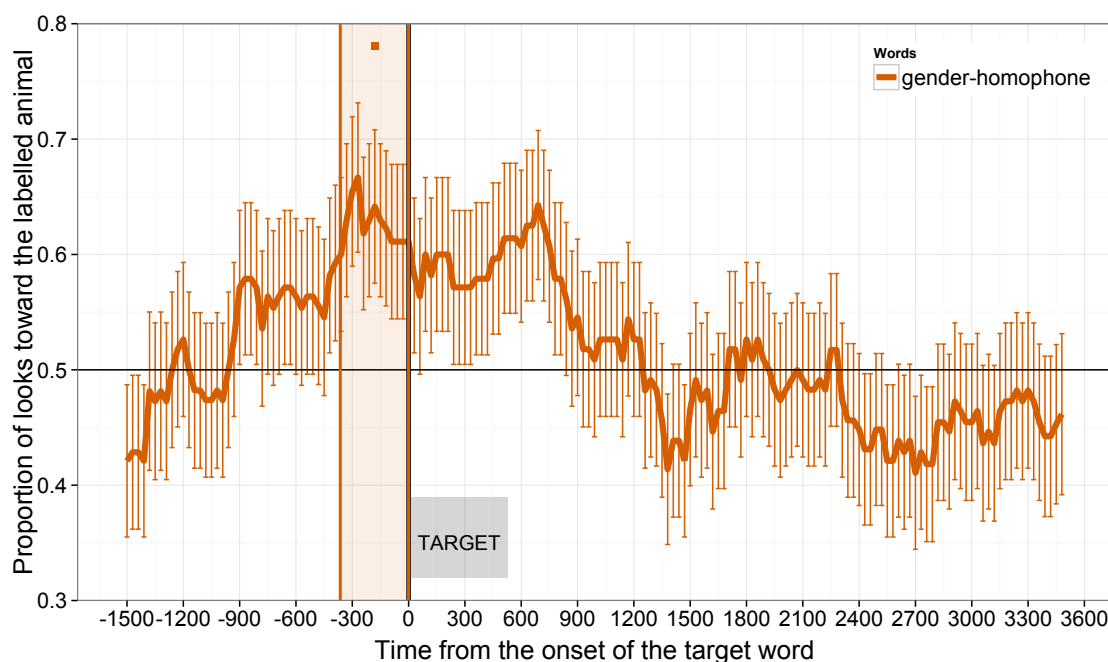
*Novel words.* The 4 animal-homophones were taught with a different gender (therefore gender-homophones): *une chat, une loup, un poule, un mouche* (cat, wolf, hen and fly) instead of *un chat, un loup, une poule, une mouche*. The average duration of the novel words in the test sentences was 530ms for the gender-homophones.

**Measure and Analysis.** Similar to Experiment 1.



## Results

Figure 3.8 shows the proportion of looks towards the target around the beginning of the first target word ("Regarde le [target]! tu le vois le [target]" *Look at the [target]! Do you see the [target]*) until the end of the trial.



**Figure 3.8:** Proportion of looks towards the target picture from the beginning of the first target word (do you see the [target]?) for the gender-homophone (in orange). Toddlers failed to show any recognition of the gender-homophone.

Toddlers taught a gender-homophone patterned the same way than toddlers taught an animal-homophone (Figure 3.4 and 3.5): They looked at the target at a rate above chance but again, before target word onset (from -364ms until 0ms) though this effect was only marginally significant ( $p = 0.09$ ). This suggests that toddlers fail to learn the gender-homophones in the same way they failed to learn the animal-homophones (Experiment 2).

## Discussion

Experiment 4 isolated the effect of syntax alone in the identification of a novel meaning by using a gender-marked context. The results show that toddlers failed to take this information into account. Certainly it does not mean that syntax alone is insufficient to

distinguish between meanings of a pair of homophones, but that gender may not provide enough evidence for toddlers that a novel meaning is intended in that situation.

Yet there may be several other interpretations for toddlers' failure. One possibility is that French 20-month-olds do not yet use gender when processing spoken language. However, I believe this possibility to be unlikely: A recent study shows that 18-month-old toddlers can track the statistical dependencies between gender-marked articles and nouns (Van Heugten & Christophe, in press). Toddlers prefer to listen to article-noun sequences in which the gender-marked article matched with the gender of the noun (e.g., "la<sub>fem</sub> poussette<sub>fem</sub>", *the stroller*) than when the gender-marked article mismatched with the gender of the noun (e.g., "le<sub>masc</sub> poussette<sub>fem</sub>"). This suggests that toddlers of that age may already be sensitive to gender cues when processing speech.

Another possibility is that gender information may not bring convincing evidence that a novel meaning is intended when used to label animals. Indeed, gender marking for an animal-label corresponds to the male and female individuals of the specie (e.g., "un<sub>masc</sub> chat" /ʃa/ refers to the male cat and "une<sub>fem</sub> chatte" /ʃat/ to the female cat). Accordingly if toddlers already know that animals may have different biological genders and understand that female individuals are often preceded by a feminine article (e.g., "la") and male individuals by a masculine article (e.g., "le"), they may have considered that the original meaning was intended when presented with the novel animal, as in Experiment 2, despite being preceded by contradictory gender information. Thus, if gender-marking helps to distinguish an additional meaning for a known word form, such information may be available only when labeling non-biological entities. In order to test this hypothesis, the same experiment could be conducted on artifact-labels instead of animal-labels.<sup>15</sup>

At any rate, while toddlers of that age use gender cues when processing speech, such cues may not constitute systematic evidence to identify that a word form could map onto several meanings (as I discussed in the case of animal labels) and accordingly toddlers may fail to use it as I show here. If this is correct, this suggests that learning should not exert a pressure for across-genders homophones to be more represented than chance in the lexicon, I come back to this hypothesis in section 3.3.

### 3.2.4 Experiment 5 - Manipulating neighborhood density

The previous experiments manipulated the distance (semantic and syntactic; semantic alone; syntactic alone) between a known first meaning and a novel second meaning. Experiment 5 investigated whether the phonological context of the word in the lexicon (whether

---

<sup>15</sup>Though, one would need to find a condition where children fail on artifact-labels first.

it belongs to a dense vs. sparse phonological neighborhood) modulates the conditions in which toddlers accept a secondary meaning for a known word.

Phonological neighborhood density for a word is the number of words that differ by one addition, one deletion or one substitution (Luce, 1986). For instance, neighbors of "cat" include words such as "cap", "hat", "fat", "rat", "at", "catch", etc. Some words are said to live in a *dense* neighborhood when they have many neighbors and some words live in a *sparse* neighborhood because they have only a few neighbors. Importantly, neighborhood density has been shown to play a critical role in speech processing: Adults recognize words faster when they live in sparse neighborhoods than when they live in dense neighborhoods (Luce & Pisoni, 1998a; Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Such an inhibitory effect may also help toddlers to learn homophones as I explain below.

In previous experiments, toddlers were taught a novel meaning for a known word form with a high phonological neighborhood density in toddlers' lexicon. Indeed, on average the animal-homophones had a phonological neighborhood density of 3.1 and the artifact-homophone of 3.8, which is rather high considering that an average French 20-month-old toddler comprehends about 200 words (according to measures using the French CDI in previous experiments, Kern, 2007). One possibility is that learning an artifact-homophone was possible in Experiment 2 because of the semantic distance between the original meaning (an artifact) and the novel meaning (an animal) of the word. Yet, another non-mutually exclusive possibility for their success, is the use of word forms with high phonological neighborhood density in this experiment (e.g., "bain" *bath*) which may have activated others words in toddlers' lexicon (e.g., "pain" *bread*, "main" *hand*), thus inhibiting strongly the known meaning and favoring the possibility of a novel meaning in these conditions.

Experiment 5 explored whether neighborhood density modulates the learning effect observed in Experiment 2. Toddlers were taught that a novel animal label was homophonous with an artifact noun they already know (as in Experiment 2) but this time the label was from a sparse phonological neighborhood in toddlers' lexicon (**a sparse-homophone**: Semantically distinct from the original meaning but whose word form has a low phonological neighborhood density; e.g., "un livre", *a book*). If phonological neighborhood density helped toddlers to learn the artifact-homophones in Experiment 2, then toddlers should fail to learn the sparse-homophones. On the reverse, if toddlers are able to learn the sparse-homophones as well as the artifact-homophones in Experiment 2, this would suggest that increasing the semantic distance between the meanings of a pair of homophones is enough to learn these meanings, independently of the phonological neighborhood density of the word form.

## Method

**Participants.** Sixteen French 20-month-olds took part in this experiment (range = 19;1 months to 21 months, mean = 20;1, SD = 0;6, 9 boys). Six additional children were replaced because of fussiness during the experiment resulting in more than 50% of trials with missing eye tracking data ( $n = 3$ ), refusal to wear the sticker necessary for eye-tracking ( $n = 1$ ), no increase in average proportion of looks towards the target during familiar-word trials ( $n = 1$ ) and technical problem ( $n = 1$ ).

**Apparatus, procedure and design.** Similar to Experiment 4.

**Material.** Similar to Experiment 4 except for the set of sparse-homophones used in the teaching phase.

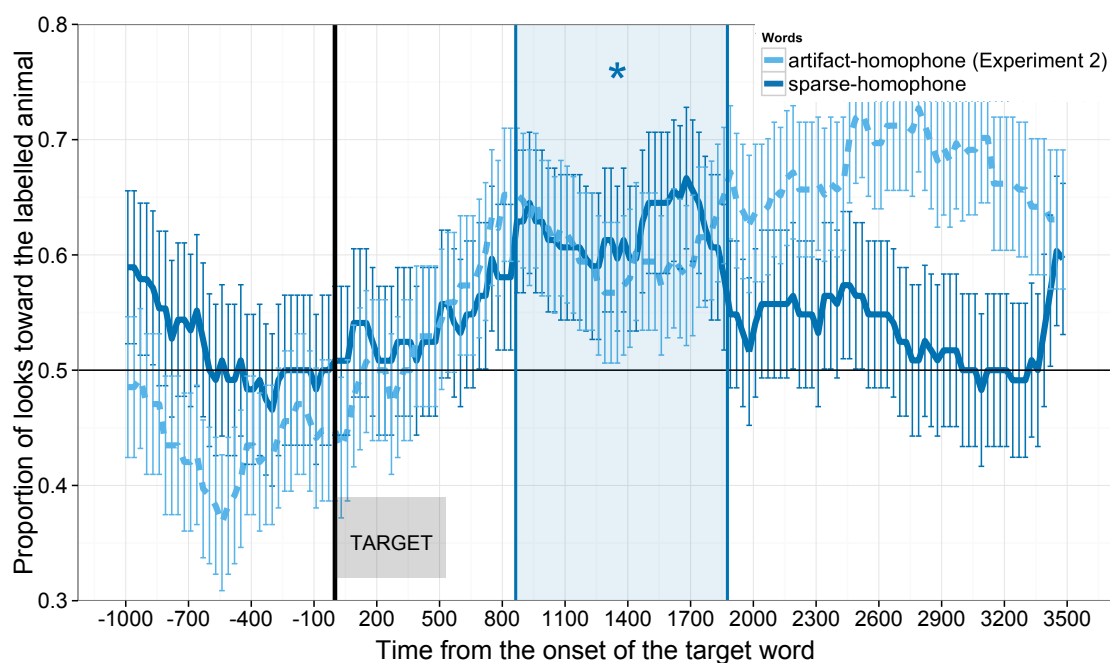
*Novel words.* I chose 4 phonological word forms of nouns labeling artifacts that toddlers of that age are likely to know (according to the CDI of previous studies).

The 4 sparse-homophones were also all monosyllabic words: *livre*, *fleur*, *fraise*, *sieste* (/livʁ/, /flœʁ/, /fʁɛz/, /siɛst/) meaning: book, flower, strawberry and nap. These words had an average phonological neighborhood density of 0 in children's lexicon (irrespective of the syntactic category of the word). The average duration of the sparse-homophones in the test sentences was 726ms.

**Measure and Analysis.** Similar to Experiment 1.

## Results

Figure 3.9 shows the proportion of looks towards the target around the beginning of the first target word ("Regarde le [**target**]! tu le vois le [target]" *Look at the [**target**]*! *Do you see the [target]*) until the end of the trial.



**Figure 3.9:** Proportion of looks towards the target picture around the beginning of the first target word (do you see the [target]?) for the sparse-homophone (in dark blue). The replication of Experiment 2 (Figure 3.3) has been reproduced here in dashed line (light blue) for convenience).

Toddlers taught a sparse-homophone behaved like toddlers taught an artifact-homophone: They looked to the correct referent at a rate above chance (from 864ms until 1876ms;  $p < .05$ ) suggesting that they learnt the sparse-homophone.

There was no statistical difference in learning a sparse-homophone and a (dense) artifact-homophone in the same experimental design: I compared the replication of Experiment 2, in which children are tested only on the animal-homophone, with the sparse-homophone condition ( $p = 0.15$ ).

## Discussion

Toddlers successfully learnt a sparse-homophone, and behaved in the same way as if they learnt a (dense) artifact-homophone. This suggests that phonological neighborhood density does not exert a major influence on homophone learning.

However, despite the lack of statistical significance, there seems to be a learning advantage for dense artifact-homophones compared to sparse artifact-homophones (Figure 3.9) suggesting that an influence of neighborhood density on learning may be visible with more statistical power. Yet such an advantage might be also compatible with another explanation:

Dense artifact-homophones are composed of more frequent sound sequences than sparse artifact-homophones (cumulative bigram log-probability  $\log P = -3.38$  for dense artifact-homophones;  $\log P = -5.69$  for sparse artifact-homophones<sup>16</sup>) and there is evidence that toddlers and children find it easier to learn words composed of frequent segments than words composed of infrequent segments (Graf Estes & Bowen, 2013; Storkel & Maekawa, 2005).

At any rate, the present results suggest that the effect of phonological neighborhood density is smaller than the effect of semantic distance.

### 3.2.5 Conclusions

An important part of the word learning process requires children to identify what counts as a novel word and what does not. This is especially a challenge when learning homophones, where the same phonological form is used to refer to several distinct meanings. Here, I investigated different sources of information that may help children to identify when a novel meaning for a known word is appropriate. Specifically I manipulated 1) both the syntactic and the semantic distance between the novel word and its familiar homophone across experiments (Experiment 1-4) and 2) the position of the word form in the phonological network of the mental lexicon (Experiment 5).

Experiment 1 showed that toddlers had no problem learning homophones when their meanings are realized in different syntactic categories: "an eat" was a good label for a novel animal despite children knowing the meaning of the verb "to eat". Yet, this does not tell us whether the syntactic distinction (noun/verb) or the semantic distinction (object/action) between the two meanings was important to learn homophones. Experiment 2 and 3 showed that a semantic distinction between the two meanings of a pair of homophones was sufficient to trigger learning: Toddlers learnt easily that "a potty" could also label a novel animal but failed in a condition where the novel animal was labelled "a cat". However, Experiment 4 failed to show that the syntactic context alone, when different meanings of a homophone are cued by different genders, is sufficient to learn a second meaning for a known noun: Toddlers failed to learn that a novel animal could be called "une<sub>em</sub> chat" *a cat* despite the existence of the article "une<sub>em</sub>" that indicates that "chat" (a masculine noun in French) may not be used in its known meaning. Finally, Experiment 5 suggests that the phonological density of the word form of the homophone does not have a major influence on establishing that this word form maps onto several meanings.

The word learning process, thus, does not seem to exclusively rely on identifying those word forms that have no lexical entry and associate them a meaning (e.g., Carey, 1978b). As I

<sup>16</sup>This was calculated using a ngram model on the set of word types in the French lexicon, taken from the Lexique database; (New, Pallier, Brysbaert, & Ferrand, 2004)

showed were, deciding what counts as novel word or not depends on the context in which the word form appears, even when this word form already has a lexical entry in children's lexicon. In particular, the linguistic context plays a prominent role in constraining lexical access to the existing word and thus impacting the likelihood that this word form could convey a novel meaning in that context. Similarly, when the two meanings of the same word form are semantically distinct, it makes it easier to identify that a novel meaning is intended.

Taken as a whole, this set of studies suggests that learning homophones is less challenging for toddlers than one might think: When the two meanings of a pair of homophones appear in contexts that are sufficiently distinct, they are learnt as easily as non-homophones. I conclude that creating a novel lexical entry depends on multiple sources of information coming from the lexicon and the parsing system.

### 3.3 Similar-sounding words and homophones in the lexicon

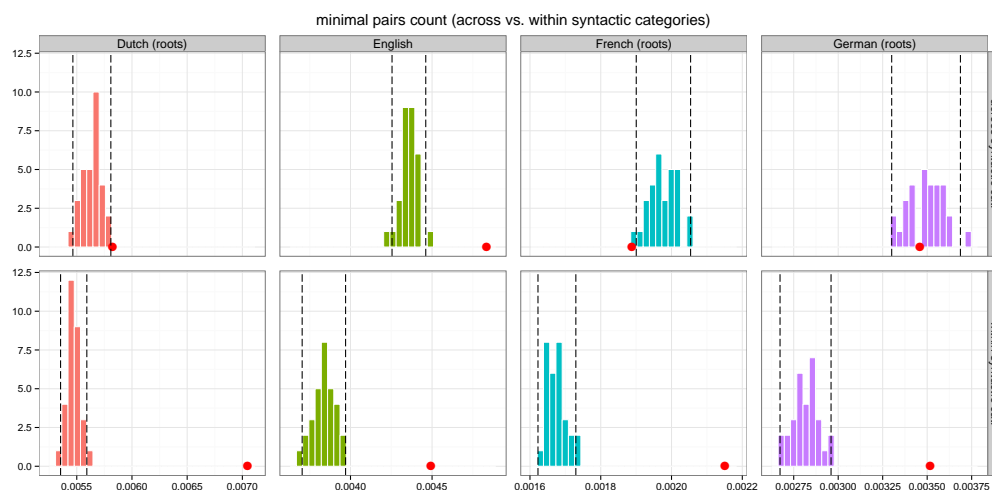
The results obtained in both section 3.1 and 3.2 suggest that the process of creating a novel lexical entry is mediated by multiple sources of information coming from the lexicon (e.g., lexical-semantic relationships) and the parsing system (e.g., expecting a noun or a verb in a given linguistic context). As a result, minimal pairs of words and homophone pairs are less challenging for children than previous studies suggested, at least whenever each member of the pair appears in distinct syntactic or semantic contexts. One important question is whether the structure of the lexicon reflects these constraints on learning: Is it the case that members of a homophone pair, or a minimal pair, are more distant from one another than would be expected by chance alone? Such a result would suggest that the lexicon might be shaped by learning constraints.

#### 3.3.1 Minimal pairs in the lexicon

Learning neighbors of familiar words is difficult for toddlers, but as I showed, this difficulty disappears when the novel words appear in contexts that are sufficiently different from their known neighbors (either syntactically or semantically or both) (section 3.1). If learnability influences language changes, then this constraint on early lexical acquisition might have a long-lasting impact on the overall structure of the lexicon. Do lexicons avoid similar-sounding words? And when similar-sounding words do occur, are they preferentially distributed across syntactic or semantic categories to improve their learnability (and their recoverability)?

The quantitative analyses of the lexicon done in section 2.1 suggest that natural lexicons contain many similar sounding words, more than what would be expected by phonotactics and morphological regularity. In addition, lexicons display a small but significant tendency for similar sounding words to be also more semantically similar (section 2.2, Monaghan, Shillcock, Christiansen, & Kirby, 2014). Relevant to the results with toddlers, in section 2.1, I also looked at whether lexicons favor similar-sounding words (i.e., minimal pairs) to appear *across* syntactic categories rather than *within* the same syntactic category (see Figure 3.10 copied below for convenience). These results show that all 4 languages (Dutch, English, French and German) have more minimal pairs *within* categories than would be expected by chance. In sum, lexicons show an advantage for similar sounding words that are closely related.





**Figure 3.10:** These histograms show the distribution of the probability of getting a minimal pair within and across syntactic categories compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of simulated lexicons. All 4 languages are significantly more likely to have minimal pairs within categories (bottom row) than would be expected by chance.

At first sight, this might appear at odds with children's difficulty for learning similar sounding words. Yet, learning is certainly not the sole pressure that could influence the phonological structure of the lexicon. As discussed in chapter 2, word form similarity may be preferred for other cognitive reasons. Indeed, similar-sounding words have been shown to be easier to remember and produce for adults (e.g., Vitevitch, 2002; Vitevitch, Chan, & Roodenrys, 2012; Vitevitch & Stamer, 2006) and preschoolers (e.g., Storkel & Lee, 2011; Storkel & Morrisette, 2002). Similarly, the result that words of the same syntactic category share more phonological properties than with words of different classes (see also Kelly, 1992) makes it easier to group words into categories and may help the acquisition of syntactic or semantic categories (Monaghan, Christiansen, & Fitneva, 2011).

Why does word learning not actively exert a selective pressure for words that are phonologically dissimilar over the course of language evolution? There may be two possible explanations for this. First, children may eventually learn phonological neighbors through repeated exposure and more varied learning contexts. Thus banning these words from appearing in the lexicon may not be required since it is not a major impediment to learning. Second, the present results suggest that the creation of a novel lexical entry depends on children's ability to use the context in which the novel word appears to constrain its possible meanings. Over the course of development, children develop their lexicon as well as their parsing abilities. As a result they become more sensitive to finer-grained contexts that could help them to identify more easily when a novel meaning is intended, and this even in cases when its associated word form is similar to a word form they already know. For instance, in a sentence such as "Do you want to drive the tog?", children will likely

interpret "tog" as a type of vehicle since they could use its immediate context (i.e., "drive") to attribute properties ("could be driven") to the novel word that makes it unlikely to be confused with "dog".

Importantly, the observation that difficulties in learning phonological neighbors do not scale up to the overall structure of the lexicon does not mean that learning does not exert an influence at all. Yet its influence may be less critical than the other cognitive pressures mentioned above for this particular phenomenon. While learning difficulties may not be visible in the *static* structure of the lexicon, they may be reflected in the *dynamics* of early lexical growth: Novel words may preferably be added along dimensions that allow them to be easily distinguishable from already existing words. Previous work looking at the dynamic growth of the lexicon only focused on how phonological similarity or semantic similarity of single words to other words in the rest of the lexicon may *separately* influence word learning (Carlson, Sonderegger, & Bane, 2014; Hills, Maouene, Maouene, Sheya, & Smith, 2009; Steyvers & Tenenbaum, 2005, but see Regier 2005). For instance, Hills et al. (2009) show that the order of acquisition of nouns depends on the semantic or the phonological connectivity to other words in the lexicon: The more connected the word, the earlier it is acquired. Yet as we have seen, looking at a single dimension independently of the others may not account for the learning pattern observed here. Thus, it remains open to question whether the growing lexicon reflects the influence of the learning system.

#### 3.3.2 Homophones in the lexicon

Learning homophones may be difficult even for preschoolers (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997), yet as I showed this difficulty is reduced when both meanings of a pair of homophones have different syntactic categories or cover distinct concepts. One interesting question is thus whether these learnability advantages translate into the overall structure of the lexicon: Are there more homophones from different syntactic categories than same-category homophones in lexicons? Similarly, are members of a homophone pair more likely to be semantically distant in languages? Interestingly the present results also suggest that grammatical gender and neighborhood density do not help toddlers to identify whether a given word form maps onto several meanings. Following the same idea, this suggests that these factors might not exert a major influence on the organization of homophony within the lexicon.

To investigate these questions, I extracted the pairs of homophones in the lexicon of 4 languages (Dutch, English, French and German) and looked at their distributions in both the syntactic (grammatical categories and gender) and semantic spaces. I then compared these

distributions to random baselines that simulate how homophone pairs should be distributed under random conditions if there were no cognitive pressure (including learning).<sup>17</sup>

#### Method

**Lexicons.** I used the phonemic lexicons of 4 languages: Dutch, English, German (extracted from CELEX, Baayen, Piepenbrock, & van H, 1993) and French (extracted from Lexique, New et al., 2004). For each language, I defined a lexicon as the set of the most 10,000 frequent word forms.

To identify homophone pairs in each lexicon, I took all the pairs of words that share the same phonological form but are from different lemmas according to their lemma code in CELEX and Lexique. This procedure eliminated homophones coming from the same root but instantiated by different categories (e.g., to fight/a fight) or where one of the forms has a silent morphological marker (e.g., chien/chiens *dog/dogs*, which are pronounced in the same way in French).

#### Measures.

*Syntactic category.* I used the Part Of Speech (POS) tags in CELEX for Dutch, English and German and in Lexique for French to count the number of homophones within the same syntactic category (e.g., animal-"bat"/baseball-"bat") and the number of homophones across different categories (e.g., a park/to park).

*Gender.* I used the gender information tags provided in CELEX for Dutch and German and in Lexique for French to count the number of noun-noun homophones within the same gender (e.g., "avocat" meaning *avocado<sub>masc</sub>* or *lawyer<sub>masc</sub>*) and across different genders (e.g., mur/mûre, *wall<sub>masc</sub>/blackberry<sub>fem</sub>*). Note that there are 3 grammatical genders in Dutch and German (feminine, masculine, neutral) and 2 in French (feminine, masculine). English was not concerned by this measure as it is not a gender-marked language.

*Semantic Similarity.* I applied Latent Semantic Analysis (LSA, Landauer & Dumais 1997) on Wikipedia for each language using the **Gensim** package (Rehurek, Sojka, & others, 2010) (see section 2.2). Thus, each word of the Wikipedia corpus was modeled as a vector in a multidimensional space. The semantic similarity between two words is the cosine of the angle between the two word vectors. A value close to 1 indicates that two words are close in meaning, whereas values close to 0 indicate that the meanings are not related.

---

<sup>17</sup>Note that I did not study (yet) the influence of neighborhood density on homophones. However one previous study suggests that shorter words and more phonotactically probable words (which also have more phonological neighbors, see Mahowald, Dautriche, Gibson, & Piantadosi *submitted*) have more meanings than longer and less phonetically likely words (Piantadosi, Tily, & Gibson, 2012). Yet, a future study should examine whether there is an influence of phonological neighborhood density on the number of meanings beyond confounding factors such as frequency, phonotactic probability and length.

I computed the semantic similarity between the two members of homophone pairs that shared their syntactic category (excluding across-categories homophones). Yet, because LSA is computed over an orthographic corpus, only homophones that are written differently can be distinguished with this measure, therefore I looked only at these homophones that have different orthography (between 200 and 400 pairs for the 4 languages under study).

#### Random baselines.

*Syntactic category.* For each language, I shuffled the syntactic categories within each word length. I then evaluated the number of pairs of homophones that fell across categories for this random configuration.

*Gender.* For each language, I extracted the subset of nouns from the lexicon and shuffled their grammatical gender within each word length. I then evaluated the number of pairs of homophones that have a different gender for this random configuration.

*Semantic similarity.* Similarly, I randomly shuffled the LSA vectors within all words of the same length and computed the average cosine similarity of this configuration of form-meaning mappings for all pairs of homophones.

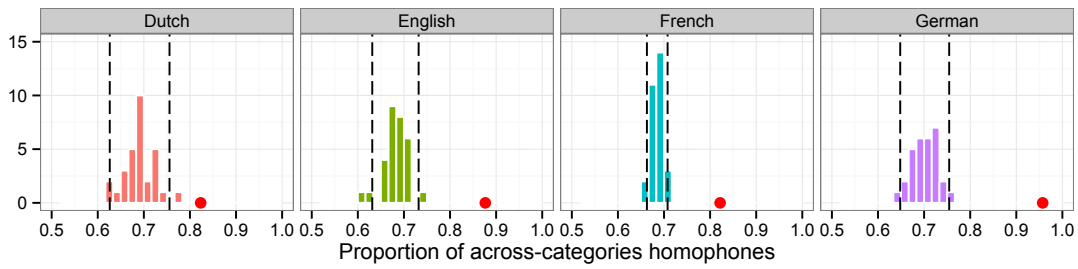
Each random baseline was repeated 30 times in order to obtain a chance distribution for the given measure.

## Results

### Across-categories homophones in the lexicon

I first considered the proportion of homophone pairs that are distributed across syntactic categories. If there are more homophones across syntactic categories than expected by chance, the proportion of across-categories homophones should be greater in the lexicons than in the random baselines.

Figure 3.11 shows how the random baselines (the histograms) compare to the lexicons (the red dots). Crucially, all histograms fall to the left of the red dot, which means that all lexicons have more across-categories homophones than expected by chance (all  $ps < .001$ ). Note that the proportion of across-categories homophones (ranging from 0.8 to 0.9 across the 4 languages) is greater than the proportion of same-category homophones (the complementary proportion). This suggests that there is a pressure for homophones to be distributed *across* syntactic categories rather than *within*.

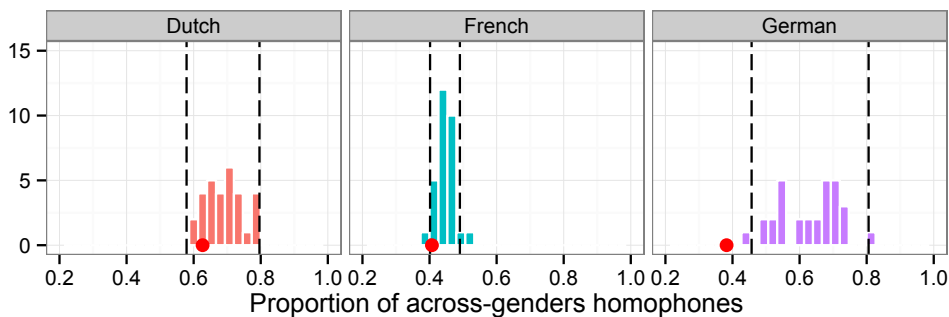


**Figure 3.11:** These histograms show the distribution of the proportion of across-categories homophones compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of random baselines. All 4 languages have significantly more across-categories homophones than expected by chance.

### Across-genders homophones in the lexicon

I then considered the proportion of noun-noun homophone pairs that are distributed across different genders. Because English is not a gender-marked language, this analysis focuses on Dutch, German and French.

As shown in Figure 3.12, the proportion of across-genders homophones is lower than chance (German) or undistinguishable from chance (Dutch and French), suggesting that there is no pressure for distributing noun-noun homophones across grammatical genders unless there are phonological correlates of gender-marking that make it difficult to get different-gender homophone pairs – but would make it easier to learn and remember gender itself.

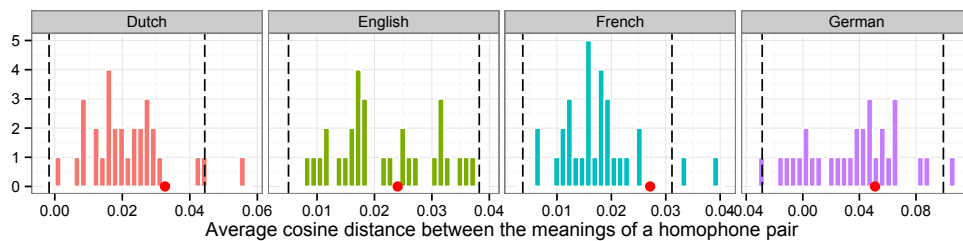


**Figure 3.12:** These histograms show the distribution of the proportion of across-genders homophones compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of random baselines. Lexicons do not have more different-gender homophones than expected by chance.

### Semantically unrelated homophones in the lexicon

Finally, I looked at the average semantic similarity of all pairs of homophones from the same grammatical category and whose members have a different orthography (see the Method section).

As seen in Figure 3.13, the average semantic similarity between meanings of a homophone pair does not differ from chance across all 4 languages. I discuss one possible explanation for this result in the Discussion.



**Figure 3.13:** These histograms show the distribution of the average semantic distance between members of a pair of homophones compared to the real lexicon (the red dot). The dotted lines represent 95% confidence intervals derived from the distribution of random baselines. The members of the pairs of homophones in these 4 lexicons are not more semantically distant than expected by chance.

### Discussion

The distribution of homophones in the lexicon is mainly in line with the results with toddlers: 1) there are more across-categories homophones in the lexicon than same-category homophones and it is also the case that toddlers learn homophones easily when they span different syntactic categories; 2) the distribution of homophone pairs across grammatical genders is not different from chance and toddlers appear to not be influenced by gender when learning homophones. However, 3) the average semantic similarity between the two members of a pair of homophones is as low as the chance level of semantic relatedness, yet toddlers find it easier to learn homophones when their meanings are sufficiently distinct.

Point 3) above seems to contrast with the results with toddlers. Yet, the result of the lexical analysis should not be surprising. Recall that the chance level was constructed by randomly assigning a meaning (i.e., a vector) to each word involved in a homophone pair. Yet, how likely is it for two randomly chosen meanings in the lexicon to be similar? The intuition is that it is pretty low. As a result, the chance level already measures the absence of similarity between two meanings. Thus if homophones are indeed semantically distinct, then it is no surprise that their average semantic distance does not differ from such chance level.

It is also important to consider the methodological limitations of these results. All these analyses depend on the coding scheme used for lemmas: Which word form counts as a lemma and which word forms are derived from it. This was essential to spot homophones which, by definition, belong to different lemmas. Yet it certainly misses homophones on the way. For instance, "to run" and "a run" are coming from the same lemma according to CELEX. Yet "a run" could mean, inter alia, a score in baseball while "to run" could mean to manage a place. This is a case where these distinct meanings could well be considered as homophones and not as derived meanings of the same base (polysemes).<sup>18</sup>

Despite these limitations, the present results show that there are some correspondences between what makes homophones easy to learn and how they are organized in the lexicon. Certainly this does not imply that learning difficulties, and learning facilities, translate directly into lexical structure. The distribution of homophones in the lexicon is also compatible with a pressure for communication: If words can be easily distinguished in context, there is more chances that the message will be transmitted accurately. The aim of the present work was not to distinguish between these pressures (I discuss possibilities of doing so in the General Discussion) but to point out that the homophones that are currently in the language display properties that make them learnable. This suggests, tentatively, that homophones that did not display these properties may have been eliminated from the lexicon across language evolution.

---

<sup>18</sup>Note that the boundary between certain polysemous words (called irregular polysemous) and homophones is unclear (Rabagliati & Snedeker, 2013)

## 3.4 Summary and Discussion

At the beginning of this chapter I laid out that similar-sounding words and ambiguous words were a challenge for children who need to determine what counts as a novel word and what does not (*the identification problem*).

In **section 3.1**, I revisited children's failure to learn similar-sounding words (learning "tog" when "dog" is already in their lexicon Swingley & Aslin, 2007). I proposed that toddlers may resist considering "tog" as a novel object label because its neighbor "dog" is also an object, such that both words share too many commonalities between them. To increase the likelihood that a novel form such as "tog" could be interpreted as a novel word, I manipulated the *syntactic context* of the form. In particular, 18-month-olds were taught object labels that were phonological neighbors of a familiar noun (as "tog" was, for "dog"), or neighbors of a familiar verb (like teaching "kiv", a neighbor of "give"). Children successfully learnt the verb-neighbors but failed to learn the noun-neighbors. Thus, manipulating the linguistic context by placing a verb-neighbor in a noun syntactic frame indicated to children that a new meaning was appropriate for the novel word form. Learning neighbors of familiar words is difficult for toddlers, but this difficulty disappears when the novel words appear in contexts that are sufficiently different from their known neighbors.

In **section 3.2**, I investigated whether children's ability to learn homophones depends also on the context these words are presented in. I showed that 20-month-olds are willing to learn a second meaning for a word they know, provided that the two homophones are sufficiently distant syntactically (e.g. "an eat" is a good label for a novel animal), or semantically (e.g. "a sweater" for a novel animal), but not when they are close (e.g. "a cat" for a novel animal). This suggests that when the two members of a homophone pair appear in contexts that are sufficiently distinct (either syntactically or semantically) they are learnt as easily as non-homophones.

As the present results suggest, minimal pairs and homophones may be less problematic than previously thought, at least when each member of the pair appears in different syntactic or semantic contexts that indicate to children that a meaning distinction is necessary.

In **section 3.3**, I evaluated whether similar-sounding words and homophones are distributed in the lexicon in a way that makes them more distinctive, thus more learnable. Interestingly enough, my results show that despite being more learnable, there is no pressure for minimal pairs to appear across distinct syntactic categories in the lexicon and to be semantically distinct (see section 2.2). In contrast, such a pressure exists for homophones, that is, there are more across-categories homophones than expected by chance and that homophones were as semantically dissimilar as any random pair of words in the lexicon.



Note that this trend was observed for all 4 languages under study (Dutch, English, French and German) suggesting that these results are robust cross-linguistically (though further investigation with more language families needs to be done in order to confirm this).

How can we explain that minimal-pairs and homophones are distributed differently in the lexicon? I suggest that these differences reflect the influence of different functional pressures associated with language acquisition and language use. There may be a greater advantage for minimal pairs to be more semantically similar and clustered within the same syntactic category than for homophones. I suggested that a compressible lexicon, or clumpy lexicon, would be advantageous for speech production (i.e., re-use of common articulatory sequences), memory (i.e., less sound sequences to remember) and learning categories (i.e., phonological similarity may facilitate the identification of syntactic and semantic classes). Additionally such a clumpy lexicon would also be advantageous to segment speech into words (see the Introduction 1.2.1, Altvater-Mackensen & Mani, 2013). A pressure for clumpiness would prevail over a pressure for distinctiveness because adults, and children as I showed here, are able to use fine-grained contextual cues to access the relevant meaning of the word successfully (see the relevant discussion in section 3.3). On the contrary, homophones show an advantage of distinctiveness over clumpiness, suggesting that because of form-identity, there is no choice but for these words to be distinctive in meaning in order to be learnable and transmitted with accuracy.

In sum, the lexicon is the theater of functional tradeoffs: Several pressures seem to be at play in the lexicon and influence differently the set of word forms present in the language and how they associate to meanings. An important future direction for this work will be to determine exactly how such functional tradeoffs arise in the course of language use and its acquisition depending on the phenomenon (e.g., minimal pairs, homophones) under study.

w



## 4 Theories of word learning and homophony: what is missing, what do we learn

Learning the meaning of a word is not an easy task, though children make it appear this way. Certainly, after a few presentations children are able to correctly identify the appropriate referent of a novel word they just have been taught (see Chapter 3). Yet, it is unclear what exactly they have "learned" about the word and which lexical representation they have formed. Suppose that a speaker uses "banana" to refer to the fruit the child is eating, what can the child infer about the word "banana"? At this point, even assuming that the child understood which object is referred to by the word in context (a non-obvious problem, see section 1.2.3), the meaning of "banana" is still undetermined (Quine, 1960). Certainly "banana" could refer to the set of all bananas and only bananas but many other meanings are consistent with that one experience: the set of all fruits, the set of all yellow objects and so on.

Existing theories of word learning have stressed the importance of prior knowledge or constraints about possible word meanings to constrain the learning problem faced by the child (e.g., Bloom, 2001; Goodman, 1955; Markman, 1989). For instance, children assume that novel labels refer to whole objects rather than parts of the object (*the whole object constraint*, Markman, 1991), to objects of the same type (*the taxonomic constraint*, Markman & Hutchinson, 1984) and to unnamed objects (*the mutual exclusivity constraint*, Markman & Wachtel, 1988). These constraints on word learning have specifically addressed the problem of learning unambiguous words where a single form is used to refer to a single meaning.

I propose that current accounts of word learning face massive challenges which can be revealed by trying to incorporate more word learning phenomena into the picture. In this chapter, I focus in particular on the role of homophony: what are the factors that lead language learners to postulate homophony for a new word? and what does homophony reveal about the word learning algorithm? Specifically, in **Section 4.1**, I investigate whether observing a "gap" in conceptual space between the learning exemplars for a given word or the intervention of other lexical items in that gap, lead adult learners to postulate

#### 4 Theories of word learning and homophony: what is missing, what do we learn

---

homophony for a word. In **Section 4.2**, I take this results further by looking at children's acquisition of homophones.

## What homophones say about words

Dautriche Isabelle

Chemla Emmanuel

Laboratoire de Sciences Cognitives et Psycholinguistique, (DEC-ENS/EHESS/CNRS), Paris,  
France

### Acknowledgements

Funding: the European Research Council under FP/2007-2013-ERC n°313610, ANR-10-IDEX-0001-02, ANR-10-LABX-0087, Direction Générale de l'Armement (PhD program Frontières du Vivant). We thank Anne Christophe, Paul Egré, Jean-Rémy Hochmann, Alexander Martin, Philippe Schlenker, Benjamin Spector, Brent Strickland

### Abstract

Homophones are word forms associated with two separate meanings. Two sets of experiments documented two factors that lead adults to postulate homophony for newly learned words: the gap in conceptual space between the learning exemplars for a given word and the intervention of other lexical items in that gap (Experiment 1). These effects were modulated by zeugmas, linguistic manipulations coherent with the presence/absence of homophony (Experiment 2). We show how homophony yields a challenge to current accounts of word learning, which share, explicitly or implicitly, a “convexity constraint”: learners seek to associate a given phonological form with a single meaning, whose extension is convex in conceptual space. Homophones, however, cover disconnected areas in conceptual space and therefore call for an explicit way to integrate them into current models of word learning.

Keywords: word learning; concepts; word meaning; lexical representation; homophony; psycholinguistics

Learning the word “cat” implies associating the sequence of sounds /kaet/ to the set of all cats and only cats. Quite generally one description of the meaning of a content word is its “extension”, i.e. the set of all entities to which that word refers (an idea discussed in detail in the tradition of formal semantics at least since Frege, 1892). But language learners need to infer the extension of a word based on a set of exemplars that surely do not exhaust that extension. The underlying inference problem would be unsolvable without prior knowledge, most notably some that could constrain the hypothesis space, which is the set of potential meanings for words (e.g., Bloom, 2001; Goodman, 1955; Keil, 1989; Markman, 1989; and Mitchell, 1980 for a formal proof).

Such priors have been described over different forms of conceptual structure (e.g., Tenenbaum, Kemp, Griffiths, & Goodman, 2011) in association with different models of meaning inference (e.g., Bayesian inference, Xu & Tenenbaum, 2007; Hypothesis elimination, Siskind, 1996; Associative learning, Regier, 2005). Irrespective of these differences, all current accounts assume that those concepts that happen to have word forms associated with them are *convex*, that is, they form a coherent cluster in conceptual space (e.g., Gardenfors, 2004). For instance, an arbitrarily motivated concept such as DOG OR TABLE is not a possible concept because it does not form a coherent class of objects (see Murphy & Medin, 1985 for the idea of “conceptual coherence”). Crucially, all current accounts of word learning transpose the idea of *concept* convexity into an explicit or implicit “convexity constraint” on *words*: learners seek to associate a given phonological form with a *convex* extension. That is, if A and B can be labeled using the sound /kaet/, then all objects falling in between A and B in conceptual space can also be labeled with the word /kaet/.

There is evidence suggesting that language learners may start with a convexity constraint. For instance, toddlers and preschoolers prefer to extend a novel word (e.g., “blicket” designating a dog) to an object of the same kind (e.g., a cat) rather than to an object of a different kind (e.g., a bone) (e.g., Markman & Hutchinson, 1984; Waxman & Gelman, 1986; see also the “shape bias”, showing that infants extend a label on the basis of the shape, Landau, Smith, & Jones, 1988). Hence, children seem to expect that the extension of a given label groups together objects that share a common property. If such a property is prioritized in the way we organize the world (we discuss the importance of some properties over others to define conceptual spaces further in the General Discussion), then it follows that extensions of words should follow a convexity constraint.

Our goal is to study how homophones fit into this picture, because they present a challenge to the convexity constraint. Indeed, words do not have to point to individual

concepts. A homophone is a phonological form associated arbitrarily with *several* meanings (contrary to polysemy, see e.g., Rabagliati & Snedeker, 2013; Rabagliati, Marcus, & Pylkkänen, 2010), which together form a discontinuous set in conceptual space. For instance, the English word “bat” applies both to the convex concept of ANIMAL BATS and to the convex concept of BASEBALL BATS, but, regardless of how the conceptual space is constructed, not all intervening objects sharing a common property of animal bats and baseball bats count as bats. However, all current approaches implement the convexity constraint to reduce the hypothesis space and therefore mechanically predict that when encountering this word form that applies to animal bats and baseball bats, English learners should conclude that *bat* applies to any intervening object, as would words that apply very broadly, such as “thing” or “stuff”. The very existence of homophony in human languages thus shows that learners do not adhere blindly to a convexity constraint. This challenges the details of any account that rely on the convexity constraint to reduce the hypothesis space.

Concretely, our point of departure will be work by Xu and Tenenbaum (2007). One advantage of their study is that it implements the convexity constraint in a predictive model, but it also provides the means to test it in a non-circular way. To do so, they first gathered similarity judgments between pairs of objects, and inferred a tree-structure over the whole set. This tree structure represents the taxonomy between the objects: different dogs are close together and form a subtree, mammals form a (bigger) subtree, etc. Such a hypothesis space reflects the taxonomic assumption (Markman, 1989) that requires words to label the nodes of a tree-structured hierarchy of natural concepts, in line with developmental data (e.g., Keil, 1989; Markman & Hutchinson, 1984; Markman, 1989; Waxman & Gelman, 1986). Crucially, Xu and Tenenbaum used this structured conceptual space to test a model of word learning according to which the extension inferred for a given word label should be a set of objects with no gap in conceptual space and which minimally includes all exemplars. Accordingly, they demonstrate that, when exposed to a set of learning exemplars, participants generalize the extension of the exemplars’ label to the smallest subtree that contains all these exemplars (their convex hull). In other words, participants pick the smallest generalization that satisfies the convexity constraint.

The present study explores the situations that lead language learners to postulate homophony for a new word using the word learning paradigm used by Xu and Tenenbaum. In Experiment 1, we manipulate two factors that should invite learners to favor a homophone interpretation of a novel label:

a) The *size of the gap*, in conceptual space, that separates different learning exemplars of a given word. To learn a homophone, language learners are exposed to a



discrete set of learning exemplars. For instance, for the word *bat*, they would observe several animal-bats and several baseball bats. However if the underlying true concept were the broad category that encompasses animal-bats, baseball-bats and all intervening objects (e.g., “thing”), then presumably learners would not observe exemplars confined to two corners of this set. Rather, they would observe a *set* of learning exemplars randomly (uniformly) sampled from the broad category. Observing exemplars clustered at two distant positions in the hypothesis space, i.e., observing a large gap between the exemplars may boost the likelihood that the exemplars are sampled from two independent categories, favoring a homophone interpretation.

b) *The intervention of other lexical items in that gap.* Evidence for homophony may also come from other words in the lexicon. There has been much evidence that words and their underlying concepts mutually constrain each other. For instance, language learners assume that words do not overlap in meaning (the “mutual exclusivity effect”; e.g., Markman & Wachtel, 1988). Having evidence that an additional label point towards an intervening region of the conceptual space (e.g., between animal-bat and baseball bats) may help learners discover more subtle configurations about how words map onto meanings.

Our results show that participants refrain from associating a label to a broad concept encompassing all the exemplars. Yet it does not entail that learners postulate homophony in these cases. We address this question more directly in Experiment 2. All in all, our results suggest that the effects documented in Experiment 1 are the footprints of homophony. This shows that current accounts of word learning face new challenges when incorporating homophony into the picture and that homophony can reveal (some of) the existing constraints on how words are associated with concepts in general.

#### **Experiment 1: gap in conceptual space and overall structure of the lexicon**

We used a word learning paradigm à la Xu and Tenenbaum (2007): participants were exposed to a new label through a couple of learning exemplars and asked whether the label should be extended to test items. We introduced a) a large gap in conceptual space between learning exemplars b) an intervening exemplar with a different label in that gap. We predicted that these two manipulations would lead to a breaking point after which participants would violate the convexity constraint, i.e., exclude items in the gap from the extension of the label.

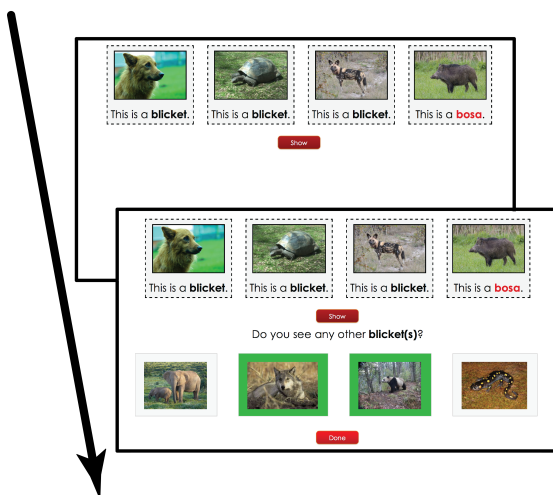
#### **Method**

**Participants.** One hundred and five adults were recruited through Amazon’s Mechanical Turk (45 females; M = 33 years; 102 native speakers of English) and

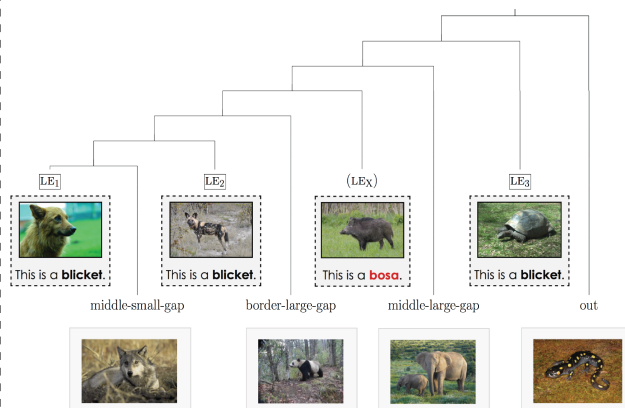
compensated \$0.50 for their participation. We excluded participants for lack of engagement in the task (criterion: participants who selected no test item in more than 50% of the “attractive” trials, in which at least 3 items should have been selected, see below;  $n = 0$  in Experiment 1A,  $n = 16$  in Experiment 1B) and participated in both versions of the experiment or in a previous pilot version ( $n = 3$  and  $5$ ). This resulted in 41 participants in Experiment 1A and 40 participants in Experiment 1B. Data collection was stopped when each of the experiment had at least 40 participants. The number of participants was established before data collection began.

**Procedure and display.** Participants were tested online. They were instructed that they would be exposed to words from an alien language and would have to select images that correspond to those words. In the instructions, participants were shown an example of a trial with pictures and a label that would not appear during the test. In each trial, participants first saw 3 or 4 learning exemplars, presented as the combination of a picture and a sentence. The first three learning exemplars (referred to as  $LE_1$ ,  $LE_2$  and  $LE_3$  below) were presented in random order and labeled with a novel word, e.g., *blicket*, via a prompt of the form “This is a blicket” underneath each of them. The fourth learning exemplar ( $LE_x$  below), if present, was labeled with another novel word highlighted in red, as in e.g., “This is a *bosa*” and was always the right-most exemplar. Once participants pressed a button “Show”, they would see a set of 4 pictures below the learning exemplars and be asked to select from these test items which one(s) could be labeled with the first novel word: “Do you see any other *blicket(s)*?” (see Figure 1.1). They responded by clicking to select none, one or multiple test items. When a picture was selected, its frame became green. Participants could unselect their choice by clicking on it again. Once a response was validated, the set of selected pictures was recorded and the test continued to the next trial.

1. Screenshots from Experiment A1



2. Schema of the structure of a trial in conceptual space



**Figure 1.** 1) Screenshots from Experiment 1A. Participants first see the 3 learning exemplars for the word “blicket” and one optional learning exemplar for the word “bosa”. After pressing the “show” button they then see the test pictures and are asked to find the other blickets. Once the pictures are selected (green frame), participants submit their answers by pressing the “done” button. 2) Schema of the structure of a trial in conceptual space. The first row of pictures corresponds to the learning exemplars ( $LE_1$ ,  $LE_2$ ,  $LE_3$ ,  $LE_X$ ) and the second row to the test items. The intervening item  $LE_X$  appeared only in half of the test trials (hence the parentheses).

**Conditions.** Each participant saw 12 test trials and 10 filler trials.

*Test trials.* The structure of test trials is represented schematically in Figure 1.1, the key factor is how the learning exemplars ( $LE_1$ ,  $LE_2$ ,  $LE_3$  and optionally  $LE_X$ ) were spread in conceptual space (here a tree-structure) and how the test items were distributed between them. As shown in Figure 1.2, there were two gaps between the exemplars: one small gap between  $LE_1$  and  $LE_2$  and one much larger gap between  $LE_2$  and  $LE_3$ . Test items were picked somewhere in the middle of the first small gap (*middle-small-gap*), of the large gap (*middle-large-gap*), in the large gap but close to the corresponding exemplars (*border-large-gap*) or out of all the exemplars altogether (*out*).

Six of the test trials, “Gap trials”, were designed solely to test the effect of the size of a gap between learning exemplars. They displayed three learning exemplars ( $LE_1$ ,  $LE_2$ ,  $LE_3$ ) associated with a to-be-learned label. According to the convexity constraint, participants should select all test items in the minimal subtree containing all learning exemplars, but we

expected that participants would be willing to violate this constraint and exclude *middle-large-gap* (or not as much as *middle-small-gap*).

Another 6 test trials, “Gap+Intervention trials”, had a fourth learning exemplar with a secondary label (the  $LE_x$  *bosa* exemplar in Figure 1). The convexity constraint applies to single lexical entries and is in principle blind to the rest of the lexicon, but we expected that participants would select the *middle-large-gap* test item less in these trials with an intervening label than in the test trials without this intervening label.

*Filler trials.* One filler trial was presented first so that participants could familiarize themselves with the task (with no particular indication of it however). Nine other fillers were randomly interspersed between the test trials. 6 “attractive” fillers were designed such that participants would select at least 3 of the 4 test pictures (3 of these filler trials contained three learning exemplars, all with the same label as in the Gap test trials, and 3 others included a fourth learning exemplar with a secondary label as in the Intervention test trials). 3 “repulsive” fillers implemented the opposite bias: participants were expected to select one or no test picture.

**Materials.** Our stimuli relied on a set of to-be-learned labels and taxonomically organized objects.

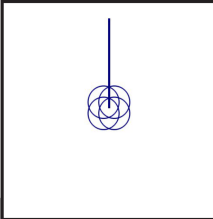
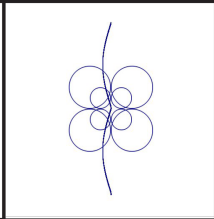
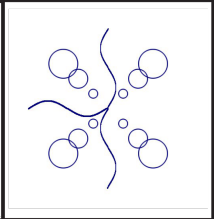
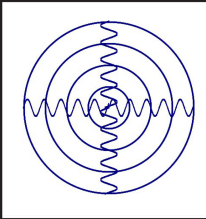








*Labels.* 28 phonotactically legal non-words of English were used for both experiments and were not repeated across trials.

*Objects in conceptual space.* We tested participants on two sets of objects organized into drastically different taxonomic hierarchies: natural objects, with a similarity measure based on phylogenetic trees (Experiment 1A) and artificial objects constructed in a parametric fashion, so that a similarity measure between these objects can be defined in a canonical way (Experiment 1B; Figure 2). Objects from this artificial taxonomy do not exist such that the actual lexicon of our participants cannot influence our experimental results.

One important difference with Xu and Tenenbaum’s paradigm is that our conceptual space did not rely on subjective, experimentally-gathered similarity judgments, but rather on objective similarity measures: one based on the distance in the phylogenetic tree and the other based on the parameterization of the objects. Surely these measures are only a proxy for participants’ representation of the similarity relationships between the objects. Yet, any effect that can be detected from these imperfect objective measures will retrospectively validate that it is a good approximation of the underlying subjective measure. We describe

#### 4 Theories of word learning and homophony: what is missing, what do we learn

the two sets of objects at the basis of Experiments 1A and 1B, their structure, and how our experimental conditions were obtained in each case in the Supplemental Material. The experimental material for both experiments is available at [https://osf.io/u473e/?view\\_only=33576a1ac18746b08d7e3fcc96e10e9a](https://osf.io/u473e/?view_only=33576a1ac18746b08d7e3fcc96e10e9a)

					
<b>Parameters</b>	Core pattern	4 overlapping circles 	4 tangent circles 	4 circles 	1 circle 
	Core pattern occurrences	1	2	3	4
	Size of the core pattern	25%	50%	75%	100%
	Number of radial lines	1	2	3	4
	Number of bumps in the radial lines	0 	1 	2 	8 

**Figure 2.** Examples of the artificial stimuli used in Experiment 1B, out of a set of 1024 possible unique combinations obtained from 5 parameters (core pattern, core pattern occurrences, size of the core pattern, number of radial lines, number of bumps in the radial lines) with 4 levels each.

**Presentation and trial generation.** The order of the trials as well as the pairing between the labels and the set of learning exemplars was fully randomized and differed for each participant. All trials were generated automatically following the algorithmic constraints described in the Supplemental Material for each stimuli type.

**Statistical analysis.** In a mixed logit regression (Jaeger, 2008), we modeled the selection of a test item (coded as 0 or 1) for each experiment (natural or artificial stimuli). Both models included two categorical predictors with their interaction: Test Item (*middle-small-gap*, *border-large-gap*, *middle-large-gap*, *out*) and Trial Type (Gap vs. Gap+Intervention) as well as a random intercept and random slopes for both Test Item and Trial Type for participants. We coded our predictors such that selection of *middle-large-gap* for Gap trials served as a baseline (unless otherwise mentioned) against which we compared a) responses to the other test items, b) the responses to *middle-large-gap* in Gap+Intervention trials.

All analyses were conducted using the lme4 package (Bates & Sarkar, 2004) of R. Ceiling effects (both in the choice of *middle-small-gap* and *out*, see Figure 2) impacted the log-estimation behind the logit models such that even obvious effects revealed by mere visual inspection of the data were not captured by the analysis. To get rid of these ceiling effects in a highly conservative way we introduced random noise: we ran the same analysis on a modified dataset where we randomly changed 5% of the responses within each Trial Type and each Test Item.

## Results

Figure 3 reports the average proportion of selection of each test item by Trial Type (Gap vs. Gap+Intervention trials) and Experiment (1A or 1B).

For Gap trials (Figures 3.1a and 3.2a), we replicate the minimal category effect seen in previous results (i.e., Xu & Tenenbaum, 2007) showing that participants are more likely to select a test item belonging to the category which is minimally consistent with the exemplars (*middle-small-gap*, *border-large-gap*, *middle-large-gap*) than a test item outside of this category (*out*), both for Experiment 1A ( $\beta = -1.56$ ,  $z = -8.80$ ,  $p < .001$ ) and Experiment 1B ( $\beta = -5.17$ ,  $z = -11.46$ ,  $p < .001$ )<sup>1</sup>. Crucially, the size of the gap between learning exemplars modulated the convexity constraint. That is, participants selected *middle-small-gap* items more than *middle-large-gap* items both in Experiment 1A ( $M_{middle-large-gap} = 0.59$ ,  $M_{middle-small-gap} = 0.98$ ;  $\beta = 0.83$ ,  $z = 3.85$ ,  $p < .001$ ) and in Experiment 1B ( $M_{middle-large-gap} = 0.44$ ,  $M_{middle-small-gap} = 0.94$ ;  $\beta = 3.04$ ,  $z = 9.55$ ,  $p < .001$ ). Participants were sensitive to the distribution of the learning exemplars with natural stimuli but also with unfamiliar stimuli. This latter case shows that familiarity with the categories (e.g., mammals, carnivores, animals) and possible existing labels for them cannot fully explain the results.

For Intervention trials (Figures 3.1b and 3.2b), we first replicate the effect described above: participants were sensitive to the size of the gap between the exemplars, that is, they selected *middle-small-gap* more than *middle-large-gap* in Experiments 1A ( $M_{middle-large-gap} = 0.43$ ,  $M_{middle-small-gap} = 0.97$ ;  $\beta = 3.82$ ,  $z = 9.75$ ,  $p < .001$ ) and in Experiment 1B ( $M_{middle-large-gap} = 0.32$ ;  $M_{middle-small-gap} = 0.82$ ;  $\beta = 2.70$ ,  $z = 10.32$ ,  $p < .001$ ). Crucially, we expected that the

<sup>1</sup> We Helmert-coded the predictor Test Item to compare the choice of *out* to the choice of the rest of the test items as a group.

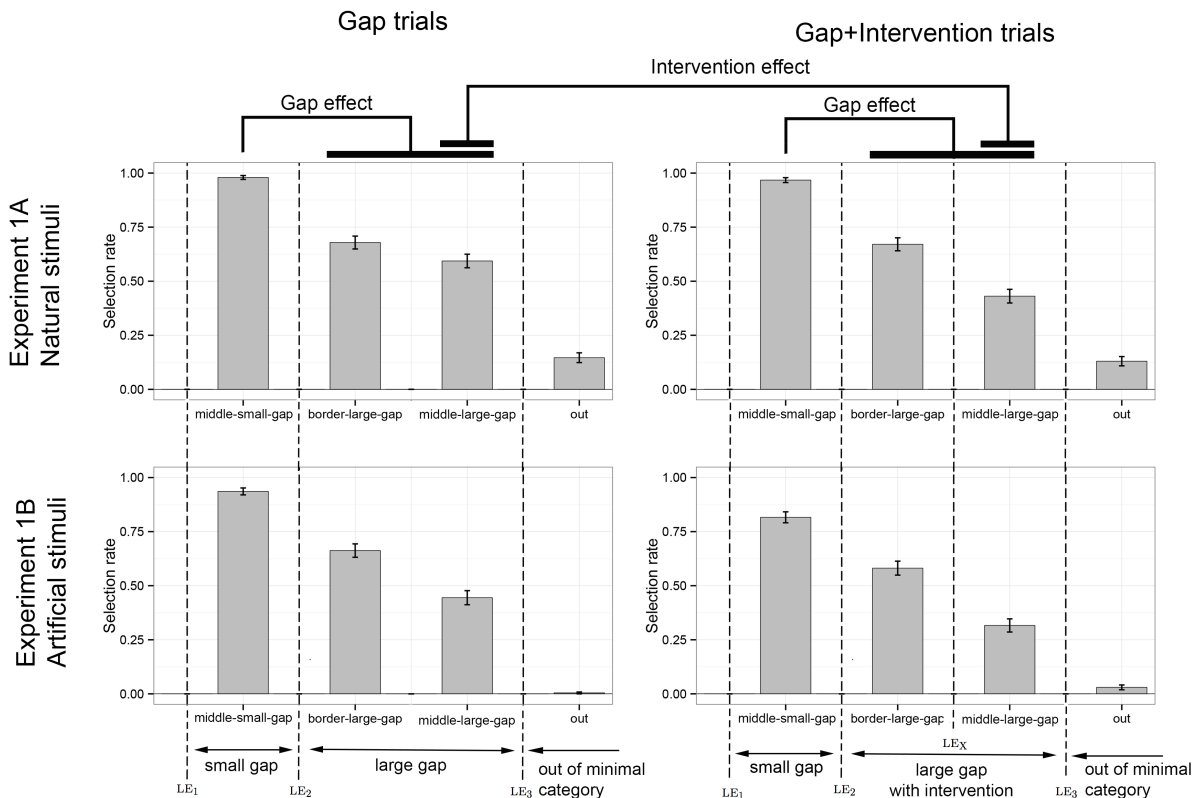
#### 4 Theories of word learning and homophony: what is missing, what do we learn

presence of an intervening item would increase participants' violation of the convexity constraint.

Indeed, in Experiment 1A, participants selected *middle-large-gap* less in Gap+Intervention trials than in Gap trials ( $\beta = -0.68, z = -3.65, p < .001$ ). Yet, the presence of an intervening lexical item did not affect the choice of any other test items (all  $ps > 0.7$ ) leading to an interaction effect: the difference between the selection rate of *middle-small-gap* and *middle-large-gap* was greater in Gap+Intervention trials than in Gap trials ( $\beta = 0.74, z = 2.62, p < .01$ ).

In Experiment 1B, participants similarly selected *middle-large-gap* less in Gap+Intervention trials than in Gap trials ( $\beta = -0.52, z = -2.18, p < .05$ ). But we should pause and note that the same was true for *middle-small-gap* items ( $\beta = -1.21, z = -3.68, p < .001$ ; here the intercept reflected selection of *middle-small-gap* in Gap trials). This was because the intervening exemplar  $LE_x$  was sometimes close to *middle-small-gap* (and even closer than it was to *middle-large-gap*), thus introducing an independent reason not to select *middle-small-gap* in these intervention trials.

Overall, we did observe that intervening labels block the extension of a word to the minimal category including all observed exemplars, even though this effect was polluted for artificial stimuli.





**Figure 3.** Proportion of choice of each test item averaged across Experiment 1A with natural objects (upper panel) and Experiment 1B with artificial objects (lower panel) for each trial type (Gap vs. Gap+Intervention trials). The x-axis follows (with some simplification) the structure in conceptual space: the position of the learning exemplars is indicated among the bars for the test items with the dashed lines. Error bars indicate standard errors of the mean.

## Discussion

We highlighted two factors that disturb the association of a word form to the single category that minimally includes all its learning exemplars: a) the size of the gap between the exemplars; b) the presence of intervening lexical items. There may be three potential interpretations for these results:

1) Participants associated a label to two meanings that *each* satisfies the convexity constraint. That is, participants postulated homophony, a non-immediate way to bind labels and concepts.

2) Participants associated a label with a set covering entities from several *disjoint* concepts (e.g., as in DOG OR TABLE), either because meaning discontinuity is acceptable or because the specific experimental task that we propose led them to do so.

3) Participants did not associate the new word with a meaning at all. Instead, they simply went by similarity of the test items to the learning exemplars: they selected more the objects close to the exemplars (*middle-small-gap*) than to the objects further away from them (*middle-large-gap*). The role of the intervening label may be harder to account for in this view, but one may imagine some strategic effect such that if an object is close to some irrelevant object X, it will decrease the tendency to say that this object belongs to a set that was not said to contain X.

Experiment 2 was designed to distinguish between these three interpretations.

## Experiment 2: linguistic manipulations

Homophones interact with linguistic constructions in a characteristic way. Zeugmas are the typical rhetorical device used to pun on the different senses of ambiguous words (e.g., Cruse, 1986; Zwicky & Sadock, 1975) and have been extensively used as a test to distinguish words with an extension that covers a broad category from polysemous and ambiguous words (e.g., Cruse, 1986; Geeraerts, 1993). Consider for instance “John and his driving license expired last Thursday” (Cruse, 1986), where the verb “expire” has two distinct, but related, senses (i.e. “died” and “not valid anymore”). If the zeugmatic sentence is acceptable, it shows that the relevant word is polysemous or ambiguous (the two meanings are distinct) rather than vague (the boundary between meanings are indistinct).



Interestingly, zeugmas can be used to distinguish between a homophone, where a label applies to two convex concepts, and a word associated with a disjunctive meaning, where a label would apply to a disjoint concept. For instance, if “blicket” maps onto a disjunctive concept, such as DOG OR TABLE, it should be possible to use a plural sentence “these are two blickets” when pointing to a dog and a table, while it would be zeugmatic to say “these are two bats”, pointing at one animal-bat and one baseball-bat. This is explained in a theory of homophones in which two words, with different meanings, share the same form: one cannot use a single phonological form to refer to both meanings at the same time. However, different *tokens* of the phonological form may pick out different meanings: it may therefore be more natural to say in a situation as above “This is a bat (pointing at the animal-bat), this is *also* a bat (pointing at the baseball-bat)”.

We will use these two constructions to test whether the effects we documented in the experiments above are the signatures of homophony. If participants postulated homophony, the *plural* zeugmatic construction, which is not compatible with homophony, should increase the tendency to form a single convex category encompassing all learning exemplars (as dictated by the convexity constraint in the absence of homophony), compared to the *also* construction. This would be evidence that participants did not postulate that a label could map onto a discontinuous concept and that our effects are not solely driven by similarity, since the similarity of the test items to the exemplars is held constant across the two linguistic constructions.

#### **Method**

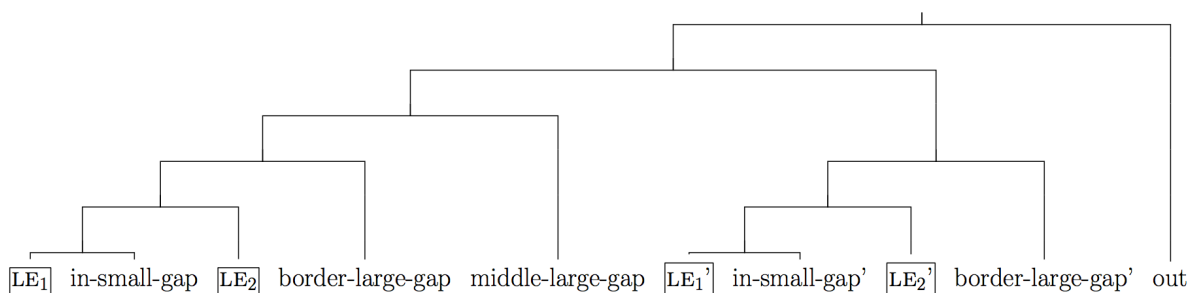
**Participants.** Ninety adults were recruited through Amazon Mechanical Turk (28 females; M = 30 years; 87 native speakers of English) and were compensated \$0.50 for their participation. We excluded subjects who participated in both conditions of the experiment ( $n = 3$ ). This resulted in 44 participants in the *also*-condition and 43 participants in the *plural*-condition. Data collection was stopped when each of the conditions had at least 40 participants. The number of participants was established before data collection began.

**Procedure and display** were similar to Experiment 1, except that each trial now included 4 learning exemplars and 6 test items.

**Conditions.** Each participant saw 8 test trials and 16 filler trials.

**Test trials.** As schematized in Figure 4, each test trial contained 4 learning exemplars ( $LE_1$ ,  $LE_2$  and  $LE_1'$ ,  $LE_2'$ ). We implemented symmetry in the distribution of learning exemplars such that there were two small gaps (between  $LE_1$  and  $LE_2$  and between  $LE_1'$  and  $LE_2'$ ) and one

large gap (between the two pairs of exemplars). This distribution of exemplars in conceptual space may favor the construction of sharp boundaries over two disjoint categories (see also discussion of the “size principle” in the General Discussion). The position of the six test items is shown in Figure 4. Two test items were placed inside the small gaps (*middle-small-gap* and *middle-small-gap'*), two items just outside of the minimal subtrees  $S(\text{LE}_1, \text{LE}_2)$  and  $S(\text{LE}_1', \text{LE}_2')$  containing each pair of exemplars (*border-large-gap* and *border-large-gap'*), one item inside the large gap (*middle-large-gap*, either attached to  $S(\text{LE}_1, \text{LE}_2)$  or to  $S(\text{LE}_1', \text{LE}_2')$ ) and one item outside of the minimal subtree containing all four learning exemplars (*out*).



**Figure 4.** Schema of the tree-structure of the items used in a trial for Experiment 2. The boxed items correspond to the learning exemplars.

The 8 test trials were created according to the schema in Figure 4, but their mode of presentation differed across the two conditions. In the *also*-condition, the four learning exemplars were presented in pairs: the left pair was labeled with a given word (e.g., “These are two blickets”) and the right pair with the same word using *also* (e.g., “These are also two blickets”). In the *plural*-condition, the four learning exemplars were ordered in pairs as in the *also*-condition but the four exemplars were grouped together in a gray frame and labeled at once via a plural sentence (e.g., “These are four blickets”; see Figure 5).

We expected that participants would select the test items *middle-large-gap* and *border-large-gap* more in the plural-condition than in the *also*-condition, because homophony is less of an option while using the plural construction.



**Figure 5.** Possible learning exemplars for a test trial as presented in 1) the plural-condition and 2) the also-condition.

*Filler trials.* 16 filler trials were interspersed, half of which were visually similar to the test trials of the plural-condition (Figure 5.1) and the other half were visually similar to the test trials of also-condition (Figure 5.2; but with a different label for the two pairs of objects and, of course, no *also* in the description).

**Material.** We used the same set of objects as in Experiment 1A and the same labels.

**Presentation and trial generation.** The experiment always started with 3 filler trials. All trials were generated pseudo-randomly following the constraints described in the Supplemental Material.

**Statistical analysis.** As before, we modeled the selection of a test item in a mixed logit model including two categorical predictors with their interaction: Test Item (*middle-small-gap*, *border-large-gap*, *middle-large-gap*, *out*) and Linguistic Condition (Plural vs. Also) as well as a random intercept and a random slope for Test Item on participants. The selection of *middle-large-gap* in the plural-condition served as a baseline (unless otherwise mentioned).

## Results

The results are presented in Figure 6. The pairs  $(LE_1, LE_2)$  and  $(LE_1', LE_2')$  played symmetric roles, we accordingly collapsed responses for *middle-small-gap* and *middle-small-*

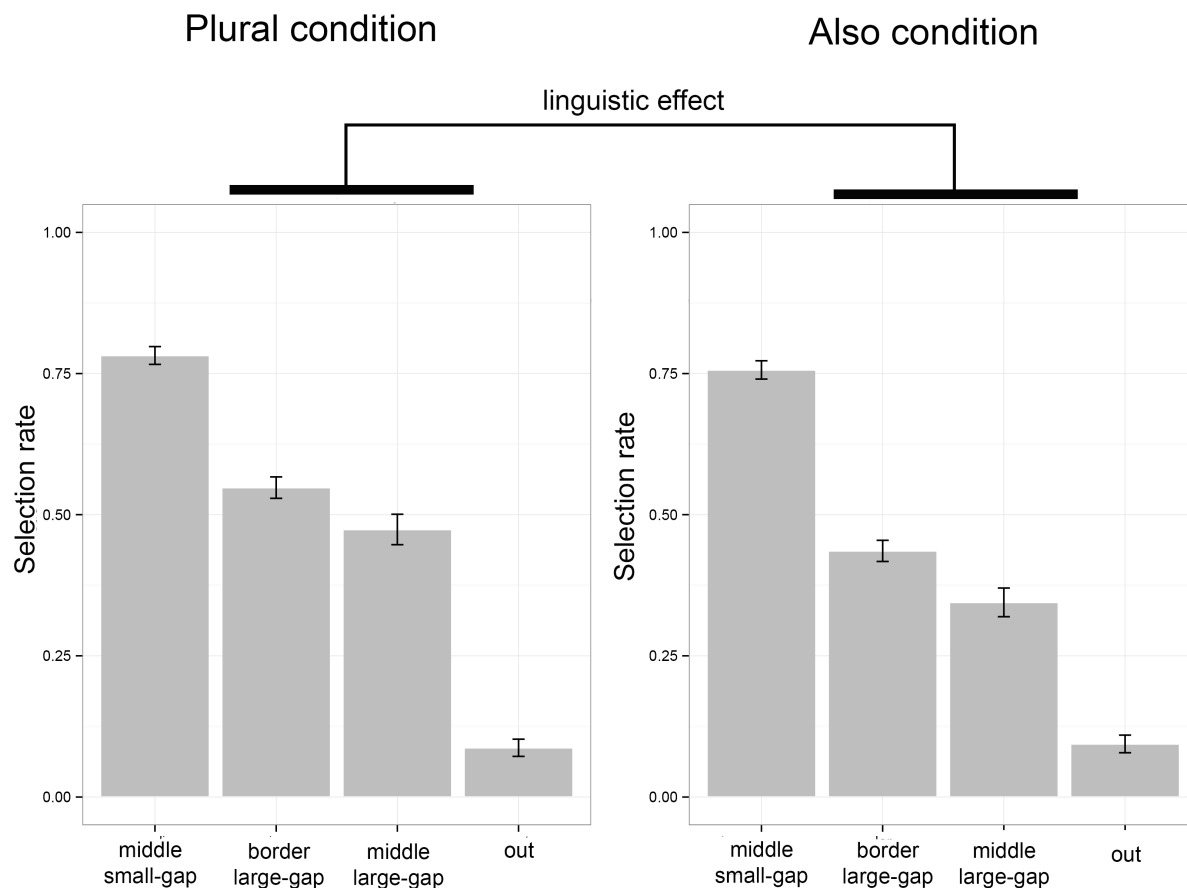
*gap'* and responses for *border-small-gap* and *border-small-gap'* (practically ignoring the prime sign in the report).

First, the results confirm the existence of a gap effect. Participants showed sensitivity to the sampling distribution of the exemplars, in that they selected more *middle-small-gap* than *middle-large-gap* in both the also-condition ( $\beta = 2.20$ ,  $z = 11.26$ ,  $p < .001$ ) and the plural-condition ( $\beta = 1.70$ ,  $z = 8.59$ ,  $p < .001$ ).

Interestingly, the minimum path (in terms of number of branches) between the learning exemplars  $LE_1$  and  $LE_2$  was smaller in Experiment 1A (mean for  $d(LE_1, LE_2) = 3.65$ ), compared both to  $(LE_1, LE_2)$  and  $(LE_1', LE_2')$  in Experiment 2 (mean for  $d(LE_1, LE_2) = 7.16$ ) (see Supplemental material). Accordingly, we found a cross-experiment gap effect such that *middle-small-gap* was less selected in Experiment 2 ( $M = 0.75$ ) than in Experiment 1A ( $M = 0.95$ ).

Our critical expectation concerned the comparison between linguistic presentations. Test items in the gap between  $(LE_1, LE_2)$  and  $(LE_1', LE_2')$  were selected more often in the plural-condition than in the also-condition: this was true both for *middle-large-gap* ( $\beta = 0.73$ ,  $z = 2.24$ ,  $p < .05$ ) and *border-large-gap* ( $\beta = 0.64$ ,  $z = 2.13$ ,  $p < .05$ ), resulting in an interaction effect: the difference between the selection rate of *middle-small-gap* (serving as a baseline) and the combined selection rate of *middle-large-gap* and *border-large-gap* was greater in the plural-condition than in in the also-condition ( $\beta = 0.46$ ,  $z = 2.02$ ,  $p < .05$ ).<sup>2</sup>

<sup>2</sup> We Helmert coded the predictor Test Item to compare the choice of *middle-small-gap* to the choice of the *middle-large-gap* and *border-large-gap* as a group.



**Figure 6.** Proportion of selection of each test item averaged across 1) The plural-condition using a linguistic construction discarding the possibility of homophony and 2) the also-condition using a linguistic construction more suitable to homophony. Error bars indicate standard errors of the mean.

### Discussion

When presented with a plural construction (e.g., “These are blickets”), participants were more likely to associate the word to a category that spans over all the exemplars than when they were presented with a construction compatible with homophony (e.g., “This is a blicket and this is *also* a blicket”). This effect suggests that the gap effect documented in Experiment 1 is the footprint of homophony mapping two words with the same phonological form onto two convex concepts and not of the association of a single word to a discontinuous category.

Certainly, in line with previous results (e.g., Goldstone, 1994), participants were guided in part by similarity: they extended a label more to an object close to the exemplars (*middle-small-gap*) than to an object further away (*middle-large-gap*) and they extended the label less to *middle-small-gap* in Experiment 2 than they did in Experiment 1 due to a greater

distance between the learning exemplars in Experiment 2.<sup>3</sup> Yet, such similarity effects cannot explain the main result of Experiment 2 since the linguistic manipulation is realized holding constant similarity relations among learning exemplars and test items. The amount of similarity-driven generalization in participants' responses could be quantified (see Xu and Tenenbaum for a model comparison of rule- vs. similarity-based model), but it is sufficient for our purposes to note that it cannot account for the entirety of the present effects, which are driven by linguistic manipulations alone in Experiment 2.<sup>4</sup>

One may also ask whether the plural/also effect is linked to the very specific linguistic constructions involved or whether it is merely driven by the visual, two-part presentation that co-varies with these constructions in our experiments. Importantly, a visual effect (i.e., a visual "zeugma") would make the same point as a more specific linguistic construction effect: all that matters for our argument is that there is room for two tokens of the same phonological form, either because two tokens are indeed present, or simply because the presentation introduces different labeling events.

### General Discussion

We documented two factors that reduce the tendency to map a phonological form onto a single, convex extension, an explicit or implicit assumption about learners in all current accounts of word learning: a) the size of the gap in conceptual space between learning exemplars; b) the presence of an intervening label for entities in that gap. These effects were modulated by linguistic manipulations coherent with the presence/absence of homophony. We submit that when encountering novel words in such situations, learners prefer to postulate homophony to preserve concept convexity, whereby a label applies to two convex concepts, rather than accepting that a label applies to a single discontinuous concept.

In the following we first come back to the very idea of a conceptual space, how to access such an abstract construct. Second, we move to the level of words and show how the current study of homophony is relevant to current accounts of word learning broadly, and why other phenomena should be subjected to the same scrutiny. Finally, our results are

---

<sup>3</sup> Note that this is also compatible with the size principle documented by Xu and Tenenbaum (2007): since the boundaries of the categories defined by the pairs of exemplars ( $LE_1$ ,  $LE_2$ ) and ( $LE_1'$ ,  $LE_2'$ ) were less sharp than in Experiment 1A, the correct level of generalization was more uncertain in Experiment 2 than in Experiment 1A.

<sup>4</sup> While we demonstrate that participants are sensitive to the linguistic constructions in which the words enter, we cannot tell whether the also construction alters the results in one direction (towards homophony), or whether the *also* construction pushes in the opposite direction (against homophony), or both.

based on adult data only and we discuss their relevance for children in the process of learning their native language.

##### **How to work with concepts**

A provocative question at this point is whether the notion of conceptual space is useful. One worry is that there may not be a stable metric between abstract entities across contexts (Tversky, 1977), such that two entities can be made arbitrarily similar by changing the dimension under consideration. For instance, one may consider that a Ferrari and a VW Beetle are closer to one another than a Ferrari and a diamond, but this similarity relation may reverse if the context involves paying attention to the value of entities. However, we submit that some dimensions are privileged: they are more stable across contexts, by default, and infants are biased to pay more attention to them (e.g., Poulin-Dubois, Lepage, & Ferland, 1996). For instance, animacy may be a property that is *privileged* in that sense, over say color, to categorize objects. It does not always have to be the case, but on average this will create the basis for a stable set of privileged features to (partly) provide a structure for conceptual space (see also Barsalou, 1983 for the notion of *ad hoc* categories and Keil, 1981; Osherson, 1978 for the idea of concept naturalness).

The next worry then is to decide how one can objectively assess what the actual, “privileged” metric in conceptual space is. Xu and Tenenbaum (2007) gathered subjective judgments of similarities, independently from the categorization task. In our studies, we decided on a structure of conceptual space prior to using it for our test. Specifically, our notion of convexity relied on phylogenetic trees and on an arbitrary metric over a multi-dimensional space of visual features. The hope was that there would be a sufficiently good matching between these idealized conceptual spaces and what participants would actually take to be the relations between the relevant entities. Since participants had access to the entities only through visual representations, one may worry that we over-evaluated the chances that perceptual features could determine concepts. Perceptual features as such may not be the determinant of conceptual structure, since concepts may be defined by non-observable properties. Several developmental studies show that, indeed, children prefer to draw inferences based on category membership than inferences based on perceptual appearances (e.g., Gelman & Coley, 1990; Gelman & Markman, 1987; Graham, Kilbreath, & Welder, 2004). Nevertheless, we use perceptual similarity as a proxy to reflect conceptual structure and follow previous work in that respect (see Medin & Schaffer, 1978; Nosofsky, 1986; Shepard, 1964; Smith & Medin, 1981; Xu & Tenenbaum, 2007). Interestingly, we note that young children may also use such a proxy in their earliest word meaning inferences (Graham & Poulin-Dubois, 1999; Landau et al., 1988).

The current inquiry was based on the hope that conceptual space could be approximately circumscribed by objective or scientifically based properties (e.g., phylogenetic trees). There surely has to be *some* correlation between such an objectively based categorization and actual, subjective categorization (e.g., Atran, 1998). Most importantly, the fact that our results come out the right way suggests *a posteriori* that our simplifying hypotheses are acceptable to a sufficient degree: our results could not be obtained if our assumptions to approximate the underlying conceptual structure were inappropriate.

### Challenges for accounts of word learning

The above discussion only refers to concepts, not to words. A natural assumption is that one word would map to one concept, but it does not have to be so. For instance, a word could map onto a set of concepts, as if there was a word meaning DOG OR TABLE (where DOG and TABLE here are supposed to be disjoint concepts). Our study of homophones shows that this does not happen. Instead, when a word could potentially have such a disjunctive, discontinuous meaning, a homophone is created. As a result, the convexity condition of concepts invades the level of words, but this comes at the cost of the existence of homophones, which constrains the learning device, as we will now discuss.

All current accounts of word learning presuppose a convexity constraint whereby word forms map onto a single meaning that ought to be convex in conceptual space. But this assumption bans homophony from the system. Specifically, it seems that incorporating homophony in the best current views of word learning (i.e., Bayesian approaches of word learning, Xu & Tenenbaum, 2007) requires that the learning system allows for this from the start: children would come to the world with a learning mechanism able to learn non-homophones, just like the modern models of word learning, but there would have to be a different learning mechanism to track and learn homophones, one part of the system that is not currently described. If the architecture of the learning system were such that it separated homophones and non-homophones so sharply, it would be a very strong prior that amounts to saying that children know, innately, that their to-be-learned language will include homophones. If this prior cannot be eliminated (by tweaking the learning inference component, see our proposition regarding the possibility of a *sampling effect* below), the innate expectation of homophony would be a striking example of innate linguistic knowledge.

The problem is more general. Current models of word learning incorporate built-in constraints to reduce the hypothesis space. Sometimes these constraints not only reduce the hypothesis space, but also ban phenomena that are outside of the reduced hypothesis space. This is the situation we revealed for the convexity constraint and homophony. In principle, there are two solutions to this problem: one could hope that later refinements will



be able to get rid of the constraint or, if this is not obtained, one could take this difficulty at face value and postulate that the system implement independent learning modules (e.g., one for homophones and one for non-homophones), implying that the distinction between homophones and non-homophones is at least innately “expected”. Let us illustrate with another example. Xu and Tenenbaum (2007) propose a rational use of co-occurrences of words and objects to learn content words. Function words, however, occur in all sorts of contexts and may co-occur with all possible objects, in principle. Hence, the model predicts that words like “the” or “and” mean the same as “thing” or “stuff”, which also co-occur with any kind of object. Arguably, learners deploy a different strategy to learn content words and function words (see relatedly Piantadosi, Tenenbaum, & Goodman, 2012 for the use of a different strategy for learning numerical concepts). Yet before separating hypothesis spaces for content words and hypothesis spaces for function words, one would need to propose a mechanism that separates function and content words for rational reasons (see, e.g., Hochmann, 2013 for *empirical* facts that could support this rational mechanism). But the fact that this separation is triggered in the first place may have to be implemented in the prior part of the learning system, thus making it an innate component that languages contain both function words and content words.

In sum, current word learning accounts break the learning problem into manageable pieces of the puzzle, studying object labels, ambiguous words, functions words or numerical concepts separately. A reconciliation of these pieces into a single solution may be technically easy; one could say that the system “expects” these differences. But it has rich consequences because in the absence of a more complete picture, it amounts to postulating that subtle and quite specific phenomena such as the distinction between function words and content words or the existence of homophony have an innate basis.

### **Early language acquisition**

Through the study of homophones, our studies uncover several factors that play an important role in revealing the existing constraints on how words associate with concepts in general. An important open question is whether these factors influence word learning during the earliest stages of word acquisition. While studying adults may inform us about the general strategies involved in word learning (Markson & Bloom, 1997), children have different cognitive resources and biases and may consequently use different strategies. We detail four relevant factors that could lead to the emergence of homophony in children and leave their study open for future research:

- 1) *Concept convexity*. Adults refrain from associating a label to a broad concept when positive evidence is missing for a large gap within the concept. The observation that a label

applies to a discontinuous extension triggers the formation of novel word representations that are compatible with the convexity constraint. Do children also expect words to refer to coherent and convex concepts and, if so, what representation do they adopt when the convexity constraint is not met? Plunkett, Hu, and Cohen (2008) offer a relevant study in which they presented 10-month-old infants with exemplars of a word forming a gap in conceptual space: the presence of a similar label was enough for children to extend the label to all intervening items in that gap. Yet, they only tested rather small gaps, which may very well be before the breaking point of the convexity constraint.

2) *Sampling effect.* Xu and Tenenbaum (2007) document a “size principle” according to which the sharpness of a concept is a function of the number of learning exemplars, for both children and adults. We showed an effect of the *distribution* of the learning exemplars in conceptual space: observing exemplars clustered at two distant positions in the hypothesis space boosted the likelihood that the exemplars were sampled from two independent categories. Children are sensitive to the size principle; they may also show sensitivity to such a “distribution principle”, a possibility that we are currently exploring.

3) *The structure of the semantic lexicon.* When confronted with a new word, adults consider the existence of *other* (potentially unknown) words. Specifically, they generalize a word A less to a new object if this new object comes in the vicinity of a concept labeled by a word B. This demonstrates that learners have expectations about the structure of the semantic lexicon as a whole and priors about how words may share the conceptual space. This new kind of evidence against individual word-by-word learning is coherent with simpler, so-called “mutual exclusivity effects” (Markman & Wachtel, 1988), according to which a new word should not occupy the same conceptual space as a known word. Interestingly, this effect has to be modulated by other factors, since some words surely overlap in conceptual space (e.g., compare *cat* and *animal*). To our knowledge, priors over the whole lexicon are missing from current word learning computational models – and their implementation raises immediate challenges.

4) *Linguistic factors:* Adults’ generalization was modulated by the linguistic construction in which words were presented. While we used linguistic constructions as a linguistic test for homophony, these constructions may also be used to discover homophony (noting that homophones never appear in plural constructions but may appear in some more appropriate constructions such as the *also* construction we documented). Whether children are able to pick up on this is an empirical question, both because they may not be sensitive to these linguistic factors (effectively this would otherwise be a case of linguistic bootstrapping of homophony) or because the relevant facts may be too sparse in their input, e.g., if homophones cover distant concepts, it is unlikely that these two concepts will be mentioned within the same learning situation.

### Summary

In this work, we showed that a word is more likely to yield homophony if: (a) it is learnt from exemplars leaving an important gap between them (in conceptual space), (b) this gap in conceptual space is occupied by other words. We submit that encountering novel words in such situations may trigger forms of word representations which comply with concept convexity. More generally, we argue that incorporating homophony and other challenging word learning phenomena into current word learning accounts, will provide a better understanding of learners' implicit knowledge and assumptions about how word forms map onto meanings.

### References

- Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21(04), 547–569.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, 11(3), 211–227.
- Bates, D., & Sarkar, D. (2004). *Ime4 library*. Accessed.
- Bloom, P. (2001). Précis of How children learn the meanings of words. *Behavioral and Brain Sciences*, 24(06), 1095–1103.
- Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.
- Frege, G. (1892). Ueber Begriff und Gegenstand. *Vierteljahrszeitschrift Fuer Wissenschaftliche Philosophie*, 16, 192–205.
- Gardenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2), 9–27.
- Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)*, 4(3), 223–272.
- Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, 26(5), 796.
- Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, 1532–1541.
- Goldstone, R. L. (1994). Arguments for the Insufficiency of Similarity for Grounding Categorization.
- Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
- Graham, S. A., Kilbreath, C. S., & Welder, A. N. (2004). Thirteen-Month-Olds Rely on Shared Labels and Shape Similarity for Inductive Inferences. *Child Development*, 75(2), 409–427.
- Graham, S. A., & Poulin-Dubois, D. (1999). Infants' reliance on shape to

generalize novel labels to animate and inanimate objects. *Journal of Child Language*, 26(02), 295–320.

Hochmann, J.-R. (2013). Word frequency, function words and the second gavagai problem. *Cognition*, 128(1), 13–25.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, 88(3), 197–227.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. MIT Press.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299–321.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, 16(1), 1–27.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20(2), 121–157.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385(6619), 813–815.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.

Osherson, D. N. (1978). Three conditions on conceptual naturalness. *Cognition*, 6(4), 263–289.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2), 199–217.

Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2), 665–681.

Poulin-Dubois, D., Lepage, A., & Ferland, D. (1996). Infants' concept of animacy.

*Cognitive Development*, 11(1), 19–36.

Rabagliati, H., Marcus, G. F., & Pykkänen, L. (2010). Shifting senses in lexical semantic development. *Cognition*, 117(1), 17–37.

Rabagliati, H., & Snedeker, J. (2013). The Truth About Chickens and Bats: Ambiguity Avoidance Distinguishes Types of Polysemy. *Psychological Science*.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29(6), 819–865.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, 1(1), 54–87.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1), 39–91.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Harvard University Press Cambridge, MA.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Tversky, A. (1977). Similarity features. *Psychological Review*, 84, 327–352.

Waxman, S., & Gelman, R. (1986). Preschoolers' use of superordinate relations in classification and language. *Cognitive Development*, 1(2), 139–156.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

Zwicky, A., & Sadock, J. (1975). Ambiguity tests and how to fail them. *Syntax and Semantics*, 4(1), 1–36.

## Word learning: homophony and the distribution of learning exemplars

Isabelle Dautriche\*, Emmanuel Chemla, and Anne Christophe

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS), Ecole Normale Supérieure, PSL Research University, Paris, France

### Acknowledgements

This work was funded by grants from the European Research Council (FP/2007-2013-ERC n°313610), the Région Ile-de-France, the Fondation de France, LabEx IEC (ANR-10-LABX-0087), IdEx PSL (ANR-10-IDEX-0001-02), the ANR ‘Apprentissages’ (ANR-13- APPR-0012), and a PhD fellowship from Direction Générale de l’Armement (DGA). We thank Anne-Caroline Fievet and the language group of the LSCP.

### Abstract

How do children infer the meaning of a word? Current accounts of word learning assume that children expect a word to map onto exactly one concept whose members form a coherent category. If this assumption was strictly true, children should infer that a homophone, such as “bat”, refers to a single superordinate category that encompasses both animal-bats and baseball-bats. The current study explores the situations that lead children to postulate that a single word-form maps onto several distinct meanings, rather than a single superordinate meaning. Three experiments showed that adults and 5-year-old French children use information about the sampling of learning exemplars (and in particular the fact that they can be regrouped in two distinct clusters in conceptual space) to postulate homophony. This unexplored sensitivity and the very possibility of homophony are critically missing from current word learning accounts.

**Keywords:** word learning; homophony; concepts; constraints

---

\*Correspondence concerning this article should be addressed to Isabelle Dautriche, Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 29 rue d’Ulm – P.J., 75005 Paris – France. E-mail: [isabelle.dautriche@gmail.com](mailto:isabelle.dautriche@gmail.com)

### Introduction

To learn a word, language learners must draw a link in their mental lexicon between a phonological form and its meaning. While many words conform with a one-to-one mapping between form and meaning, this is not always the case: a homophone is a phonological form associated arbitrarily with several meanings, each of which corresponds to a concept. For instance, the word form “bat” applies both to the concept ANIMAL BAT and to the concept BASEBALL BAT. Hence, homophones present children with a non-standard word learning situation, for which they need to discover that there is a decoupling between linguistic signals and concepts.

In order to examine what kind of challenge homophony brings into the word learning task, let us consider a typical word learning situation. Children do not observe associations between words and concepts; rather, they observe the co-occurrences of word forms and exemplars of their associated concepts. Thus, one major problem for the learner is to infer the meaning of the word from a set of exemplars that is consistent with an unbounded number of possible meanings (Quine, 1960). Existing theories of word learning have stressed the importance of prior knowledge to constrain the learning problem faced by the child (e.g., Bloom, 2001; Goodman, 1955; Markman, 1989). Such priors have been described at the level of concepts (what are the possible concepts our mind is ready to entertain) and at the level of word forms (what are the constraints on possible *form*-concept configurations).

All current accounts of word learning (associative learning accounts, e.g., Regier 2005; Yu & Smith 2007; hypothesis elimination accounts, e.g., Pinker 1989; Siskind 1996; Bayesian accounts, e.g., Frank et al. 2009; Piantadosi et al. 2012; Xu & Tenenbaum 2007) assume that learners rest on two main assumptions: First, the structure of the language is clear, that is, involves transparent, uniform mappings between forms and concepts, leading to one-to-one correspondence across these domains (a constraint at the level of word forms). Second, those concepts are *convex* (a constraint at the level of concepts), that is, whose members form a group that share a common set of properties that holds them to be contiguous in conceptual space (see further the notion of concept convexity in Gärdenfors, 2004; and the related notion of conceptual coherence in Murphy & Medin, 1985). For instance, a concept such as CAR OR WATER, is not a proper candidate for a concept because its members would be drawn from two disjoint sets which do not form a convex cluster of entities in conceptual space. Yet there is a sense in which this convexity constraint does not translate at the level of word forms: if a phonological form is used to refer to two objects A and B, it does not imply that all objects between A and B in conceptual space can also be labeled by the same form, as in the case of homophones (see discussion in Dautriche & Chemla, *submitted*). In other words, the extension of the meaning(s) of a word form is not necessarily convex.

Evidence for a convexity constraint at the level of concepts comes from several experimental studies (Dautriche & Chemla, *submitted*; Xu & Tenenbaum, 2007). In these studies, the conceptual space is defined over a tree-structured representation of entities by clustering a set of entities based on their similarity. Subtrees correspond to categories that words could label at different levels of granularity (e.g., cat, feline, mammal, animal). When exposed to a set of learning exemplars uniformly sampled from a category, adults extend the label to the minimal subtree including all the exemplars (Xu & Tenenbaum, 2007). For example, when presented with three “feps” labeling three Dalmatians, adults readily extend “fep” to the set of all Dalmatians, would they be presented with a Dalmatian, a Labrador and a German-shepherd they would extend the label to the set of all dogs. Yet, when presented with exemplars clustered at two distant positions in conceptual space, such as {two primates, one mushroom}, adults did not extend the label to



all objects falling within a convex category encompassing all exemplars (i.e. LIVING BEINGS), rather, they preferred to restrict the label to members of two disjoint subcategories (i.e. PRIMATE and MUSHROOM). Furthermore, this effect is strengthened by the manipulation of specific linguistic evidence that signals or bans homophony, such that the occurrence in some particular constructions (so called zeugmas, Zwicky & Sadock 1975). The interpretation is the following. When encountering novel words, adults are first guided by the idea of a one-to-one mapping between words and concepts, and convexity of concepts thus yields word forms with convex extensions. Yet, when observing evidence against the convexity of the set of entities that may be labelled by a given word form (i.e. the presence of two distinct convex clusters of learning exemplars in conceptual space), adults postulate homophony, that is they prioritize concept convexity over the possibility that a word form maps onto a single concept (Dautriche & Chemla, *submitted*).

An important question is whether concept convexity influences word learning during language development. Do children also expect word forms to refer to convex concepts? In case this seems untenable, do they also know about the backup strategy, homophony? Much developmental work suggests that children start with the assumption that one word form maps onto exactly one concept (Slobin, 1973, 1975) and that they follow some convexity constraint in that they expect concepts to group objects that share a common property. For instance, once children have mapped one form to an object, they will extend it to other objects that share the same ontological kind (*the taxonomic constraint*, Markman & Hutchinson, 1984), or the same shape (*the shape bias*, Landau et al., 1988). Yet, it remains open to question how children react to cases where concept convexity is challenged (see however Plunkett et al., 2008), i.e. when there is evidence that the extension of a word form is not convex: do they postulate homophony to maintain concept convexity (an important assumption to form concepts in the first place), do they prioritize sticking to a one-to-one mapping assumption between word form and meaning (an important assumption to learn words), or do they simply ignore the presence of two smaller convex clusters of exemplars and save their prior assumptions about both word forms and concepts by simply postulating a broad meaning for the word in question?

Relevantly, it has been experimentally demonstrated that 3- to 9-year-old children have difficulty in learning homophones (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997). In particular they find it more difficult to learn a second meaning for a word they know (thus a homophone, e.g., learning that the familiar word form "door" also labels an unfamiliar object) than to learn a completely novel word (e.g., learning that "blick" labels an unfamiliar object). This suggests that children are slower to learn secondary meanings of homophones than to learn novel words, consistent with the idea that children prefer to preserve a one-to-one mapping between forms and meanings. Yet another possibility is that homophone learning is difficult in these conditions because the homophones are chosen such that one meaning is already available to children and thus interferes with their ability to learn a novel meaning for the same word form. Indeed even when both meanings of a pair of homophones are known to children before the experiment, they find it difficult to retrieve the less frequent meaning of the pair when the most frequent meaning is activated, but have less trouble to do so when the task provides greater contextual support for the less frequent interpretation (Beveridge & Marsh, 1991; Campbell & Bowe, 1977; Rabagliati et al., 2013). That is, in experiments where a second meaning novel meaning is taught for a known word form, the competition between the known first meaning and the novel secondary meaning may mask children's ability to consider homophones as a possible option.

The present study addresses exactly this key problem by testing the simultaneous acquisition of two meanings for a single word form. When learning a homophone, such as "bat", learners will observe several exemplars of animal-bats and several exemplars of baseball-bats, all linked to the same word form "bat".



In such a case, if learners hypothesize that this word form applies to a single, convex concept, as most word forms do, they would never discover homophony. Instead, they could consistently postulate that “bat” refers to some superordinate, coherent category encompassing both animal bats and baseball bats, just like a word like “thing” does. However, if “bat” was indeed linked to such a broad category, it is likely that learners would have observed many things that are called “bat” but are neither animal bats nor baseball bats (a *uniform* distribution of exemplars drawn from the superordinate category of “things”) rather than having observed only animal bats and baseball bats (a *bimodal* distribution of exemplars within the superordinate category).

We thus ask whether children capitalize on the *sampling distribution of the learning exemplars* to postulate homophony. To our knowledge, this is the first study that looks at the acquisition of multiple meanings for a new word by children. In Experiment 1, we combined two tests (inspired from Srinivasan & Snedeker, 2011 and Xu & Tenenbaum, 2007) to replicate previous results with adult participants, circumscribing the situations in which homophony emerges. In Experiment 2, we exported our experimental procedure with children, showing that they also refrained from associating a label to a broad set of entities encompassing all learning exemplars when they form two distinct convex clusters. Experiment 3 provides a control with adults to discard the possibility of a superficial explanation for part of our effect. Altogether, our results suggest that children by the age of 5 use information about the sampling distribution of learning exemplars to discover whether a novel word form is associated with one or several meanings. Just like adults, children expect that words, but not word forms, refer to convex concepts and form lexical representations that follow this constraint, in essence showing early awareness that homophony is a possibility in natural languages.

### Experiment 1

The experiment consisted of two testing phases: the *extension test* and the *representation test*. The extension test was similar to Xu & Tenenbaum (2007) and Dautriche & Chemla (*submitted*): participants were taught novel labels from an alien language (e.g., “blicket”) for animal categories and were asked to extend this label to test items. We manipulated whether the set of exemplars they observed formed either a uniform or a bimodal distribution of the minimal superordinate category encompassing all the exemplars. If participants are sensitive to this sampling information, we predicted that they should be less likely to extend the label to all objects that are in the superordinate category when the exemplars form a bimodal distribution compared to when they form a uniform distribution.

During the representation test, we tested whether participants represented the meaning(s) of the word they have just been taught as two separate lexical entries (i.e. homophony) or as a single lexical entry. The procedure was similar to Srinivasan & Snedeker (2011): participants were taught that a subset of the examples previously shown were wrongly labelled and are in fact labelled by *another* word in that language (*the corrected label*, e.g., “these are not blickets, these are feps”). When the exemplars formed a bimodal distribution (as in the case of homophones, e.g., two animal-bats and two baseball-bats), the corrected label corresponded to one of the meanings of the initial word (e.g., “fep” labelled the two animal-bats). When the exemplars formed a uniform distribution (e.g., one animal-bat, one tree, one car, one-baseball bat), the corrected label applied to two of the exemplars (e.g., “fep” labelled one animal-bat and one car). Participants were then tested on their extension of the corrected label. If a bimodal distribution of exemplars is sufficient to trigger homophony, and assuming that participants can rely on the independence of the two meanings of a homophone, they should restrict the corrected label to the subcategory for which they have

evidence (e.g., “fep” refers to animal-bats) and not extend it to the broader category (e.g., exclude baseball-bats). On the opposite, if participants readily extend the corrected label to the unattested meaning (or at least as much as in the uniform condition), this would suggest that they interpreted the initial word as referring to a broad category, and not as a homophone.

## Method

### Participants

Nineteen adults were recruited from Amazon Mechanical Turk (6 Females;  $M = 37$  years; all native speakers of English) and were compensated \$0.4 for their participation. One additional participant was excluded because he did not provide any answer.

### Procedure and display

Adults were tested online. They saw the pictures of two aliens, one blue and one red, both coming from the same planet. They were instructed that they would be exposed to words from their language and would have to select images that correspond to those words. In the instructions, they saw an example of a trial with the pictures and the label used for the training trial.

The trials followed the time course schematically represented in Figure 1. In the learning phase, 4 learning exemplars were displayed as a combination of a picture and a prompt underneath each of them (e.g., “This is a blicket”), allegedly pronounced by the blue alien who was pictured at the bottom of the screen.

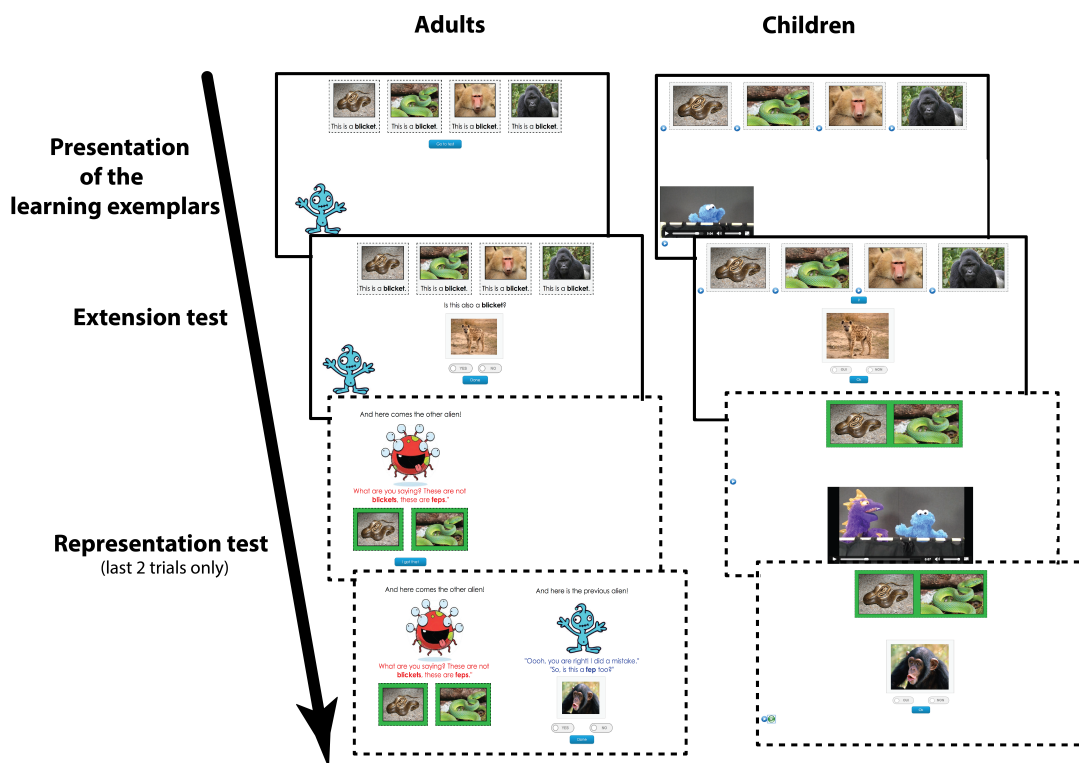
The *extension test* started as soon as adults pressed a “Go to test” button placed below the exemplars. Participants were presented with 12 test pictures displayed one-by-one below the 4 learning exemplars and asked whether this test item could be labelled by the novel word (e.g., “Is this also a blicket?”). Participants could answer by clicking on a “yes” or “no” button on the screen. When the response was “yes”, the picture frame became green, if it was “no”, the picture frame became red. Once the response was validated by participants by pressing a “Done” button, the test continued to the next test picture.

In the last two test trials, the extension test was followed by a *representation test*. On the left side of the screen, participants saw 2 of the 4 learning exemplars, highlighted in a green frame, with the red alien appearing with the prompt “What are you saying these are not blickets, these are feps!”<sup>1</sup> Once participants pressed the “I got that” button, the blue alien re-appeared on the right side of the screen recognizing his mistake and asking whether the corrected label could apply to 3 novel test pictures presented one-by-one (e.g., “Oooh you are right! I made a mistake, Is this a fep too?”). Participants validated their answer by pressing a “Done” button before moving to the next test picture.

At the end of the experiment, there was a final questionnaire asking participants about their age, native language and country.

---

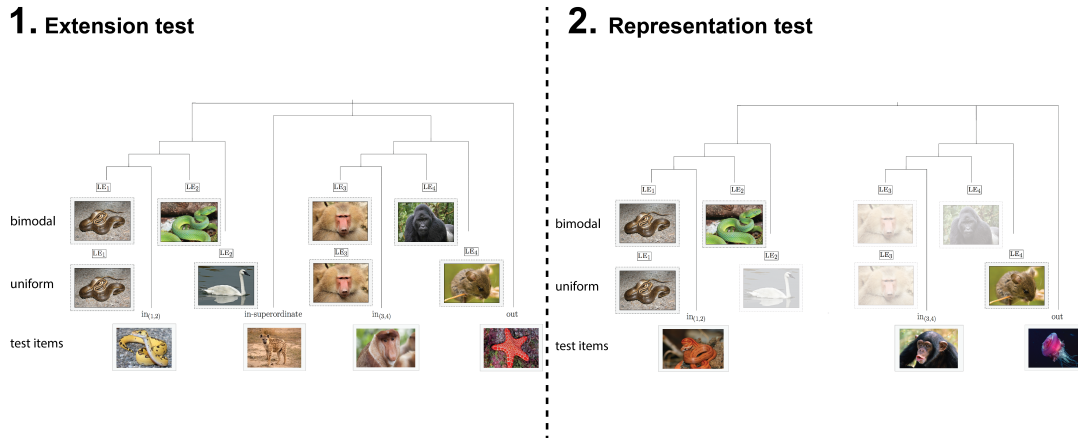
<sup>1</sup>We chose to rename two of the learning exemplars instead of one following a pilot study conducted with adults. When a single exemplar was re-labelled, adults were confused whether the corrected label applied only to this specific instance or to the whole category of the initial label.



**Figure 1: Time course of a trial in experiments 1 and 2.** Adults were tested online while children were tested in their preschool. The mode of presentation differed between the two age groups: adults learned words through written prompts while children saw videos. The time course was the same for both groups. Participants first saw the 4 learning exemplars for the novel word presented by one of the aliens. In the extension test: participants then saw 12 test pictures presented one-by-one and were asked whether the image corresponds to the word just learned. In the last two trials, the extension test was followed by a representation test: first participants saw two of the learning exemplars being renamed with another label (the corrected label) by a second alien. Then participants see another set of 3 test pictures presented one-by-one and are asked whether each of them can also be named with the corrected label.

## Conditions

Each participant saw 1 training trial and 4 test trials: 2 *uniform* and 2 *bimodal* trials. For a simple and schematic explanation, we refer the reader to Figure 2 which represents the structure of test trials and introduces visually the associated terminology.



**Figure 2: Schema of the structure of a trial in conceptual space for the extension test (1) and the representation test (2).** The first row of pictures corresponds to the configuration of the learning exemplars ( $LE_1$ ,  $LE_2$ ,  $LE_3$ ,  $LE_4$ ) in the bimodal condition and the second row to the configuration of the learning exemplars in the uniform condition. The third row corresponds to the test items.

**Training trial.** The training trial was the same for all participants: the 4 learning exemplars were the pictures of 4 animals (a dog, a goat, a pig and a cow) and the 4 test items were two animals (a cat, a horse) and two plants (a tree, a pumpkin). It was designed so that participants understand readily the task, inviting them to extend the novel label to the two animals but not to the two plants.

**Test trial.** The key factor differentiating the two test conditions (*uniform* and *bimodal*) concerns the distribution of the learning exemplars ( $LE_1$ ,  $LE_2$ ,  $LE_3$ ,  $LE_4$ ) in conceptual space (here a tree-structure).

- In the *uniform trials*, the learning exemplars formed a uniform distribution sampled from a superordinate category such that all learning exemplars are about the same distance from one another.
- In the *bimodal trials*, the learning exemplars formed a bimodal distribution sampled from two independent subcategories belonging to the superordinate category such that they formed two clusters of exemplars: ( $LE_1$ ,  $LE_2$ ) and ( $LE_3$ ,  $LE_4$ ).

During the extension test (Figure 2.1), the 12 test items were either:

- *out*: out of the superordinate category formed by the 4 exemplars (4 items).
- *in*: in one of the two subcategories (2 items in between  $LE_1$  and  $LE_2$  and 2 in between  $LE_3$  and  $LE_4$ )
- *in-superordinate*: in the superordinate category but not in any subcategory (4 items)

During the representation test (Figure 2.2), another label (the corrected label) applied to 2 of the learning exemplars:  $LE_1$  and  $LE_4$  in the uniform trials and  $LE_1$  and  $LE_2$  in the bimodal trials. The 3 test items were either:

- *out*: out of the superordinate category formed by the 4 exemplars of the initial word.

- $in_{(1,2)}$ : in the subcategory formed by  $LE_1$  and  $LE_2$ .
- $in_{(3,4)}$ : in the subcategory formed by  $LE_3$  and  $LE_4$ .

### Materials

Our stimuli relied on a set of to-be-learned labels and taxonomically organized objects.

**Labels.** We chose 7 phonotactically legal non-words of English that were not repeated across trials: 5 for the initial labels *blicket*, *smirk*, *zorg*, *moop*, *tupa* and 2 for the corrected labels *kaki*, *fep*.

**Objects in conceptual space.** Participants were tested on a set of 100 animals organized into a taxonomic hierarchy extracted from NCBI (<http://www.ncbi.nlm.nih.gov>) to obtain an objective measure of similarity between the different items as in (Dautriche & Chemla, *submitted*).<sup>2</sup> For each item, we selected 3 color photographs showing the animal in its natural background.

### Presentation and trial generation.

The order of the trials as well as the pairing between the labels and the set of learning exemplars was fully randomized and differed for each participant. We created 2 lists of trials, such that each uniform trial had a corresponding bimodal trial in the other list. That is  $LE_1$ ,  $LE_3$  and the test items were common between a pair made of a uniform and a bimodal trial, while  $LE_2$ ,  $LE_4$  varied to make the trial uniform or bimodal. All trials were generated automatically following the algorithmic constraints described in the supplemental material and selected following pilot data on adults (the full list of trials is available in the supplemental material). Participants were randomly assigned to one of the two lists of trials.

### Data analysis

Analyses were conducted using the `lme4` package (Bates et al., 2014) of R (R Core Team, 2013). In a mixed logit regression (Jaeger, 2008), we modeled the selection of a test item (coded as 0 or 1) independently for the extension test and the representation test. The extension test model included two categorical predictors with their interaction: Test Item (*out*, *in*, *in-superordinate*) and Sampling Condition (uniform vs. bimodal) as well as a random intercept and random slopes for both Test Item and Sampling Condition and their interaction for participants and trial pairs.

The representation test model included as well two categorical predictors with their interaction: Test Item (*out*,  $in_{(1,2)}$ ,  $in_{(3,4)}$ ) and Sampling Condition (uniform vs. bimodal) with a random intercept and random slopes for Sampling Condition for participants.<sup>3</sup> As can be seen in Figure 4, participants were at ceiling in selecting  $in_{(1,2)}$  in the bimodal condition. This impacted the log-estimation behind the logit model for the representation test, such that the estimates and the standard errors calculated by the model

---

<sup>2</sup>Species for which there is a clear difference between scientific and subjective popular taxonomy were excluded (e.g., sea mammals: counter-intuitively, dolphins are closer to elephants than to sharks).

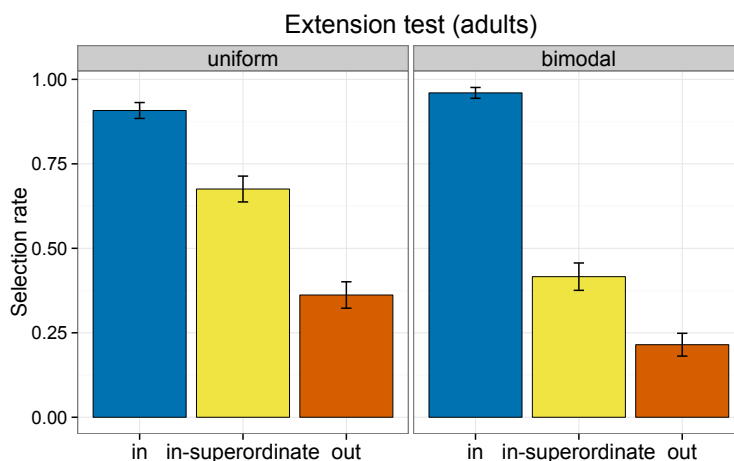
<sup>3</sup>There was no random slopes for Test Items because it leads to implausible estimates and standard error even after correction (see thereafter). There was also no random effect for trial pairs since the two last trials could be different for each participant.

were implausibly big. To get rid of these ceiling effects in a highly conservative way we introduced random noise: we run the same analysis on a modified dataset where we changed randomly 10% of the responses given for  $in_{(1,2)}$  in the bimodal condition (2 responses).

## Results

### Extension test

Figure 3 reports the average proportion of selection of each test item by sampling condition (uniform vs. bimodal) during the extension test.



**Figure 3:** Proportion of choice of each test item during the extension test averaged for each trial condition (uniform vs. bimodal) for adults in Experiment 1. Error bars indicate standard errors of the mean.

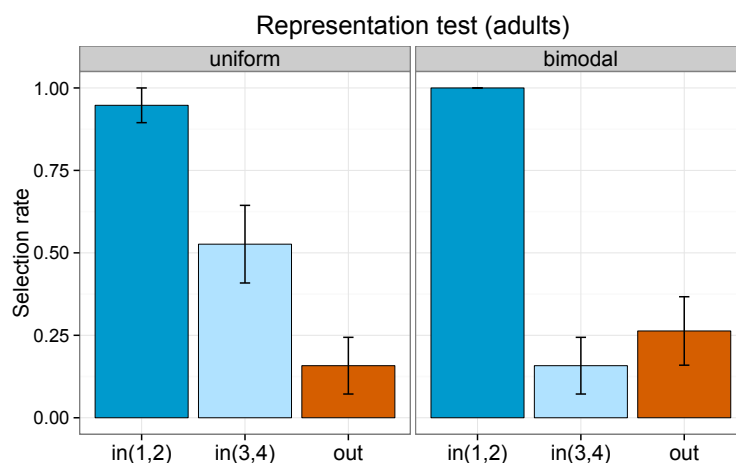
Participants were sensitive to the distribution of the learning exemplars: they selected more *in-superordinate* items in the uniform condition than in the bimodal condition ( $M_{uniform} = 0.68$ ,  $SE = 0.04$ ;  $M_{bimodal} = 0.45$ ,  $SE = 0.04$ ;  $\beta = -3.38$ ,  $z = -2.48$ ,  $p < .05$ ). We note that the distribution of the learning exemplars also affected the selection rate of *out* items: they selected more *out* items in the uniform compared to the bimodal condition ( $\beta = 1.79$ ,  $z = 2.03$ ,  $p < .05$ ). Yet, this is expected following the size principle documented by Xu and Tenenbaum (2007): the boundaries of the superordinate category defined by the 4 learning exemplars in the uniform condition are less sharp (for an equal number of exemplars) than the boundaries of the subcategories in the bimodal condition. As a result, there is more uncertainty about the correct level of generalization in the tree-structured hierarchy leading to a slightly higher selection rate of *out* items in the uniform condition.

However the sampling distribution of the exemplars modulated participants' responses beyond the size principle effect since we observe two interaction effects: the difference between the selection rate of *in-superordinate* items and *in* items was greater in the bimodal than in the uniform condition ( $\beta = 5.70$ ,  $z = 3.23$ ,  $p < .01$ ); similarly the difference between the selection rate of *in-superordinate* items and *out* items

was marginally smaller in the bimodal than in the uniform condition ( $\beta = -1.58, z = -1.67, p < .1$ ). This suggests that participants were more inclined to extend the label to all objects in the superordinate category including all the exemplars when the exemplars formed a uniform distribution than when they formed a bimodal distribution.

## Representation test

Figure 4 reports the average proportion of selection of each test item by trial condition (uniform vs. bimodal) during the representation test.



**Figure 4:** Proportion of choice of each test item during the representation test averaged for each trial condition (uniform vs. bimodal) for adults in Experiment 1. Error bars indicate standard errors of the mean.

Crucially, the distribution of the exemplars influenced participants' representation of the initial word: in the bimodal condition, participants were less likely to extend the corrected label to unattested items ( $in_{(3,4)}$ ) than in the uniform condition ( $M_{uniform} = 0.53, SE = 0.12; M_{bimodal} = 0.16, SE = 0.08; \beta = -5.05, z = -3, p < .01$ ). This is compatible with a homophonous representation of the initial word in the bimodal condition, where only one of the two meanings ( $in_{(3,4)}$ ) has been affected by the later correction.

## Discussion

The sampling distribution of the exemplars modulated adults' interpretation of a novel word: when the exemplars formed a uniform distribution, participants were more likely to extend its label to all objects falling in the superordinate category containing all the exemplars than when the exemplars formed a bimodal distribution (extension test).

Yet, one may worry that participants did not form a lexical representation but rather extended the label based on similarity to the learning exemplars: they selected more *in-superordinate* items in the uniform than in the bimodal condition simply because, on average, *in-superordinate* items may be closer to the

learning exemplars in the uniform than in the bimodal condition. Yet that's not the case: recall that in our similarity space (Figure 2), the common ancestor of *in-superordinate* items with any pair of learning exemplars, i.e. (LE<sub>1</sub>, LE<sub>2</sub>) or (LE<sub>3</sub>, LE<sub>4</sub>), is the same in the uniform and the bimodal condition. Certainly participants' responses were in part guided by similarity of the test items to the learning exemplars: in the bimodal condition during the extension test, adults selected *in-superordinate* items at a higher rate than *out* items ( $\beta = -2.30, z = -1.99, p < .05$ ). Yet it is sufficient for our purpose to note that this does not account for the entirety of our effect.<sup>4</sup>

An interesting question is whether observing the exemplars in a bimodal distribution was sufficient for participants to form homophonous form-meaning representations. Indeed, there may be two interpretations of the results of the extension test:

1. Participants formed words' representations that respect concept convexity. In the bimodal condition, participants postulated homophony: they associated the novel word with two independent meanings, each corresponding to a convex concept (e.g., PRIMATE and SNAKE).
2. Participants accepted that a word's meaning could be a set of disconnected concepts: in the bimodal condition they associated the novel word to a single, disjoint concept (e.g., PRIMATE OR SNAKE).

The results of the representation test favor the first possibility. When presented with a bimodal distribution of exemplars (e.g., 2 primates and 2 snakes) labelled by a single word "blicket", participants interpreted the corrected label based on its taught meaning alone (e.g., the 2 snakes but not the 2 primates) suggesting that they preferentially understood "blicket" as a word form associated with two homophonic words, rather than as a single word with a single discontinuous meaning.

All in all, we replicate previous results (Dautriche & Chemla, *submitted*) showing that the distribution of learning exemplars interacts with constraints on concept convexity to form different form-meaning representations. When the exemplars form a uniform distribution, participants are more likely to associate the word to a single convex meaning that encompasses all the learning exemplars (and every entity in between them). Yet, when the exemplars form a bimodal distribution, participants prefer to postulate homophony such that the novel word is associated to two convex meanings, rather than to a single, broad convex meaning or to a single discontinuous meaning. The critical question then, is, do children also postulate homophony when there is evidence that the exemplars of a word are distributed in two convex clusters in conceptual space? Experiment 2 investigated this question by adapting the design of Experiment 1 to French preschoolers.

<sup>4</sup>Note that similarity to the exemplars also played a role in the representation test, in which participants selected more  $in_{(1,2)}$  items than  $in_{(3,4)}$  items in the uniform condition ( $\beta = -4.01, z = -2.61, p < .01$ ) simply because in the tree structure for uniform trials (Figure 2),  $in_{(1,2)}$  is closer to LE<sub>1</sub> than  $in_{(3,4)}$  to LE<sub>4</sub>, thus more likely to be selected as an instance of the corrected label.



## Experiment 2

### Method

#### Participants

Twenty-one 5-year-old monolingual French speaking children (5;1 to 6;1,  $M_{age} = 5;6$ , 10 girls) were tested in a public preschool in Paris. Their parents signed an informed consent form. Three additional children were tested but not included in the analysis because they systematically responded *yes* ( $n = 1$ ) or *no* ( $n = 2$ ) without even looking at the test pictures or responding before they appeared on the screen.

#### Procedure and display

The experiment was identical to Experiment 1 except that we used videos instead of written prompts with children (see Figure 1).

Children were tested individually in a quiet room in their preschool. During the experiment, children sat next to the experimenter, in front of a computer and wore headphones to listen to the stimuli. Before the experiment began, children watched a video where two alien puppets introduced them with the task. The two puppets presented themselves as Boba and Zap, and told the children that they were coming from another planet where they speak a different language, so they would teach the children words of their language. Once a child demonstrated to the experimenter that (s)he understood the task, the experiment started. In each trial, children saw 4 learning exemplars, presented one-by-one as the combination of a picture and a video of Boba labeling the picture with a non-word “Ça, on appelle ça une bamoule!” *This, we call it a bamoule*. Each learning exemplar was displayed on the screen when the experimenter clicked on a button. Once children saw all 4 learning exemplars, they saw a last video where Boba asked them to repeat the word. This was to ensure that children were on task and for the experimenter to know which word was used by the puppet (as the words were randomly assigned to a set of learning exemplars and the experimenter could not hear the stimuli).

During the *extension test*, children were presented with the 12 test pictures displayed one-by-one below the 4 learning exemplars. For each of them, the experimenter asked: “Est ce que tu penses que ça s’appelle une bamoule?” *Do you think it is called a bamoule?* When the child answered, the experimenter clicked accordingly on the “yes” or “no” button on the screen. Children could change their mind if they wanted within a few seconds after their answer or longer if the experimenter saw that they were still hesitating. Once a response was validated by the experimenter, the test continued to the next test picture.

In the last two test trials, the extension test was followed by a *representation test*. Children saw 2 of the 4 learning exemplars grouped in a frame together with a video where the second puppet, Zap, scolded the first one, Boba, for using the wrong word for the two exemplars displayed “C’est pas des bamoules ça! C’est des torbas!” *These are not bamoules! These are torbas* (the whole script for this video can be found in the supplemental material). Boba then, acknowledged his mistake. During the dialogue, the frame of the pictures was blinking in green. At the end of the video, the experimenter asked the child what happened and replayed the video if the child did not understand the video or did not remember the novel word. Children were then tested whether the corrected label could apply to 3 novel test pictures. Each of the test picture

was displayed below the two learning exemplars and the child was asked by the experimenter “Est ce que tu penses que ça, ça s’appelle un torba?” *Do you think that this is called a torba?*

At the end of the experiment, there was a final video where the two puppets said good-bye to the child. The whole Experiment lasted about 15 min. All sessions were audiotaped.

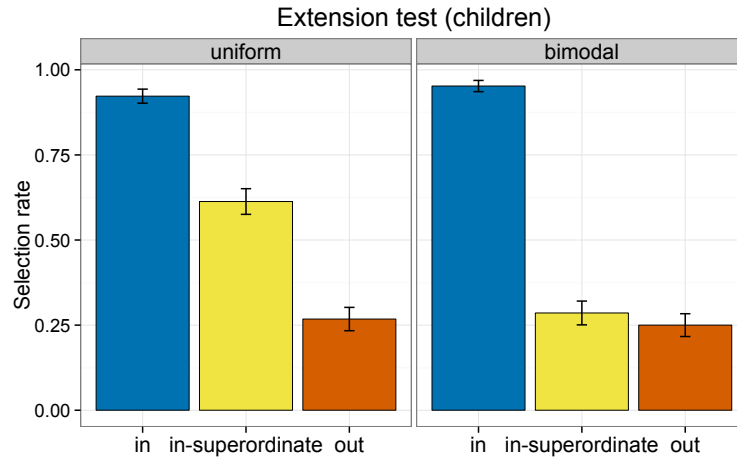
### **Conditions, Materials, Presentation and trial generation, Data analysis.**

Everything was similar to Experiment 1, except that the experiment was in French and hence the set of non-words consisted of 9 phonotactically legal non-words of French: *bamoule* was always used in the training trial. From the remaining 8 non-words that were used in the test trials, half of them were bisyllabic and were used as the initial label (*toupa, fimo, lagui, yoshi*), the other half were trisyllabic and were used as the corrected labels (*midori, cramoucho, didolu, baboocha*). The difference in the number of syllables was introduced to make it easier for children to distinguish between the initial and the corrected labels.

## **Results**

### **Extension test**

As shown in Figure 5, children and adults behaved in the same way: children selected more *in-superordinate* items in the uniform condition than in the bimodal condition ( $M_{uniform} = 0.61, SE = 0.04$ ;  $M_{bimodal} = 0.28, SE = 0.04$ ;  $\beta = -2.56, z = -3.73, p < .001$ ). The distribution of the learning exemplars did not affect any other test items for children (all  $ps > 0.7$ ) resulting in two interaction effects: the difference between the selection rate of *in-superordinate* items and *in* items, was greater in the bimodal than in the uniform condition ( $\beta = 2.79, z = 2.30, p < .05$ ); similarly the difference between the selection rate of *in-superordinate* items and *out* items, was smaller in the bimodal than in the uniform condition ( $\beta = -3.28, z = -4, p < .001$ ). This suggests that children were more likely to extend the label to all objects in the minimal subtree containing all the exemplars in the uniform condition compared to the bimodal condition.



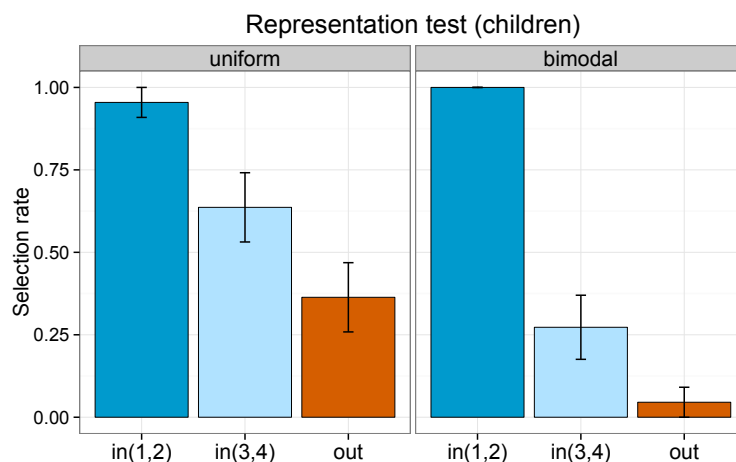
**Figure 5:** Proportion of choice of each test item during the extension test averaged for each trial condition (uniform vs. bimodal) for children. Error bars indicate standard errors of the mean.

For uniform trials, we replicate previous results (Xu & Tenenbaum, 2007): children were more likely to extend the label to all objects in the minimal subtree containing all the exemplars than to objects that are out of the subtree (*in* vs. *out*,  $\beta = -5.39, z = -6.28, p < .001$ ; *in-superordinate* vs. *out*,  $\beta = -3.01, z = -3.82, p < .001$ ). We note that the distance of the test items to the learning exemplars affected children's responses: they selected more *in* items than *in-superordinate* items ( $\beta = -2.29, z = -2.80, p < .01$ ) suggesting that children's extension of the label may be in part guided by similarity of the test items to the learning exemplars.

For bimodal trials, children were more likely to extend the label to objects that were in one of the two subcategories (*in*) than to other objects that were either in the superordinate category containing all the exemplars but out of the subcategories (*in-superordinate*,  $\beta = -5.12, z = -4.61, p < .001$ ) or out of the superordinate category (*out*,  $\beta = -4.85, z = -8.01, p < .001$ ). There was no difference between the selection rate of *in-superordinate* items and *out* items for children ( $p > 0.7$ ) suggesting that children excluded *in-superordinate* items from the extension of the word.

## Representation test

The distribution of the exemplars influenced children's representation of the initial word (see Figure 6): in the bimodal condition, children were less likely to extend the corrected label to unattested items ( $in_{(3,4)}$ ) than in the uniform condition ( $M_{uniform} = 0.64, SE = 0.10$ ;  $M_{bimodal} = 0.28, SE = 0.10$ ;  $\beta = -1.66, z = -2.51, p = .01$ ).



**Figure 6:** Proportion of choice of each test item during the representation test averaged for each trial condition (uniform vs. bimodal). Error bars indicate standard errors of the mean.

While children and adults behaved statistically the same way ( $p > 0.2$ ), there were some visible differences. In particular, there was a main effect of Sampling Condition for children: children were more likely to select test items in the uniform than in the bimodal condition ( $\chi(1) = 8.20, p < .01$ ). They even selected more *out* items in the uniform than in the bimodal condition ( $\beta = 2.56, z = 2.20, p < .05$ ; no such difference was observed for adults,  $p > 0.2$ ). The selection rate of *out* items in the uniform condition is similar to their selection rate during the extension test. One may wonder why children selected it even less in the bimodal condition during the representation test. Children may be driven by a “size principle” (Xu & Tenenbaum, 2007): what determines whether the hypothesized extension of a label will have sharp boundaries is the number of consistent hypotheses with the exemplars. In the representation test, the category defined by the two exemplars is rather narrow (e.g., snakes) ensuring that very few meaning hypotheses are possible for the corrected word. During the extension test, there were more meaning hypotheses possible for the initial word: despite the fact that two exemplars were presented for each of the two subcategories (e.g., INSECTS and PRIMATES), the possibility that the initial word corresponded to the minimal superordinate category encompassing all the exemplars could still be entertained leading to a bigger uncertainty about the boundaries of the categories of the initial word compared to the boundaries of the corrected word.

## Discussion

These results suggest that children, just like adults, postulate homophony for a word when the learning exemplars formed a bimodal distribution, i.e. that the meaning of that word form is best represented as two independent convex clusters rather than a big cluster encompassing all the exemplars.

At this point, we would like to point out a weakness in the second of our tests, the representation test. Because the category including the re-labelled exemplars is wider in the uniform case (e.g., a snake and a mouse) compared to the bimodal case (e.g., two different kinds of snakes), this may be sufficient for children to extend the corrected label to more test items in the uniform compared to the bimodal condition: indeed, a chimp is more likely to be a “fep” when “fep” labels a snake and a mouse than when “fep” labels

two snakes. This could happen independently of what children had learned about the meaning of the initial word “blicket”. Thus, although our results reflect exactly what we could expect to observe if children had hypothesized that “blicket” was a homophone when observing a bimodal distribution of its exemplars, we cannot entirely rule out the possibility that children, and adults, responded to the representation test independently from the extension test, in which case the representation test would tell us nothing about what had been learned initially. To test whether participants responded to the representation test independently from the extension test, i.e., whether participants’ extension of the corrected label was independent from their representation of the initial word, we ran a control experiment with adults, in which we tested both participants’ representation of the initial word not only by testing their extension of the corrected label (“fep”) but also by testing their *updated* extension of the initial label (“blicket”).

### Experiment 3

In Experiment 3, adults learned a novel word, e.g., “blicket” (the initial label), from a set of exemplars, and were instructed that some of the exemplars were not “blickets” but “feps” (the corrected label).

When “blicket” applies to a bimodal distribution of exemplars (e.g., two animal-bats and two baseball-bats) and “fep” corresponds to one of the meanings of the initial word (e.g., the two animal-bats), if participants recruit what they have learned from “blicket”, they should restrict not only the corrected label to the subcategory for which they have evidence (e.g., “fep” refers to animal-bats and not to baseball-bats) as in Experiment 1, but they should also update their representation of “blicket” (e.g., “blicket” now refers to baseball-bats and not to animal-bats). On the opposite, if they readily extend “blicket” to animal-bat items, although these items are now labelled as “feps”, then this will suggest that their responses for “fep” and “blicket” are independent. When “blicket” applies to a uniform distribution of exemplars (e.g., one animal-bat, one tree, one car, one-baseball bat) and the corrected label applies to two of the exemplars (e.g., “fep” labelled one animal-bat and one car), if participants recruit their lexical representation of “blicket” during the representation test, they should, as in Experiment 1, be more willing to extend “fep” to all “blickets” (e.g., the broader category of animal-bat, tree, car and baseball bat) but crucially they should also update their lexical representation for “blicket” (e.g., excluding now the broader category of animal-bat, tree, car and baseball bat from it). On the contrary if participants’ lexical representation of “blicket” is unaltered, this would suggest that participants have independent lexical representations for “fep” and “blicket”.

In sum, for both conditions if participants’ extension of the corrected label, “fep”, is uninformed by what they learned about “blickets”, then their representation of “blicket” should be untouched. On the other hand, if participants recruited their lexical representation of “blicket” during the representation test, then they should update this lexical representation by excluding all the “feps” from it.

## Method

### Participants

Twenty adults were recruited from Amazon Mechanical Turk (8 Females;  $M = 33$  years; 19 native speakers of English) and were compensated \$0.4 for their participation.

## Procedure and display

The experiment was identical to the adult version in Experiment 1 except that during the representation test we tested their representation of the corrected word (as in Experiment 1) but also their representation of the initial word.

During the representation test, as before, on the left side of the screen the red alien labelled 2 of the 4 learning exemplars by another word “What are you saying these are not blickets, these are feps!”. On the right side of the screen, the blue alien seen during the extension text appeared with a test item and two prompts 1) “Oooh you are right! I made a mistake. Is this a fep too?” (representation test for the corrected word) and below it 2) “Is this a blicket?” (representation test for the initial word). There was a “yes” or “no” button below each question to record participants’ answer.

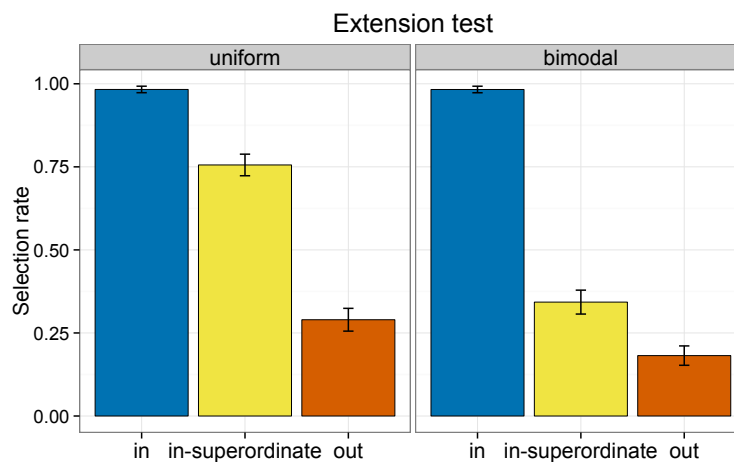
## Conditions, Materials, Presentation and trial generation, Data analysis.

Similar to Experiment 1.

## Results

### Extension test

Figure 7 reports the average proportion of selection of each test item by trial condition (uniform vs. bimodal) during the extension test.



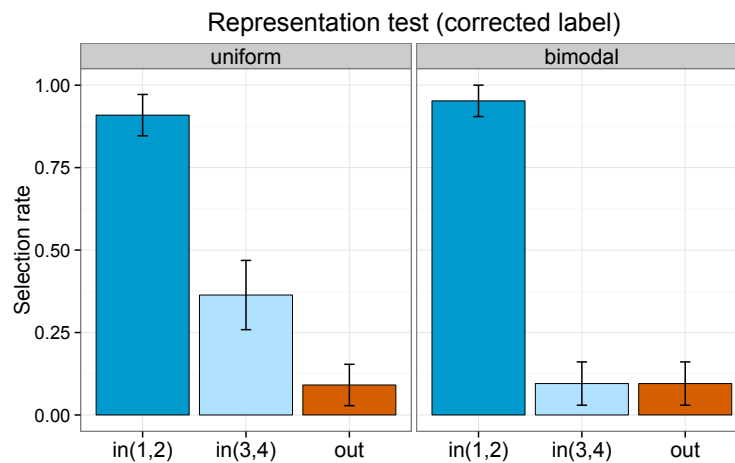
**Figure 7:** Proportion of choice of each test item during the extension test averaged for each trial condition (uniform vs. bimodal) in Experiment 2. Error bars indicate standard errors of the mean.

The extension test replicated the results of Experiment 1: participants’ responses were modulated by the distribution of the learning exemplars: participants chose more *in-superordinate* items in the uniform than in the bimodal condition ( $M_{uniform} = 0.75$ ,  $SE = 0.04$ ;  $M_{bimodal} = 0.34$ ,  $SE = 0.03$ ;  $\beta = -6.31$ ,  $z =$

$-3.57, p < .001$ ). The sampling distribution did not affect the choice of any other test items ( $ps > 0.1$ ). As a result, the difference between the selection rate of *in-superordinate* and *in* items was greater in the uniform than in the bimodal condition ( $\beta = 3.68, z = 1.98, p < .05$ ) and the difference between the selection rate of *in-superordinate* and *out* items was smaller in the bimodal condition compared to the uniform condition ( $\beta = -5.33, z = -2.51, p < .05$ ).

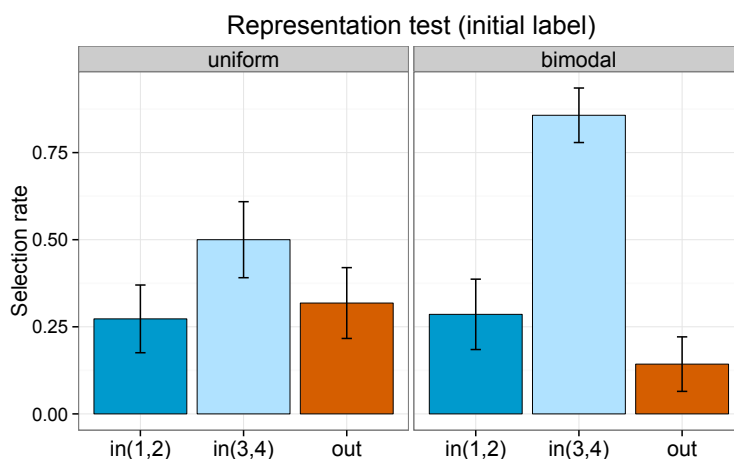
## Representation test

Figure 8 reports the average proportion of selection of each test item by trial condition (uniform vs. bimodal) during the representation test for the corrected label. This test also replicated the results of Experiment 1: in the bimodal condition, participants were less likely to extend the corrected label to unattested items ( $in_{(3,4)}$ ) than in the uniform condition ( $M_{uniform} = 0.36, SE = 0.10; M_{bimodal} = 0.1, SE = 0.06; \beta = -2.26, z = -2.19, p < .05$ ). The sampling distribution of the exemplars did not affect the responses to other test items ( $ps > 0.5$ )



**Figure 8:** Proportion of choice of each test item during the representation test averaged for each trial condition (uniform vs. bimodal) when participants were tested on the extension of the corrected label. Error bars indicate standard errors of the mean.

As shown in Figure 9, when tested on the initial label, participants selected more  $in_{(3,4)}$  in the bimodal than in the uniform condition ( $M_{uniform} = 0.5, SE = 0.11; M_{bimodal} = 0.86, SE = 0.08; \beta = 2.28, z = 2.70, p < .01$ ). This suggests that the sampling distribution of the exemplars affected participants' representation of both the corrected and the initial words. The sampling distribution did not affect the response of any other test items ( $ps > 0.1$ ).



**Figure 9:** Proportion of choice of each test item during the representation test averaged for each trial condition (uniform vs. bimodal) when participants were tested on the initial label. Error bars indicate standard errors of the mean.

In the uniform condition, participants refrained from applying the initial label to any of the test items that were within the superordinate category formed by the 4 exemplars ( $in_{(1,2)}$  and  $in_{(3,4)}$ ). This was evidenced by the fact that the selection rate of  $in_{(1,2)}$  and  $in_{(3,4)}$  as representatives of “blickets” was not different from their choice of the *out* items ( $ps > 0.15$ ). This suggests that the corrected word overrode the representation of the initial word: all items previously labelled as “blickets” are now “feps”, thus “blicket” is not associated with any meaning.

In the bimodal condition, participants refrained from applying the initial label to test items that belonged to the subcategory of items that had been relabelled with the corrected label: they did not choose  $in_{(1,2)}$  as representatives of the initial label more often than *out* items ( $p > 0.2$ ); however, they readily extended the initial label to test items that were outside the subcategory of relabelled items ( $in_{(3,4)}$  vs. *out*:  $\beta = -4.52, z = -4.46, p < .001$ ). In other words, when exposed to two primates and two snakes all labeled “blicket”, if the two primates are relabelled “feps”, participants still consider that the new instances of snakes are valid instances of “blicket”.

## Discussion

When the initial word was corrected by another word (e.g., “these are not blickets, these are feps”), participants took into account what they have learned about the initial word (“blicket”) to comprehend the extension of the corrected word. We replicated the results from Experiment 1: in the uniform condition, participants were more likely to associate the corrected word to the superordinate category that spans all 4 learning exemplars than in the bimodal condition. In addition, participants also updated their representation of the initial word. Specifically, in the uniform condition, participants refrained to associate the initial label to all items that were previously labelled by it (as all are now in the extension of the corrected word, “fep”). Yet, when the exemplars formed a bimodal distribution, i.e. were taken from two clusters of exemplars  $C_1$  and  $C_2$  such that the corrected label applied only to one of the clusters of exemplars  $C_1$ , participants still considered test items belonging to  $C_2$  as valid instances of the initial label. This suggests that adults in



Experiment 1 recruited the representation of the initial word during the representation test.

However, the alternative explanation, i.e., that the representation of the corrected label “fep” is computed on the basis of the learning exemplars for “fep” only (independently of the representation of the initial label “blicket”), cannot be entirely ruled out. Participants could have extended the corrected label “fep” according to the distribution of its exemplars in the representation test (i.e., extend it to more test items in the uniform condition than in the bimodal condition simply because the category covered by the two exemplars in the representation test is wider in the uniform case) and then used this information to decide how “blicket” should be extended. Clearly if participants decide that a given item *I* is an instance of “fep”, they will be more willing to exclude it from the extension of “blicket” and this independently of how they decided that *I* is a “fep”. However, it should be noted that *out* items are excluded both from the extension of “fep” and the extension of “blicket”, suggesting that when responding to “blicket” in the representation test, participants are still influenced by their responses for “blicket” during the extension test (where they also excluded *out* items from the extension of the word). So this would suggest that, while participants may not recruit their initial representation of “blicket” to extend “fep” to test items in the representation test, they still recruit it (together with what they have learned about “fep”) to find the extension of “blicket” in the representation test. While we cannot entirely dismiss the possibility that participants would selectively attend to their initial representation of “blicket” in the representation test when responding to “blicket” but not to “fep”, our data provide little support for this hypothesis.

We conclude thus that the most likely interpretation of our results in Experiment 1 is that adults recruited the representation of the initial word during the representation test. Adults preferred, in the first phase, to postulate that the novel word carries homophony when it is learned from exemplars in a bimodal distribution. Given the similarity of the results between Experiments 1 and 2, one would be tempted to extend this interpretation to the children results from Experiment 2. At this point, however, a note of caution is necessary. One limitation of the present data is that we used our conclusion with adults to rule out a possible confound for Experiment 1 and 2 with adults and children. Yet we cannot exclude the possibility that this very explanation may still underly children’s response pattern. We leave it for future research to establish a more direct argument to explain both children’s and adults’ performance in these experiments.

## General Discussion

Children are sensitive to the sampling distribution of the learning exemplars when learning words (as in Xu & Tenenbaum, 2007). Yet, we demonstrate that this interacts with the kind of form-meaning representation children are ready to entertain. Observing a bimodal distribution of learning exemplars for a novel word indicated to our participants that the word was likely to have several meanings. Importantly, our results suggest that these meanings were stored *separately*, suggesting that children’s representation of the novel word in these conditions is very much similar to homophony (Srinivasan & Snedeker, 2011). This extends previous results from adults (Dautriche & Chemla, *submitted*) and suggests that when observing a bimodal distribution of exemplars for the same word form, children generate form-meaning representations, such as homophony, that respect concept convexity.

Current word learning accounts have documented and modeled paradigmatic cases of word acquisition, where a single form is associated with a single meaning. We pursued that enterprise by showing that less standard situations, such as homophony, can help highlight the key role of factors such as the sampling

distribution of exemplars. It also helps better understand the priors that may constrain and guide word acquisition, as we detail below.

### Missing factors in word learning accounts

Previous studies have shown that children are sensitive to sampling principles when learning words. Xu & Tenenbaum (2007) describe a *size principle*: children’s confidence in the boundary of the set of entities associated with a word increases as they observe more learning exemplars, even if they are all identical. Here we showed that children were sensitive to the distribution of learning exemplars in conceptual space, another statistical principle presumably following from the assumption that the exemplars of a word are sampled *randomly* from the underlying category (c.f. Xu & Tenenbaum, 2007). Intuitively, in the case of homophones, we expect the label to occur with a set of exemplars  $S$  drawn from two distinct subcategories  $X_1$  and  $X_2$ . Thus,  $S$  would take the form of a *bimodal* distribution within the single superordinate category  $X$  encompassing the two subcategories (i.e.  $X$  is the minimal well-formed category such that  $X_1 \cup X_2 \subset X$ ). Yet if the word were to be associated with the whole  $X$ , we would expect the exemplars that are associated with the label (i.e.,  $S$ ) to be *uniformly* distributed within  $X$ . Our results suggest that children can use such sampling considerations to decide whether a word is associated with one category (standard case) or several categories (homophone), even when exposed to very few exemplars.<sup>5</sup>

There are other factors, not documented here, that may interact with the expectation that concepts are convex and could help children to identify that a word has several meanings. First, evidence for homophony may come from other words in the lexicon. Adults are less likely to extend a label (e.g., “blicket”) to an entity, even if this entity falls right between the learning exemplars for the label, when this entity also falls close to some entity labelled by another word (e.g., “fep”) (Dautriche & Chemla, *submitted*). Intuitively, the interfering label provides further evidence for the presence of two distinct clusters of exemplars in conceptual space, that are separated by another concept labelled by another word (“fep” in the example). This suggests that learners have expectations not only about how words occupy the conceptual space, but also about how they *share* the conceptual space. Similarly, children assume that word extensions are mutually exclusive (Markman & Wachtel, 1988), and may thus possibly use the presence of other words in their lexicon together with a constraint on concept convexity to discover that a word is likely to have several meanings.

Second, some linguistic constructions may be helpful to discover homophony (or the absence of homophony). For instance, children could notice that words mapping to a single meaning commonly appear in some plural sentences where homophones never appear (e.g., “These are two bats” pointing at one baseball-bat and one animal-bat). And this is so for reasons one can understand: a single phonological form cannot be used to refer to two words at the same time, even if the two words are homophonic.<sup>6</sup> Adults have been shown to use such constructions to assess homophony (Dautriche & Chemla, *submitted*), and children may also be sensitive to such linguistic evidence.

These factors may help learners to identify words with multiple meanings. Yet, they also raise immediate challenges for current word learning accounts. For instance, we assumed until now that children understood which object is referred to by the word in context. Yet, in the real world, the label is uttered

<sup>5</sup>One may argue that the low number of exemplars was not a limitation in our task because it could be compensated by pragmatic considerations: participants may expect that the learning exemplars were not drawn at random by the aliens, but rather that the aliens in our stories were trying to be informative about the meaning of the words and chose their exemplars optimally.

<sup>6</sup>Note that such a sentence is not ungrammatical but *zeugmatic* (Zwicky & Sadock, 1975).

in a complex visual environment where the true referent is likely to be confounded with other possible referents present at the same time (the *mapping problem*, Quine, 1960). Thus it is likely that the set of exemplars for a label contains *outliers*, i.e. items that are outside of the true extension of the word, because the child would have failed to narrow down the true referent of the word. As a result, the set of exemplars would certainly form a multimodal distribution (e.g., a set of banana exemplars along with a dog exemplar – that happened to eat a banana during one of the learning event). The challenge for the learner is thus to distinguish between outliers of the true meaning of the word and representative examples of a new word meaning. General principles such as the convexity of concepts and one-to-one mapping between word forms and concepts may help discard noise of this type. However, if these general principles allow exceptions, as our study of homophony reveals, they may hardly help disentangle signal (of homophony) from noise.

### Missing priors in word learning accounts

Our results suggest that children expect meanings to be convex, and are willing to postulate homophony rather than breaking this constraint (postulating a disjoint meaning) or than enforcing that convexity constraint at all cost (postulating a broad lexical entry for problematic words).

This contradicts current word learning accounts which, technically, transpose the notion of convexity from the level of concepts to the level of word forms, assuming that word forms link to concepts in a one-to-one fashion. Accordingly, none of the current accounts allow for the possibility that children can associate word forms with multiple meanings. As a matter of fact, many developmental studies have documented that preschoolers have notable difficulties in learning homophones (Casenhiser, 2005; Doherty, 2004; Mazzocco, 1997). In these studies, the encounter of the second meaning of a homophone is simulated by using familiar words (e.g., “snake”) to refer to novel referents (e.g., an unfamiliar object). Yet, we suggest that children’s failure in these studies does not reflect an excessive reliance on a one-to-one mapping between form and meaning, but rather insufficient executive skills for such a task; as the current results show, children have no problem learning homophones when they have to learn the two meanings simultaneously. Learning homophones in our study may be easier than learning a second meaning for a known word because children do not have to inhibit a highly active word representation for one of the meanings (Khanna & Boland 2010, see also Choi & Trueswell 2010; Novick, Trueswell, & Thompson-Schill 2010). This suggests that children’s difficulty in learning homophones may have been previously overrated and that endowing the learning system with a strict one-to-one form-meaning mapping constraint cannot solely account for the mechanism underlying the acquisition of homophones. Certainly it may still be possible that a one-to-one form-meaning mapping constraint is guiding word learning at earlier stages of language development, as 5-year-olds may already have learned to relax this constraint to accommodate more challenging form-meaning mappings, such as homophony. Yet, if this is the case, current word learning accounts should be able to explain how children depart from this default assumption.

The present work thus has important implications for current word learning accounts. When a label seems to apply to a disjoint set of objects, the learner has two options: 1) postulate homophony (i.e. the possibility that a word form maps onto two distinct meanings) or 2) follow a one-to-one form-meaning mapping constraint and postulate that the label is a single word that applies to a larger set of objects (the category that covers all the positive instances of the label). Most accounts predict that 2) is the default, but this has to be refined since children eventually learn homophones (e.g., it is likely that English preschoolers

know both meanings of the word form “bat”; see also de Carvalho et al. 2015 for evidence that French 3-year-olds have acquired a certain number of homophone pairs). An important open issue then is to equip the learning system with the right built-in constraints. At this point it seems that word learning is guided by a) learners’ expectations that concepts are convex, always; b) learners’ expectation that word forms are linked to one meaning, in general; and c) the possibility that a word form maps onto several distinct meanings if a) is challenged. While a) helps to constrain the possible concepts one can entertain, b) helps to constrain the number of hypotheses children need to consider when learning words (especially while used in combination with a)). Note that this would constitute very specific priors for a general learning system: it presupposes that children already know that an essential feature of their to-be-learned lexicon is to be composed of form-meaning mappings of very specific kinds. The present study contributes to point c) above: children can entertain the possibility that a word maps onto several distinct meanings to accommodate apparent violations of a) concept convexity at the detriment of b) a one-to-one mapping between forms and meanings. This suggests that children are able to selectively trigger or silence b) as a function of the learning situation (and we documented that this could be made possible by observing the sampling distribution of the learning exemplars). One possibility for current word learning accounts would be thus to implement these three built-in constraints directly in the learning system and tweak the learning inference component to accommodate the sampling effect we document. Yet this would be a rather strong assumption that amounts to saying that the learning system expects the existence of a phenomenon as specific as homophony, from the start.

Because current word learning accounts specialized into fairly simple word learning phenomena (i.e., one form associated with a single meaning), they equipped the learner with specialized built-in constraints (e.g., a one-to-one form-meaning mapping constraint) that cannot explain learning of other more complex phenomena, such as homophony. Incorporating homophony, and potentially other less trivial word learning situations, will allow these accounts to delineate the more general priors that children bring into the word learning task. Yet, although adding different built-in constraints to these accounts may be technically easy, this may be theoretically challenging as we described above, as one would need to explain how the learning system is capable of juggling between different priors to appropriately learn different types of words.

### **Conclusion**

Homophony presents a challenge to word learners: it requires children to discover that word forms and concepts are not always in a one-to-one relation, an otherwise important assumption to restrict the search space for word meanings. The present study showed that children use information about the sampling distribution of learning exemplars (and in particular the fact that they form two distinct convex clusters in conceptual space) to infer homophony for a novel label. We argue that this unexplored sensitivity and the very possibility of homophony should be incorporated into future accounts of word learning.

### **References**

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using eigen and s4* [Computer software manual]. (R package version 1.1-2)

- Beveridge, M., & Marsh, L. (1991). The influence of linguistic context on young children's understanding of homophonic words. *Journal of Child Language*, 18(02), 459–467.
- Bloom, P. (2001). Précis of How children learn the meanings of words. *Behavioral and Brain Sciences*, 24(06), 1095–1103.
- Campbell, R. N., & Bowe, T. (1977). Functional asymmetry in early language understanding. *Salzberger Beiträge für Linguistik*, 3.
- Casenhiser, D. M. (2005). Children's resistance to homonymy: an experimental study of pseudohomonyms. *Journal of Child Language*, 32(2), 319–343.
- Choi, Y., & Trueswell, J. C. (2010). Children's (in) ability to recover from garden paths in a verb-final language: Evidence for developing control in sentence processing. *Journal of Experimental Child Psychology*, 106(1), 41–61.
- Dautriche, I., & Chemla, E. (submitted). What homophones say about words.
- de Carvalho, A., Dautriche, I., & Christophe, A. (2015). Preschoolers use phrasal prosody online to constrain syntactic analysis. *Developmental science*.
- Doherty, M. J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language*, 31(1), 203–214.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578–585.
- Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
- Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446.
- Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, 63(1), 160–193.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, 3(3), 299–321.
- Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.
- Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, 16(1), 1–27.
- Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2), 121–157.
- Mazzocco, M. M. (1997). Children's interpretations of homonyms: a developmental study. *Journal of Child Language*, 24(02), 441–467.
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3), 289.

- Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, 4(10), 906–924.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, 123(2), 199–217.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. The MIT Press.
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, 106(2), 665–681.
- Quine, W. V. O. (1960). *Word and Object*. MIT Press.
- R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.
- Rabagliati, H., Pyllkkänen, L., & Marcus, G. F. (2013). Top-down influence in young children's linguistic ambiguity resolution. *Developmental Psychology*, 49(6), 1076–1089.
- Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6), 819–865.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1), 39–91.
- Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. *Studies of child language development*, 1, 75–208.
- Slobin, D. I. (1975). *Language change in childhood and in history*. Language Behavior Research Laboratory, University of California.
- Srinivasan, M., & Snedeker, J. (2011). Judging a book by its cover and its contents: The representation of polysemous and homophonous meanings in four-year-old children. *Cognitive Psychology*, 62(4), 245–272.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420.
- Zwicky, A., & Sadock, J. (1975). Ambiguity tests and how to fail them. *Syntax and semantics*, 4(1), 1–36.

### Supplemental material

#### Trial generation

We define  $S(A, B)$  the minimal, well-formed subtree containing  $A$  and  $B$ .

In the *bimodal condition*, the learning exemplars ( $LE_1, LE_2, LE_3, LE_4$ ) were chosen such that ( $LE_1, LE_2$ ) have at least 3 more common ancestors than ( $LE_1, LE_3$ ).  $LE_2$  and  $LE_4$  were chosen such that  $S(LE_1, LE_2)$ , respectively  $S(LE_3, LE_4)$ , was the first subtree which contained at least 4 animals.

The *uniform condition* was defined from the bimodal distribution.  $LE_1$  and  $LE_3$  were kept the same but  $LE_2$  and  $LE_4$  were chosen such that  $S(LE_1, LE_2^{bimodal}) \subset S(LE_1, LE_2^{uniform})$  and  $S(LE_3, LE_4^{bimodal}) \subset S(LE_3, LE_4^{uniform})$  such that  $LE_1, LE_2, LE_3$  and  $LE_4$  were about the same distance (in number of branches) from one another.

The test items (*in, in-superordinate, out*) were chosen such that:

1. 2 *in* items belonged to the smallest subtree containing  $LE_1, LE_2, S(LE_1, LE_2)$  and 2 *in* items belonged to the smallest subtree containing  $LE_3, LE_4$ .
2. *in-superordinate* items were not in  $S(LE_1, LE_2)$ , neither in  $S(LE_3, LE_4)$  but it was in  $S(LE_1, LE_2, LE_3, LE_4)$ .
3. *out* items were not in  $S(LE_1, LE_2, LE_3, LE_4)$ .

The test trials chosen for the present set of experiments were chosen after piloting with adults and selecting the trials that maximized our chances of observing an effect with children.

#### List of trials

List of learning exemplars for each trial in each list. Note that within a list, participants were not asked to learn two words with the same target meaning. Target meanings for the uniform condition are defined according to the *out* items used for each trial. Note that all target meanings may not have a word in children's lexicon. Yet this is not a problem: we do not expect children to map these novel words on words they already know.

List 1		
Condition	LE <sub>1</sub> , LE <sub>2</sub> ,LE <sub>3</sub> ,LE <sub>4</sub>	Target meaning(s)
bimodal	gulf crayfish snake, viper, baboon, gorilla	snake; primate
uniform	clown fish, anaconda, hawk, horse	animal
bimodal	scarab, ladybug, tiger, leopard	beetle; feline
uniform	grizzly, chita, chimpanzee, porcupine	mammal

List 2		
Condition	LE <sub>1</sub> , LE <sub>2</sub> ,LE <sub>3</sub> ,LE <sub>4</sub>	Target meaning(s)
uniform	gulf crayfish snake, swan, baboon, mouse	amniote
bimodal	clown fish, yellow tang fish, hawk, owl	fish; bird
uniform	scarab, toad, tiger, robin	animal
bimodal	grizzly, panda, chimpanzee, cacajao	bear; primate

**Example of script used during the representation test of Experiment 2**

Example of dialogue between the two aliens during the representation test in Experiment 2. All dialogues followed the same frame yet displayed small variations in order to not be repetitive.



Zap: Mais, mais tu dis n'importe quoi Boba!

*But, but Boba, what you say is wrong!*

Boba: Bah, qu'est ce que j'ai dit?

*Huh, what did I say?*

Zap: C'est pas des **[initial word]** ça, c'est des **[corrected word]**!

*These are not [initial word], these are [corrected word]*

Boba: Ooooooh, tu as raison, oui, c'est des **[corrected word]**.

*Ooooooh, you are right, yes, these are [corrected word]*

Zap: Et oui c'est des **[corrected word]**, pas des **[initial word]**!

*Yes, these are [corrected word], not [initial word]!*

## 4.3 Summary and Discussion

### What is missing, what do we learn

One piece of all current word learning models that make them succeed is that they presuppose that word forms map onto a single meaning that ought to be "convex" in conceptual space. This assumption simplifies both the research question and the learning task; yet in some versions of it, it also bans homophony from the system entirely.

In this chapter, I provided the first careful look at what homophones have to say about word learning, from a theoretical and an experimental standpoint both with adults (**Section 4.1**) and with children (**Section 4.2**). On the experimental side, I showed that a word is more likely to yield homophony if: (a) it is learnt from exemplars leaving an important gap between them (in conceptual space), (b) this gap in conceptual space is occupied by other words. These results lead to conclusions beyond homophony, and about the existing constraints on how words associate with concepts in general. For instance, (b) above demonstrates that words compete with each other to occupy the conceptual space so that word learning has to be thought about as a global process over the lexicon, rather than a one-word at a time mechanism.

One of the main contribution of the present set of studies is that currently, learning in different domains (object labels, function words including functional morphemes, numerical concepts) involves different prior hypotheses on the hypothesis space of meanings (and how words maps onto concepts). Yet, this assumes implicitly that learners are already able to distinguish between several word "domains" in the first place, e.g., they are able to trigger different learning algorithm when facing function words vs. content words. Accordingly, word learning algorithm, even those that are argued to emerge from domain-general mechanisms (as in Bayesian models), are more specialized than they look like. I come back to this point further in the General Discussion.

### Learning homophones

This was the first set of studies looking at the simultaneous acquisition of the meanings of a pair of homophones and testing their representations. Recall that in section 3.2, I found that learning a secondary meaning for a known word is easier when the second meaning is semantically distinct from the original meaning of the word. To some extent this is comparable to one of the factors I uncovered here: observing a gap in conceptual space between learning exemplars increases the likelihood that we are in the presence of homophones. Yet the studies reported in section 3.2 did not look at the representation of the words: children were taught that "bath" could label a novel animal and could recognize it during the test

phase, yet it is unclear what kind of representation they entertained for "bath". Did they form another lexical entry to accommodate the novel meaning ("bath" is a homophone)? Did they broaden the original meaning of "bath" to include the novel one ("bath" would be a very broad category encompassing animals, bath items, and everything "in between")? Or did they simply consider that the word could label a set of disconnected concepts ("bath" meaning BATH OR PINK OCTOPUS)? The present set of studies suggests that toddlers are more likely to postulate homophony when there is evidence for an important gap in conceptual space between two learning exemplars of the same word form. Yet we cannot exclude the possibility that younger toddlers may lack the metalinguistic skills to dissociate the level of words from the level of concepts and would thus prefer to associate a single meaning for a given form. Clearly, the children tested in the present study (5.5 years) possess the metalinguistic abilities to conceive the relationship between form and meaning (Bachscheider & Gelman, 1995; Peters & Zaidel, 1980). An interesting open question is thus whether younger children, who have not yet been proven to possess this ability are nevertheless able to build lexical representations that conform with homophony.

## 5 General Discussion

Children start learning their first words within their first year of life and by the time they reach adulthood, they will know approximately 60,000 words (Bloom, 2001). Words are generally acquired without any training or feedback: It seems that whatever inference mechanism children use to learn the meanings of words, they eventually get it right, even when the evidence is scarce or noisy. This may suggest a tight connexion between the mechanisms that children employ when acquiring their lexicon and the way the lexicon is structured and used.

Yet, the presence of ambiguity in the lexicon challenges this view. Indeed, words can have multiple senses (e.g., homophones) and are represented by an arrangement of a finite set of phonemes that potentially increases their confusability (e.g., minimal pairs). Given that children initially appear to experience difficulty learning similar-sounding words (e.g., Stager & Werker, 1997; Swingley & Aslin, 2007) and resist learning homophones (e.g., Casenhiser, 2005; Mazzocco, 1997), lexicons containing many confusable word pairs likely present a problem for children. The fit between lexicon structure and lexical learning abilities may hence not be as good as common intuition may suggest.

Motivated by this apparent discrepancy between learning abilities and typological pattern with respect to ambiguity in the lexicon, this dissertation addresses the link between these two factors. In particular, I addressed the possibility that there may be no paradox between children's learning abilities and the presence of ambiguity in the lexicon: On the one hand the presence of ambiguity in languages may be the consequence of other functional pressures not related with the acquisition of words. On the other hand, the kind of ambiguity that is present in the lexicon is learnable; that is learning may exercise some finer-grained influence on the distribution of ambiguity in the lexicon by keeping only ambiguous words of the learnable kind. I summarize below the results obtained along these two (non-mutually exclusive) lines of work.

### 5.1 The lexicon: The arena of many functional constraints

Languages have to simultaneously satisfy constraints concerning expressive power and ease of learning and processing (in production and in comprehension). Expressive power will

lead to languages that are more complex, while ease of learning and processing tends to maximize simplicity of the linguistic code. Yet importantly, the constraints for ease of processing and learning may conflict: Speakers want many words that are easier to say, thus a more regular and compressible lexicon that maximizes the re-use of word forms and parts of word forms, while listeners and learners want many words that can be easily identified, thus a more distinctive lexicon that maximize the phonological distance between words. However, no quantitative study to date has investigated whether lexicons are more likely to be compressible or more distinctive than *chance* levels, where chance would be what the lexicon would look like in the absence of functional pressures.

In **chapter 2**, I provided such a methodology for quantifying the amount of phonological clustering present in the lexicon. My results show that natural lexicons (at least for the four languages under study) have more similar-sounding words than what would be expected based on chance alone. In addition, I show that greater phonological clustering in the lexicon may be explained (in part) by semantic factors: Across a large corpus of 101 languages, similar-sounding words tend to be semantically closer than expected by chance. This reveals a fundamental drive for compressibility in the lexicon that conflicts with the pressure for words to be as phonetically distinct as possible.

The prevalence of ambiguous words was not measured using this methodology (see the discussion in section 2.3). Yet, previous studies have shown that short, frequent and phonotactically probable words are likely to have more meanings than other, more complex and infrequent words (Piantadosi, Tily, & Gibson, 2012). Thus, all factors that facilitate lexical processing (word length, frequency, phonotactic likelihood) predict, independently, an increase of ambiguity (see for an identical pattern for similar-sounding words, Mahowald et al., *submitted*, Appendix A). This suggests that the frequency distribution of words is structured in a non-arbitrary way, which results in a maximization of the use of ambiguous and similar-sounding words.

Taken together, these results reveals two important tendencies: First, lexicons may be organized less arbitrarily than previously proposed (de Saussure, 1916; Hockett, 1969) – at least when considering the distribution of similar-sounding words and their mappings to meanings. Second, just like Zipfian distributions can be interpreted as the result of a functional trade-off between speakers' and listeners' interests regarding word length and word forms, the distribution of word form similarity in the lexicon appears to be explained by cognitive pressures: Phonological proximity benefits word production (e.g., Dell & Gordon, 2003) but is detrimental for word recognition (e.g., Luce & Pisoni, 1998b) and word learning (e.g., Casenhiser, 2005; Swingley & Aslin, 2007). By assigning similar-sounding and ambiguous word forms to more frequent and predictable meanings, and less similar forms to less frequent and less predictable meanings, languages establish a trade-off between the overall effort needed to produce words and the probability of successful transmission of a message, and thus, of successful learning of the language.

In sum, the above-chance presence of similar-sounding words in the lexicon illustrates the presence of multiple functional pressures that compete in the lexicon to minimize the global cost of the language: Similar-sounding words aid speech production (and memory) beyond the cost of perceptual confusion of these words. In other words, ease of production may weigh more heavily on the presence of similar-sounding words than ease of comprehension or learning, resulting in more compressible lexicons.

## 5.2 Ambiguity in context is not a challenge for language acquisition

Even within the learning system, phonological proximity may display at the same time some functional advantages and some functional disadvantages. To form a novel lexical entry in their lexicon, children must be able to extract a word form and associate it to a meaning. In theory, a compressible lexicon may be advantageous for learning as it reduces the amount of new information that must be represented in the lexicon. For instance, to learn a novel word such as "blick", children need to create a novel phonological representation /blik/ that needs to be associated to a novel semantic representation. Learning several meanings for the same phonological form (or re-using parts of a phonological form) may be more efficient because children only need to learn a novel semantic representation that they can associate with an already existing phonological representation (Storkel & Maekawa, 2005; Storkel, Maekawa, & Aschenbrenner, 2012). Thus, compressible lexicons may display a functional advantage as it minimizes the amount of phonological information that must be learnt and remembered. Yet, compressible lexicons may, at the same time, be functionally challenging for learning as it requires learners to create a new semantic representation when few or no phonological cues can be used to signal that a new meaning is intended (see also section 1.2).

In **chapter 3**, I showed that French 18- to 20-month-old toddlers had no problem learning object labels that were phonological neighbors of a familiar verb (e.g., learning "kiv", a neighbor of "give") but did find it difficult to map neighbors of a familiar noun onto a novel object (e.g., learning "tog", a neighbor of "dog"). This suggests that toddlers are not confused by phonological similarity per se when learning words. In fact, even in cases where the novel word is phonologically identical to a word in toddlers' lexicons (i.e., a homophone), toddlers correctly learnt the novel meaning, provided that the two homophones are sufficiently distant syntactically (e.g. "an eat" is a good name for a novel animal) or semantically (e.g. "a potty" for a novel animal). When the homophones were close on both dimensions (e.g. "a cat" for a novel animal), however, no learning was observed. These results show that toddlers recruit multiple sources of information to infer whether or not a given word form is likely to instantiate a novel meaning. More

generally, the process of creating a lexical entry seems to be mediated by toddlers' existing lexicon and their parsing abilities. This suggests that the functional disadvantage of having similar-sounding and homophonous word pairs in the lexicon may be reduced when the meanings of these words appear in contexts that can be recognized as distinct by children's developing parsing system.

**Chapter 4** formalized this conclusion for current word learning accounts, circumscribing the conditions in which homophonous word representations can emerge. Intuitively, in the case of homophones, we expect the label to occur with a set of exemplars  $S$  drawn from two distinct subcategories  $X_1$  and  $X_2$ . Thus,  $S$  would take the form of a *bimodal* distribution within the single superordinate category  $X$  encompassing the two subcategories (i.e.  $X$  is the minimal well-formed category such that  $X_1 \cup X_2 \subset X$ ). Yet if the word were to be associated with the whole  $X$ , we would expect the exemplars that are associated with the label (i.e.,  $S$ ) to be *uniformly* distributed within  $X$ . My results suggest that children and adults can use such sampling considerations to decide whether a word is associated with one category (standard case) or several categories (homophone), even when exposed to very few exemplars. In particular, I showed that a word is more likely to be considered as a homophone when its exemplars are sampled from semantic categories leaving an important gap between them in conceptual space. Yet, this kind of sampling information may be informative beyond cases of semantic distance between meanings: When the members of a pair of homophones are not from the same syntactic category, the sampling distribution of the syntactic context of the label would thus be bimodal and help learners to postulate homophony.

Interestingly, in **chapter 4**, I also proposed that the learning system is equipped with built-in constraints that allow for the existence of homophony. This is important because contrary to similar-sounding words, homophony presents an additional challenge for word learners: It requires them to discover that there is no one-to-one correspondence between word forms and the associated concepts (as discussed in the Introduction, section 1.2). Certainly, word learning is guided by a set of built-in constraints regarding the possible concepts and form-meaning configurations. This set of experiments makes the following constraints explicit: Learners 1) *always* expect concepts to be convex, that is, to consist of groups of contiguous entities in conceptual space; 2) *generally* expect words to be associated with a single meaning; 3) entertain the possibility that a word maps onto several meanings if concept convexity is challenged. These constraints show that, albeit dispreferred, homophony *can* be a possible outcome of our learning system.

Certainly, the requirements to learn similar-sounding words and homophones are different. While pairs of similar-sounding words can be distinguished at the word form level, homophones cannot: The only way for homophones to be recognized as such is to have sufficiently distinct meanings. Meaning distinctiveness *is* thus fundamental for these words to be learnt and to remain in the language. Interestingly, Bloomfield (1962) reports that in

a dialect of Southwestern France, when the Latin forms "gallus" *rooster* and "cattus" *cat* were in danger of merging into one form, "gat", another novel word acquired the meaning *rooster*, suggesting that the use of the same label for *cat* and *rooster* was unwanted and caused speakers to remap a new form onto one of these meanings. This illustrates that pairs of homophones that belong to the same semantic field tend to be eliminated during the course of language evolution. Conversely, while pairs of similar-sounding words may be easier to learn when they have distinct meanings, this is not a mandatory property for these words to be correctly recognized. In addition, while homophones require that learners form lexical representations that dissociate linguistic signals from concepts (i.e., one word form associated with several meanings), similar-sounding words do not. Meaning distinctiveness may thus be a prerequisite for children to form more complex lexical representations, such as homophony, that depart from an intuitive bias to map each phonological form to a single meaning. Yet meaning distinctiveness may be unnecessary for similar-sounding words, that still conform to a one-to-one mapping between forms and meanings.

While lexicons are compressible at the word form level, are they constrained by other dimensions that maximize meaning distinctiveness? **Chapter 3** showed that while members of a pair of homophone appear to be distinctive, i.e., homophones preferentially appear across syntactic categories rather than within and their meanings are semantically distinct, this does not seem the case for similar-sounding words, i.e., minimal pairs are more likely to appear *within* the same syntactic category and to be semantically related. Thus homophones show an advantage for meaning distinctiveness, but similar-sounding words do not. Yet as **chapters 3 and 4** suggest, this functional disadvantage may be reduced when the meanings of similar-sounding words and homophones can be distinguished in context. This suggests that toddlers' learning abilities impact the way in which homophones are distributed in the lexicon, but not by the way similar-sounding words are organized.

Importantly, this does not mean that there is no functional disadvantage associated with the presence of similar-sounding words that are also more syntactically and semantically similar: As I showed, learning novel words in these conditions is difficult. Yet, this disadvantage may be outweighed by other learning advantages: (a) similar-sounding words may be easier to spot in the speech stream (Altwater-Mackensen & Mani, 2013, see also section 1.2.1); (b) similar-sounding words sharing the same grammatical category may help children group words into categories (i.e., nouns, verbs) (Cassidy & Kelly, 1991; Monaghan et al., 2011); but also other processing advantages: (c) similar-sounding words are easier to produce and to memorize. In sum, the distribution of similar words in the lexicon, both at the word form level and at the syntax/semantic level, may simply reflect a greater functional advantage rather than the absence of functional cost.

The empirical evidence presented in **chapters 3 and 4** suggests that the learning system of young children is equipped with constraints and mechanisms that allow them to successfully learn ambiguous and similar-sounding words as long as these words can be distinguished in



a context that children can capitalize on. Thus, children can deal with ambiguity as long as distinctiveness along other dimensions that are relevant for them is maximized. In addition, I suggest that learning exercises a finer-grained influence on the distribution of ambiguity in the lexicon by selecting ambiguous words whose meaning can be easily disambiguated by the context in which they occur, while pruning out ambiguous words which are not distinguishable through their context, which makes them both hard to learn and prone to triggering misunderstandings. Interestingly the distribution of similar-sounding words in the lexicon did not seem to be affected by children learning difficulties potentially because they confer a greater advantage for speech production and memory (see **chapter 2**). I propose that the difference between the distribution of ambiguous and similar-sounding words illustrates the presence of multiple functional pressures that compete in the lexicon, and that the end result is a trade-off that minimizes the global cost for language users. Yet, while we can probably get an idea of the weight of different functional pressures from observing the structure of the lexicon (and of languages more generally), we cannot tell whether they actually explain why lexicons look the way they are. In the set of studies presented in this thesis, I examine the connexions between learning and processing abilities and lexicon structure, but this work does not provide evidence about whether there is a directional and causal relation between the way our mind is working and language. I turn to this point in the following sections.

### 5.3 How did the lexicon become the way it is?

Language is transmitted culturally from one generation to the next: First-generation speakers produce sentences, which second-generation learners use in order to infer the properties of the language. These cycles of production and inference are crucial in understanding how language has developed and evolved into the structure we observe now.

*Iterated learning* provides a framework to study the emergence of a linguistic system through cultural transmission. Iterated learning is the process by which individuals learn a language produced by a previous individual, who learnt it in the same way (Kirby et al., 2008; Kirby, Griffiths, & Smith, 2014; K. Smith, Kirby, & Brighton, 2003), and can be simulated using computational models or experiments with human participants in the lab. Kirby, Tamariz, Cornish, & Smith (2015) show that the languages that emerge from iterated learning are shaped by the processes of both cross-generation transmission (language learning) and within-generation communication (language use). Iterated learning can thus be used to isolate effects of learning (i.e., where a participant learns a language and then tries to recall it) and the effects of communication (i.e., where participants interact with one another).

Work using this paradigm has consistently shown that individuals will preferentially dis-

card forms and structures that are disadvantageous in favor of other, more advantageous words and phrases (e.g., Reali & Griffiths, 2009; K. Smith & Wonnacott, 2010). For instance, Reali & Griffiths (2009) show that over repeated episodes of learning, a lexicon with multiple labels for objects (synonyms or many-to-one form-meaning mappings) evolves into a lexicon that associates each label with a unique object. This is consistent with the observation that there is a historical tendency for languages to lose many-to-one mappings over time and that children display a bias against many-to-one mappings (Markman & Wachtel, 1988). Note that to date, no work exists investigating in such a paradigm whether lexicons evolve to be more phonologically similar and start tolerating the existence of homophones over time.

Iterated learning thus offers a promising venue for future research to understand *how* functional pressures from both language learning and language usage combine to produce the particular distribution of ambiguous and confusable words found in human languages. Imagine that individuals are taught an artificial lexicon that contains pairs of similar-sounding words and pairs of homophones that vary in their meaning distinctiveness. Our results with child learners predict that the degree to which members of homophone or minimal pairs are distinct may influence participants' learning abilities, thus affecting the transmission of these words to the next generation of learners. We might expect that, across generations, pairs of homophones which cannot be distinguished by their semantic or syntactic context will disappear from the language, while pairs of minimal pairs will tend to stay and be easier to learn if they facilitate syntactic or semantic grouping. This may provide direct evidence that 1) functional processes directly influence the distribution of those words in the lexicon, for instance by looking at whether phonological proximity tends to decrease or increase during learning and/or during communication; 2) the different patterns of distribution found for ambiguous and confusable words stem from different functional pressures that weigh differently in the process of language usage and transmission.

While looking at language evolution in accelerated lab time provides us with an impression of which functional pressures give rise to a particular structure, functional pressures may not be the only determinant of structure. Certainly, languages must adapt to constraints external to the human mind.

## 5.4 The influence of external factors on the lexicon

Several studies have previously demonstrated that the cultural process of transmitting a language, in combination with the constraints and biases of language learners and users, offers an explanation for language structure (e.g., Kirby, 1999). For example, *compositional* languages (i.e., languages in which the meaning of an expression is determined by

the meaning of its constituent expressions) emerge from unstructured languages through repeated transmission through a *learning bottleneck* (e.g., Brighton, 2002; Brighton, Smith, & Kirby, 2005; Vogt, 2005). Because language learners need to infer the whole language from limited evidence (the "bottleneck"), compositionality appears as a natural solution for language since it is the only way for learners to infer the properties of a large linguistic system with limited evidence. Thus, the process of transmission itself may give rise to structure because it constrains the kind of inferences that learners can entertain. In particular, the size of the "bottleneck" influences whether the resulting language will reflect the inductive biases or constraints of the learning system: The smaller the bottleneck the faster it will converge to the priors of the learner (e.g., Griffiths, Christian, & Kalish, 2008; Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007; Reali & Griffiths, 2009), yet when the bottleneck is large (a large amount of information is available to the learner), the language will be a combination of the learner's priors with the distribution of events in the world (the posterior, in Bayesian terms) (Perfors & Navarro, 2014). Thus, the transmission process by itself is a mediating force in the shaping of languages.<sup>19</sup>

These results are particularly interesting in relation to the distribution of homophones in languages. Many have suggested that learners expect words to be associated with a single convex meaning (a one-to-one form-meaning mapping bias). Yet, as we showed in **chapter 4**, learners offset this constraint when they are given evidence that the meaning onto which the word would map is *not* convex, allowing for the formation of homophones. Thus, when the evidence is fairly limited (hence the bottleneck is small), learners will more readily follow the more general constraint that one word should be associated to a single meaning. For instance, if learners are provided with only a small number of exemplars, revealing that the word "bat" is associated to two baseball-bats and one animal-bat, they will be likely to infer that "bat" means baseball-bat and discard the animal-bat exemplar. This may be especially true when one of the two meanings of a pair of homophones is very infrequent. Conversely, if the bottleneck is large, that is, if learners observe many exemplars for each meaning of the homophone pair, learners will have more evidence that concept convexity is not met and will postulate homophony. As such, the amount of homophony that remains in the language should depend on how easy it is for the learner to identify the two meanings (**chapters 3 and 4**) but may also depend on the size of the bottleneck – how much evidence the learner is given to observe. Similarly, languages may have evolved to impose greater phonological overlap to words that are semantically and syntactically related (**chapter 4**) such that learners would quickly group words according to their relevant syntactic or semantic categories even when the size of the bottleneck is small.

The structure of the world also interacts with learners and users of a language and influences

---

<sup>19</sup>Note however that it does not preclude the possibility that the type of response our mind is adopting to the bottleneck problem, i.e., in this example compositionality, may be grounded in human cognition.

the resulting structure. For instance, the *population size* has an influence on language complexity: Languages spoken by larger groups have larger signal inventories (this is true even in animal communicative systems, Freeberg, Dunbar, & Ord, 2012) but simpler morphology than languages spoken in smaller communities, both in real world languages (Lupyan & Dale, 2010; Nettle, 2012) and in artificial learning experiments (Atkinson, Kirby, & Smith, 2015). In addition, the *structure of the environment* influences the structure of the language: If entities in the world are grouped along a single dimension (e.g., objects are naturally grouped according to their size), the languages of iterated learning experiments evolve to reflect that structure independently of the dimension chosen (i.e., size, color, etc) and independently of the language the first generation has to learn (Perfors & Navarro, 2014).<sup>20</sup> Similarly, the *situational context* is an important driver of the space of meanings: If a meaning dimension (e.g., color, shape, motion) is made not relevant during word learning (because it does not help to infer an intended meaning), then languages will not encode this particular meaning dimension (Silvey, Kirby, & Smith, 2014). Others have reported the importance of *frequency of use* as an explanation of language structure (e.g., Bybee, 2006): High frequency patterns are learnt easily, independently of their regularity in the language, while low-frequency patterns are difficult to learn when they are irregular and thus, may disappear from the language. The classical example is the acquisition of English past tense. Irregular past forms (e.g., go → went) are more likely to be frequent in the language and when they are not frequent, they are more likely to show some regularity in their past forms (e.g., breed → bred; bleed → bled; meet → met).

Frequency may be an important factor to explain the existence of similar-sounding words and homophones in the lexicon. Recall that ambiguity and confusability is more likely to appear on frequent forms than on infrequent forms (Mahowald et al., *submitted*; Piantadosi, Tily, & Gibson, 2012; Zipf, 1949). As discussed in **chapter 2**, words with higher frequencies are beneficial for speakers. Yet, in addition, it could also be an advantage for learners: As homophones and similar-sounding words are frequent, this will provide them with enough evidence through the learning bottleneck to learn the meanings of those words, allowing thus for a greater likelihood that these words are transmitted to the next generation and remain in the language. Certainly this also imposes constraints on which homophones are most likely to stay in the languages: Homophones whose meanings occur frequently will be more likely to be transmitted than homophones with unbalanced frequencies where only the most frequent meaning will go through.

Maturation constraints can also impact the structure of language. For instance, when adults and 5-year-olds were exposed to a language with an inconsistent grammar, children *regularized* the language – their output was more consistent than their input, while adults

<sup>20</sup>Certainly, the way we structure the world may, to some extent, also depend on the language we use. For instance, the way that languages partition the color spectrum into labeled categories affects color similarity judgments, memory, and discrimination (e.g., Regier & Kay, 2009; Roberson, Davies, & Davidoff, 2000; Winawer et al., 2007).

reproduced only what they heard (Hudson Kam & Newport, 2009). Hudson Kam & Newport (2009) hypothesized that children's restricted memory capacities limit their ability to store complex forms (see also Newport, 1990), and are thus more likely to impact the structure of language compared to adults. This aligns nicely with the case of similar-sounding and ambiguous words. Maturation constraints linked to memory constraints may limit the appearance of maximally distinctive lexicons since it would be more difficult for children to learn them, as they would need to remember more phonological forms and more sound sequences, and favor compressible lexicons that have a certain amount of phonological overlap.

In the case of homophones, an additional maturational constraint seems critical for children: The ability to conceive the relationship between form and meaning. Metalinguistic awareness of that sort seems to appear between 3 and 4 years of age (Doherty, 2000). For instance, here is the quote of 4-year-old French kid, well aware of homophones (in French /vɛʁ/ means, inter alia, either the color, green; the material, glass; something to drink from, a glass; or an animal, a worm):

C'est /vɛʁ/ mais c'est pas en /vɛʁ/, enfin c'est en /vɛʁ/, mais c'est pas DU /vɛʁ/  
*This is [green] but it is not [glass], well it is in [green], but it is not made of [glass]*

If metalinguistic awareness develops late in younger children, this may limit their abilities to learn homophones. However, it is unclear exactly when this metalinguistic ability develops. Certainly children younger than 3 years of age know a few pairs of homophones (de Carvalho et al., 2014) and can readily attach a second meaning to a known word form in certain conditions (**chapter 3**). Yet, very little is known about the lexical representations of these words at that age and how children reflect about them. One possibility is that learning homophones may not have to be metalinguistic at its earliest stages: Toddlers may start building two separate lexical entries for the same word form because the context in which each meaning is used fully prevents them from accessing the other meaning(s) of the same form. Then, only later, children may realize that the mapping between forms and meanings is ambiguous. According to this view, it is crucial that the meanings of a pair of homophones are distinct for homophones to be able to remain in the language. Another possibility is that children have these metalinguistic skills from the start. Interestingly some toddlers appear to initially use certain word forms to refer to multiple referents (e.g., Vihman, 1981). For instance "cat" could be used in reference to a cat but also in reference to a dog or any object. One interesting question is whether their representation of "cat" is simply greatly underspecified or whether they are aware that the true meaning of "cat" is the set of cats and only cats, and use this word because they do not possess a better alternative in their lexicon. If the latter case is true, this would suggest that children have the ability to distinguish between words and concepts early on and may be able to entertain homophony.<sup>21</sup>

---

<sup>21</sup>Certainly this is not the only way to test whether children know that words and concepts are not related

Taken together, as I reviewed above, there are many factors, determined by functional considerations and the external world, that could explain part of the amount and the particular distribution of homophones and similar-sounding words in the lexicon. In my future work, I will continue pursuing these directions. In particular, I will investigate the role of frequency in the transmission of ambiguous and similar-sounding words in the iterative learning paradigm with adults, and I will extend the study of one-to-many form-meaning mappings to younger children, in order to gain an understanding of the maturational constraints underlying the representation of homophonous meanings.

## 5.5 Insights into the link between language and mind

I started this thesis by noting that there is a connection between the mechanisms by which we acquire and process languages on the one hand and the structure of languages on the other hand. Yet the origin of such a relationship is the topic of a long-standing (but important) debate (Chomsky, 1986; Christiansen & Chater, 2008; Evans & Levinson, 2009; Pinker & Bloom, 1990b). One view argues that this relationship has occurred because of specialized biological machinery that embodies the principles that govern natural languages (e.g., Chomsky, 1986). As such, these principles do not need to be determined by functional considerations since they are hard-wired in the learning system. A second view holds that the structure of language has evolved to fit domain-general learning and processing constraints (e.g., Christiansen & Chater, 2008). From this perspective, the learner's cognitive constraints are likely to be helpful in learning the target language because the language has evolved to conform to these biases. Both views agree that languages are shaped by the mind but differ in whether this occurs through language-specific or domain-general mechanisms. The present work does not aim to disentangle between these views but rather provides some interesting insights into these ideas that will need to be further explored.

For a long time, the presence of ambiguity has challenged the view that language is designed to fit our learning and processing needs. Yet, much evidence (e.g., Piantadosi, Tily, & Gibson, 2012; Wasow et al., 2005), including the data presented in the body of this thesis, proposes that the existence of ambiguity and confusability may find an explanation when one is considering the competing needs of learners, listeners and speakers of the language. The studies in that dissertation started to address this question by looking at a few languages and study child learners and provide an interesting starting point suggesting that ambiguity, and the way it is distributed in the lexicon, may be functionally motivated. Certainly, future work should expand this investigation further: Indeed, most of the languages studied here are historically related (thus share some of their properties)

---

in a one-to-one fashion. Yet this also relates directly to children metalinguistic abilities and provides an interesting framework to study the representation of words in early vocabularies.

and most language acquisition studies are restricted to a few languages. In future work, I plan to overcome the scarcity of data by taking advantage of the existence of big corpora (see **chapter 2**) and simulate language evolution in the lab (e.g., iterated learning) to study different learning and communicative configurations and supplement children's learning data. This will be a crucial component for disentangling the impact of different functional pressures, and understand why/how they weigh differently for different aspects of language.

Does word learning emerge from innately constrained cognitive mechanisms or from more domain-general mechanisms? Bayesian approaches are often called for to reduce learning to domain-general mechanisms. Yet, a careful look at the literature reveals that current models of word learning, including Bayesian models, assume very strong built-in constraints for the learning algorithm. These models do not take into account certain word learning phenomenon, such as homophony, when describing word learning. I suggest that this is due to a specialization of the individual word learning algorithms put forth by researchers: One algorithm can learn object labels, another one can learn function words, yet another can learn number words, etc. Yet none of these algorithms can be used to learn *all* kinds of words. This may not be a problem by itself, but while focusing on learning a single word type, current word learning accounts implicitly incorporate built-in assumptions that may or may not be justified. For instance, the existence of several learning algorithms rests on the assumption that learners already know to distinguish between different word types (e.g., function words and content words or words having several meanings versus words having only one) so that they can apply the relevant learning algorithm in a given situation. One possibility is that these accounts assume that word distinction is an innate property of the learning system. Another possibility is that children use a domain-general mechanism to distinguish between these words: For instance, infants have been shown to use word frequency as a cue to distinguish between function words and content words (Hochmann, Endress, & Mehler, 2010)<sup>22</sup>. Also, it has been shown that infants distinguish between function and content words at birth, likely based on acoustical cues only (Shi, Werker, & Morgan, 1999). At any rate, these implicit priors, and their origin, need to be specified and studied. As I have shown, one way of doing so is to look at more diverse word learning phenomena within the same word learning model to reveal and understand what may count as an actual cognitive constraint and what is merely a simplification of the learning algorithm.

In addition, this thesis provided evidence that the lexicon is strongly constrained by the kinds of *representations* our mind is willing to entertain. As we showed, learners expect concepts to be convex. That is, they expect its members to form a coherent cluster in

---

<sup>22</sup>Though whether this is domain-general or domain-specific is still debatable. On the one hand, there could be an innate constraint saying that frequent morphemes are functional. On the other hand, there could be a general appreciation of frequency: if a given word is too frequent, then it is unlikely to refer to a concrete object and therefore must be used for some other (functional) purposes.



conceptual space. Yet while this constraint is certainly not language-specific, it seems that learners assume that their to-be-learned lexicon is composed of form-meaning mappings of a very specific kind: They expect words to be mapped onto a single concept. This seems a rather specific assumption on the language they are about to learn, yet this does not necessarily imply that this is a language-specific constraint: Learners have been shown to have a preference for simplicity across a wide range of cognitive processes (Chater & Vitányi, 2003), including language learning. Accordingly, learners may expect that forms and meanings come in a one-to-one relation because that may be the simplest representation that learners could possibly find (see also Slobin, 1973, 1975, for the argument that expecting clarity between signals and their functions is an important driving principle). However, one important finding reported in the present work was that learners can entertain more complex meaning representations (homophony) in cases where there is no clear mapping between a word form and a single meaning, but only in the "right" contexts. This is a very specific built-in constraint for a general learning system that seems to expect that a linguistic phenomenon as specific as homophony may be a possibility from the start. Certainly, this calls for further investigation and leaves several open questions for future research, regarding the hierarchy of the constraints of the learning system (as I showed concept convexity seems to be prioritized over forming a one-to-one mapping between forms and meanings), and the possibility of the existence of some very specific constraints on the mapping between forms and meanings in the learning system.

## 5.6 Conclusion

This work investigated why lexicons are ambiguous. A prominent feature of this thesis has been the combined use of lexical models to quantify the amount of ambiguity in the lexicon and experimental methods in toddlers and adults to investigate what exactly enables children to learn ambiguous and confusable words. Taken together, this research suggests that ambiguous and confusable words, while present in the language, may be restricted in their distribution in the lexicon and that these restrictions reflect (in part) how children learn languages, and (in part) the existence of several other constraints on the lexicon. Taken together, this dissertation provides the basis for a research program dedicated to understanding *how* cognitive constraints coming from language acquisition and language use combine to produce the lexicons found in human languages through cultural transmission.





## References

- Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language, 19*(57), 347–358.
- Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.
- Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental science, 16*(6), 980–990.
- Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos, 20*(03), 679–685.
- Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicon. *Cognition, 128*(2), 214–227.
- Atkinson, M., Kirby, S., & Smith, K. (2015). Speaker input variability does not explain why larger populations have simpler languages. *PloS one, 10*(6), e0129463.
- Baayen, R. H., Piepenbrock, R., & van H, R. (1993). The celex lexical data base on cd-rom.
- Backscheider, A. G., & Gelman, S. A. (1995). Children's understanding of homonyms. *Journal of Child Language, 22*(01), 107–127.
- Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258.
- Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition, 127*(3), 391–397.
- Bernal, S., Dehaene-Lambertz, G., Millotte, S., & Christophe, A. (2010). Two-year-olds compute syntactic structure on-line. *Developmental science, 13*(1), 69–76.

- Beveridge, M., & Marsh, L. (1991). The influence of linguistic context on young children's understanding of homophonic words. *Journal of Child Language*, 18(02), 459–467.
- Bion, R. A., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, 126(1), 39–53.
- Bloom, P. (2001). Précis of How children learn the meanings of words. *Behavioral and Brain Sciences*, 24(06), 1095–1103.
- Bloomfield, L. (1962). *Language*. 1933. *Holt, New York*.
- Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4), 298–304.
- Bowerman, M. (2011). *Linguistic typology and first language acquisition*. Oxford University Press.
- Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, 25(4), 535–564.
- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, 8(1), 25–54.
- Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2(3), 177–226.
- Brusini, P. (2012). *Découvrir les noms et les verbes: Quand les classes sémantiques initialisent les catégories syntaxiques* (Unpublished doctoral dissertation). Université Pierre et Marie Curie.
- Bybee, J. (2006). Frequency of use and the organization of language.
- Campbell, R. N., & Bowe, T. (1977). Functional asymmetry in early language understanding. *Salzberger Beiträge für Linguistik*, 3.
- Caramazza, A., & Grober, E. (1976). Polysemy and the structure of the subjective lexicon. *Georgetown University roundtable on languages and linguistics. Semantics: Theory and application*, 181–206.
- Carey, S. (1978a). The child as word learner.

- Carey, S. (1978b). The child as word learner.
- Carlson, M. T., Sonderegger, M., & Bane, M. (2014). How children explore the phonological network in child-directed speech: A survival analysis of children's first word productions. *Journal of Memory and Language*, *75*, 159–180.
- Casenhiser, D. M. (2005). Children's resistance to homonymy: an experimental study of pseudohomonyms. *Journal of Child Language*, *32*(2), 319–343.
- Cassidy, K. W., & Kelly, M. H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*(3), 348–369.
- Cauvet, E., Limissuri, R., Millotte, S., Skoruppa, K., Cabrol, D., & Christophe, A. (2014). Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development*, *10*(1), 1–18.
- Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger Publishers.
- Chomsky, N. (2002). An interview on minimalism. *N. Chomsky, On Nature and Language*, 92–161.
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*(05).
- Conwell, E. (2015). Neural responses to category ambiguous words. *Neuropsychologia*, *69*, 85–92.
- Conwell, E., & Morgan, J. L. (2012). Is it a noun or is it a verb? resolving the ambicategoricity problem. *Language Learning and Development*, *8*(2), 87–112.
- Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*(2), 633–664.
- Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.
- Culbertson, J., Smolensky, P., & Legendre, G. (2011). Learning biases predict a word order universal. *Cognition*.

- Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in french. *Journal of memory and Language*, *42*(4), 465–480.
- Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 892.
- de Carvalho, A., Dautriche, I., & Christophe, A. (2014). *Phrasal prosody constrains online syntactic analysis in two-year-old children*. 39th Boston University Conference on Language Development - Boston, MA - USA.
- de Carvalho, A., Dautriche, I., & Christophe, A. (2015). Preschoolers use phrasal prosody online to constrain syntactic analysis. *Developmental science*.
- Delais-Roussarie, E. (1995). Pour une approche parallele de la structure prosodique: Etude de l'organisation prosodique et rythmique de la phrase française [for a parallel approach of prosodic structure: A study of the prosodic and rhythmic organisation of the french sentence]. *Unpublished Ph. D. thesis. Toulouse, France: Université de Toulouse-Le Mirail*.
- Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? *Phonetics and phonology in language comprehension and production: Differences and similarities*, *6*, 9.
- Demuth, K., & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, *35*(1), 99.
- de Saussure, F. (1916). *Course in general linguistics*. Columbia University Press.
- Doherty, M. J. (2000). Children's understanding of homonymy: metalinguistic awareness and false belief. *Journal of Child Language*, *27*(02), 367–392.
- Doherty, M. J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language*, *31*(1), 203–214.
- Escudero, P., Mulak, K. E., & Vlach, H. A. (2015). Cross-situational learning of minimal word pairs. *Cognitive science*.
- Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429.
- Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469–495.

- Fennell, C. T. (2012). Object Familiarity Enhances Infants' Use of Phonetic Detail in Novel Words. *Infancy*, 17(3), 339–353.
- Fennell, C. T., & Waxman, S. (2006). Infants of 14 months use phonetic detail in novel words embedded in naming phrases. In *Proceedings of the 30th annual boston university conference on language development* (pp. 178–189).
- Flemming, E. (2004). Contrast and perceptual distinctiveness. *Phonetically based phonology*, 232–276.
- Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578–585.
- Freeberg, T. M., Dunbar, R. I., & Ord, T. J. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1597), 1785–1801.
- Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3), 474–496.
- Gelman, S. A., & Brandone, A. C. (2010). Fast-mapping placeholders: Using words to talk about kinds. *Language Learning and Development*, 6(3), 223–240.
- Goldwater, S., & Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics* (Vol. 45, p. 744).
- Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.
- Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, 51(4), 548–567.
- Graf Estes, K., & Bowen, S. (2013). Learning about sounds contributes to learning about words: Effects of prosody and phonotactics on infant word learning. *Journal of Experimental Child Psychology*, 114(3), 405–417.
- Graff, P. N. H. M. (2012). *Communicative efficiency in the lexicon* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Greenberg, J. H. (1966). Universals of language .

- Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*(1), 68–107.
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480.
- Gutman, A., Dautriche, I., Crabbé, B., & Christophe, A. (2014). Bootstrapping the Syntactic Bootstrapper: Probabilistic Labeling of Prosodic Phrases. *Language Acquisition*, 1–25.
- Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.
- Hayes, B. (2012). *BLICK - a phonotactic probability calculator*.
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.
- Hochmann, J. R., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition*, *115*(3), 444.
- Hockett, C. F. (1969). The origin of speech. *Sci. Am*, *203*, 88–111.
- Holst, T., & Nolan, F. (1995). The influence of syntactic structure on [s] to [S] assimilation. *Papers in Laboratory Phonology, IV*, 315–333.
- Howes, D. (1968). Zipf's law and miller's random-monkey model. *The American Journal of Psychology*, *81*(2), 269–272.
- Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.
- Johnson, E. K. (2005). Grammatical gender and early word recognition in dutch. In *Proceedings of the 29th annual boston university conference on language development* (Vol. 1, pp. 320–330).
- Johnson, M., Griffiths, T. L., & Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems* (pp. 641–648).
- Juba, B., Kalai, A. T., Khanna, S., & Sudan, M. (2011). Compression without a common prior: an information-theoretic justification for ambiguity in language.

- Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native. *Journal of memory and language*, *32*(3), 402–420.
- Kamide, Y., Altmann, G., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133–156.
- Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364.
- Kern, S. (2007). Lexicon development in french-speaking infants. *First Language*, *27*(3), 227–250.
- Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford University Press.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.
- Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.
- Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.
- Koehne, J., Trueswell, J. C., & Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In *Proceedings of the 35th annual meeting of the cognitive science society. austin, tx: Cognitive science society*.
- Labov, W. (1966). *The social stratification of English in New York city*. Cambridge University Press.
- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.
- 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*(2), 211.



- Lee, C.-l., & Federmeier, K. D. (2009). Wave-ering: An erp study of syntactic and semantic context effects on ambiguity resolution for noun/verb homographs. *Journal of Memory and Language*, *61*(4), 538–555.
- Lee, C.-l., & Federmeier, K. D. (2012). Ambiguity's aftermath: How age differences in resolving lexical ambiguity affect subsequent comprehension. *Neuropsychologia*, *50*(5), 869–879.
- Lew-Williams, C., & Fernald, A. (2007). Young children learning spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–198.
- Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental phonology*, 13–44.
- Lippeveld, M., & Oshima-Takane, Y. (2014). The effect of input on children's cross-categorical use of polysemous noun-verb pairs. *Language Acquisition*(ahead-of-print), 1–31.
- Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception. Technical Report No. 6.*
- Luce, P. A., & Pisoni, D. B. (1998a). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1), 1.
- Luce, P. A., & Pisoni, D. B. (1998b). Recognizing spoken words: The neighborhood activation model. *Ear and hearing*, *19*(1), 1.
- Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.
- Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.
- Mahowald, K., Dautriche, I., Gibson, E., & Piantadosi, S. (*submitted*). Word forms are structured for efficient use.
- Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*(2), 252–272.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177–190.

- Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.
- Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. *Perspectives on language and thought: Interrelations in development*, 72–106.
- Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology*, 16(1), 1–27.
- Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2), 121–157.
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1), 29–63.
- Mazzocco, M. M. (1997). Children's interpretations of homonyms: a developmental study. *Journal of Child Language*, 24(02), 441–467.
- McKenna, P., & Parry, R. (1994). Category specificity in the naming of natural and man-made objects: Normative data from adults and children. *Neuropsychological Rehabilitation*, 4(3), 255–281.
- Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014–9019.
- Millotte, S., René, A., Wales, R., & Christophe, A. (2008). Phonological phrase boundaries constrain the online syntactic analysis of spoken sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4), 874.
- Millotte, S., Wales, R., & Christophe, A. (2007). Phrasal prosody disambiguates syntax. *Language and Cognitive Processes*, 22(6), 898–909.
- Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91–117.
- Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3), 325–347.
- Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2014). Gavagai is as gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive science*.

- Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130299–20130299.
- Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.
- Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524.
- Newman, R., Samuelson, L., & Gupta, P. (2008). Learning Novel Neighbors: Distributed mappings help children and connectionist models. In *30th annual meeting of the cognitive science society*.
- Newport, E. L. (1990). Maturation constraints on language learning. *Cognitive science*, *14*(1), 11–28.
- Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.
- Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, *4*(10), 906–924.
- Oshima-Takane, Y., Barner, D., Elsabbagh, M., & Guerriero, A. S. (2001). Learning of deverbal nouns. In *Proceedings of the 8th conference of the international association for the study of child language* (pp. 1154–1170).
- Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental bayesian model for learning syntactic categories. In *Proceedings of the twelfth conference on computational natural language learning* (pp. 89–96).
- Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, 384–402.
- Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of functional literacy*, *11*, 67–86.

- Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, *38*(4), 775–793.
- Peters, A. M., & Zaidel, E. (1980). The acquisition of homonymy. *Cognition*, *8*(2), 187–207.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. The MIT Press.
- Pinker, S., & Bloom, P. (1990a). Natural language and natural selection. *Behavioral and brain sciences*, *13*(04), 707–727.
- Pinker, S., & Bloom, P. (1990b). Natural selection and natural language. *Behavioral and Brain Sciences*, *13*(4), 707–784.
- Pinker, S., Lebeaux, D. S., & Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, *26*(3), 195–267.
- Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681.
- Quine, W. V. O. (1960). *Word and Object*. MIT Press.
- Rabagliati, H., Pykkänen, L., & Marcus, G. F. (2013). Top-down influence in young children’s linguistic ambiguity resolution. *Developmental Psychology*, *49*(6), 1076–1089.
- Rabagliati, H., & Snedeker, J. (2013). The Truth About Chickens and Bats: Ambiguity Avoidance Distinguishes Types of Polysemy. *Psychological Science*.
- Rafferty, A. N., Griffiths, T. L., & Ettliger, M. (2013). Greater learnability is not sufficient to produce cultural universals. *Cognition*, *129*(1), 70–87.
- Real, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

- Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive science*, *29*(6), 819–865.
- Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in cognitive sciences*, *13*(10), 439–446.
- Rehurek, R., Sojka, P., & others. (2010). Software framework for topic modelling with large corpora.
- Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants.
- Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, *110*(40), 15937–15942.
- Shannon, C. E. (1947). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3–55.
- Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science*, *11*(3), 407–413.
- Shi, R., & Melancon, A. (2010). Syntactic Categorization in French-Learning Infants. *Infancy*, *15*(5), 517–533.
- Shi, R., & Moisan, A. (2008). Prosodic cues to noun and verb categories in infant-directed speech. In *Buclad 32: Proceedings of the 32th annual boston university conference on language development* (Vol. 2, pp. 450–461).
- Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–B21.
- Silvey, C., Kirby, S., & Smith, K. (2014). Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions. *Cognitive Science*, n/a–n/a.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

- Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. *Studies of child language development*, 1, 75–208.
- Slobin, D. I. (1975). *Language change in childhood and in history*. Language Behavior Research Laboratory, University of California.
- Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3), 229–265.
- Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial life*, 9(4), 371–386.
- Smith, K., Smith, A., Blythe, R., & Vogt, P. (2006). Cross-situational learning: a mathematical approach. *Symbol grounding and beyond*, 31–44.
- Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3), 444–449.
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.
- Sorensen, J. M., Cooper, W. E., & Paccia, J. M. (1978). Speech timing of grammatical categories. *Cognition*, 6(2), 135–153.
- Spinelli, E., & Alario, F.-X. (2002). Gender context effects on homophone words. *Language and Cognitive Processes*, 17(5), 457–469.
- Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640), 381–382.
- Stevens, J., Trueswell, J., Yang, C., & Gleitman, L. (submitted). *The pursuit of word meanings*.
- Steyvers, M., & Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive science*, 29(1), 41–78.
- Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2), 191–211.
- Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of child language*, 32(4), 827.

- Storkel, H. L., Maekawa, J., & Aschenbrenner, A. J. (2012). The Effect of Homonymy on Learning Correctly Articulated Versus Misarticulated Words. *Journal of Speech, Language, and Hearing Research, 56*(2), 694–707.
- Storkel, H. L., & Morrisette, M. L. (2002). The Lexicon and Phonology Interactions in Language Acquisition. *Language, Speech, and Hearing Services in Schools, 33*(1), 24–37.
- Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*(5), 480–484.
- Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive psychology, 54*(2), 99.
- Swinney, D. A. (1979). Lexical access during sentence comprehension:(re) consideration of context effects. *Journal of verbal learning and verbal behavior, 18*(6), 645–659.
- Synnaeve, G., Dautriche, I., Börschinger, B., Johnson, M., & Dupoux, E. (2014). Unsupervised word segmentation in context.
- Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of verbal learning and verbal behavior, 18*(4), 427–440.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.
- Tincoff, R., & Jusczyk, P. W. (2012). Six-Month-Olds Comprehend Words That Refer to Parts of the Body. *Infancy, 17*(4), 432–444.
- Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of memory and language, 39*(1), 102–123.
- Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology, 66*(1), 126–156.
- Van Heugten, M., & Christophe, A. (in press). Infants' acquisition of grammatical gender dependencies. *Infancy*.

- Van Heugten, M., & Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Developmental Science*, *12*(3), 419–425.
- Vihman, M. M. (1981). Phonology and the development of the lexicon: Evidence from children's errors. *Journal of Child Language*, *8*(02), 239–264.
- Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747.
- Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.
- Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, *67*(1), 30–44.
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325–329.
- Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and cognitive processes*, *21*(6), 760–770.
- Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial intelligence*, *167*(1), 206–242.
- Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, *45*(6), 1611–1617.
- Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. *Morphology and the Web of grammar: Essays in memory of Steven G. Lapointe. CSLI Publications*.
- Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*(2), 179–186.
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780–7785.
- Wojcik, E. H., & Saffran, J. R. (2013). The Ontogeny of Lexical Networks: Toddlers Encode the Relationships Among Referents When Learning Novel Words. *Psychological Science*.



- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.
- Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–418.
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.
- Yurovsky, D., & Frank, M. C. (under review). An integrative account of constraints on cross-situational word learning.
- Zangl, R., & Fernald, A. (2007). Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, *3*(3), 199–231.
- Zipf, G. (1949). *Human behavior and the principle of least effort*.

---

# Word forms are structured for efficient use

Kyle Mahowald<sup>\*1</sup>, Isabelle Dautriche<sup>2</sup>, Edward Gibson<sup>1</sup>, and Steven T. Piantadosi<sup>3</sup>

<sup>1</sup>Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, CNRS, EHESS),  
Ecole Normale Supérieure, PSL Research University, Paris, France

<sup>2</sup>Department of Brain and Cognitive Science, MIT

<sup>3</sup>Department of Brain and Cognitive Sciences, University of Rochester

June 24, 2015

**Keywords:** lexicon, word frequency, phonetics, communication, efficiency

## Abstract

If natural language lexicons are structured for ease of production, the easiest-to-produce words will be used more often than words that have higher production costs. This account conflicts with predictions from noisy channel theories, which predict that the most frequent words in a language should be the most phonetically distinct in order to avoid perceptual confusion. We test these competing hypotheses using corpora of 101 languages from Wikipedia. We find that, across a variety of languages and language families, the most frequent forms in a language tend to be more phonotactically well-formed and have more phonological neighbors than less frequent forms.

## 1 Introduction

Zipf famously observed that frequent words tend to be shorter than infrequent words. (Zipf, 1935). This inverse relationship between word length and word frequency, and a closely related inverse relationship between word length and predictability in context (Piantadosi et al., 2011), has since been found across a variety of languages. These statistical patterns in the lexicon are most likely a functional product of language use (Piantadosi, 2014; Zipf, 1949): from an information-theoretic perspective (Shannon, 1948), it is more efficient for the most predictable, most frequently used codes to be short. By assigning shorter forms to more frequent and predictable meanings and longer forms to less frequent and less predictable meanings, languages establish a trade-off between the overall effort needed to produce words and the chances of successful transmission of a message. It would be onerous and slow to have to re-use long words over and over when the content is predictable. But when the content is unpredictable, longer words are likely to be more robust to noise and allow for information to be transmitted at a rate at which the message will be understood (Aylett & Turk, 2004; Levy & Jaeger, 2007; van Son & Pols, 2003).

Besides length, one important dimension by which words can vary is their phonological form. In addition to theorizing about length, Zipf (1935) also claimed that the Principle of Least Effort predicts that easily articulated sounds should be used more often in language than more difficult sounds. In this work,

---

<sup>\*</sup>For correspondence, e-mail [kylemaho@mit.edu](mailto:kylemaho@mit.edu)

we operationalize this idea in terms of *phonotactic probability*. Some words, like *cat*, are composed of sequences that are easy to pronounce. Other words in the language, like *dwarf*, are harder to pronounce because they have more unusual phoneme sequences. A side effect of phonotactic probability is that words with higher phonotactic probability tend to be more phonologically similar to one another. *Phonological neighborhood density* is a measure of the number of words in the lexicon that are phonologically similar to a given target word (Luce, 1986; Vitevitch & Luce, 1998). For instance, phonological neighbors of *cat* include *cast*, *bat*, and *at*. Critically, words with high phonotactic probability are likely to have higher phonological neighborhood density than words with lower phonotactic probability since highly probable words are likely close in phonetic space to other phonotactically probable words.

Both phonotactic probability and neighborhood density have been shown to play a critical role in language processing and acquisition, as we discuss below in more detail.

### **Phonotactic probability**

Phonotactic probability is a measure of the well-formedness of a string in a given language. For instance, in English, the word *drop* is phonotactically quite probable, *dwop* is less probable but still allowed, and *dsop* has, essentially, no probability. In this work, we will use orthographic probability, as measured by an *n*-gram model, as a proxy for phonotactic probability. Under a bigram model of orthographic probability, for instance, the probability of a string like *drop* would depend on the probability of the two-letter sequences that make up the word: *dr*, *ro*, and *rp*.

Phonotactic probability is likely related to articulatory constraints as well as other factors, and there is compelling evidence that phonotactically probable words are easier to produce and understand in language use. For instance, it has been claimed that the phonetics of languages evolve to enable easy articulation and perception (Lindblom, 1983, 1990, 1992) and the patterns of sounds observed across languages reflect articulatory constraints (Kawasaki & Ohala, 1980). Therefore, a language whose most frequent words are phonotactically probable likely requires less production effort than a language organized such that the most frequent strings are not phonotactically likely.

Less obviously, the same may also be true of comprehension and learning: phonotactically probable words are more easily recognized than less probable words (Vitevitch, 1999). And there appears to be a learning advantage for probable strings: probable strings are learned more easily by infants and children (Coady & Aslin, 2004; Storkel, 2004, 2009; Storkel & Hoover, 2010) and infants prefer high-probability sequences of sounds compared to lower probability sequences (Jusczyk & Luce, 1994; Ngon et al., 2013). All of this evidence suggests a functional advantage to using phonotactically probable words. Here, we will think of a word's phonotactic probability as a proxy for the *cost* of using that particular word.

### **Phonological neighborhood density**

Phonological neighborhood density of a word *w* is the number of words that differ from word *w* by one insertion, deletion, or substitution (Luce, 1986; Vitevitch & Luce, 1998). On one hand, words with many neighbors have an inhibitory effect on lexical access in perception (Luce, 1986; Vitevitch & Luce, 1998) and elicit lexical competition that slows down word learning in toddlers (Swingley & Aslin, 2007). Moreover, Magnuson et al. (2007) shows that high-density word onsets inhibit reading times. However, phonological similarity (a) facilitates the ease with which people produce words (Gahl et al., 2012; Stemberger, 2004; Vitevitch & Sommers, 2003); (b) supports novel word representation in working memory (Storkel & Lee, 2011) and (c) boosts word learnability in adults (Storkel et al., 2006). As Dell & Gordon (2003)

---

demonstrate in their model, phonological similarity in the lexicon challenges word recognition yet benefits word production.<sup>1</sup> This asymmetry between the effect of phonological similarity in word recognition and word production provides a window into the functional pressures that act on wordform similarity and makes competing predictions as to whether the most frequent words in a language should have more or fewer phonological neighbors than less frequent words.

### The present study

What we know about neighborhood density and phonotactic probability (which are highly correlated with each other) leads to two possible predictions about the functional organization of the lexicon. On one hand, following a noisy channel setting (Gibson et al., 2013; Levy, 2008; P. Smith, 1970) where a speaker is transmitting a message to a receiver with some probability of error along the way, one wants to make sure that the most frequent words are most perceptually distinct from each other in order to minimize the number of errors made. On the other hand, if speakers prefer to re-use common articulatory sound sequences, one should structure the lexicon such that the most frequent words consist of phonotactically likely strings. That way, the infrequent and hard-to-pronounce words only rarely need to be used.

To evaluate the extent to which the phonological forms of words may be explained by word usage, we investigated whether (a) wordforms that are orthographically probable (as measured over word types) are likely to be more frequent (by token) than wordforms that are less orthographically probable and (b) whether wordforms that are orthographically similar to other words are likely to be more frequent than phonologically more unique wordforms.

If we observe a *positive* correlation between frequency and orthographic probability and between frequency and phonological density, it would indicate that lexicons are structured so that the most commonly used words are easy to produce. In contrast, a *negative* correlation would indicate that highly-frequent words are more subject to a pressure for dispersion of word forms. Note that, because we train the phonotactic model and measure neighbors using *unique* word forms and then correlate those measures with the token frequency of those word forms, we avoid any circularity in this analysis. That is, the estimates of phonotactic probability and neighborhood density do not depend on the frequency of the word forms.

This kind of correlation has been examined in the literature before, but only for a small number of languages. Landauer & Streeter (1973) performed a similar analysis for English, and Frauenfelder et al. (1993) for English and Dutch. All found that the most frequent words in the language have higher phonotactic probability and more phonological neighbors than more infrequent words. While these results are suggestive, it is difficult to draw conclusions based on a small set of related languages. In the current study, we used orthographic lexicons from 101 typologically diverse languages downloaded from Wikipedia in order to investigate whether the relationship between phonotactic probability, neighborhood density, and frequency reflect functional constraints. We found that frequent wordforms tend to be phonotactically likely and have more neighbors than less frequent wordforms, suggesting that there is a functional pressure associated with word usage for languages to prefer phonotactically probable strings that are phonologically

---

<sup>1</sup>Sadat et al. (2014), however, find that phonological neighborhood density actually causes longer naming latencies in production and therefore has an inhibitory effect—but that there are also facilitative effects of neighborhood density in lexical access. Vitevitch & Stamer (2006) (like Sadat et al. (2014)) argue that morphologically rich languages like Spanish and French typically show an inhibitory effect for words with many neighbors in naming tasks. Chen & Mirman (2012) present a model showing that, whether or not neighborhood effects in general are facilitative or inhibitory may be task dependent. While the literature on the topic is large and complex, for our purposes, it is sufficient to acknowledge that there is at least some facilitative effect in language production associated with having an easy-to-pronounce string with many neighbors relative to a difficult-to-pronounce string with few neighbors. Language production would be inhibited if speakers had to rely mostly on difficult-to-pronounce wordforms.

**West Germanic:** Afrikaans, German, English, Luxembourgish, Low Saxon, Dutch, Scots, Yiddish, Alemannic; **Goidelic:** Irish, Scottish Gaelic; **Brythonic:** Breton, Welsh; **Hellenic:** Greek; **South Slavic:** Bulgarian, Macedonian, Serbo-Croatian, Slovene; **Albanian:** Albanian; **Iranian:** Central Kurdish, Persian, Kurdish, Mazandarani, Tajik; **Romance:** Aragonese, Asturian, Catalan, Spanish, French, Galician, Italian, Lombard, Neapolitan, Occitan, Piedmontese, Portuguese, Romanian, Sicilian, Venetian, Walloon; **West Slavic:** Czech, Polish, Slovak; **Armenian:** Armenian; **Italic:** Latin; **North Germanic:** Danish, Icelandic, Norwegian (Nynorsk), Norwegian (Bokmal), Swedish; **Baltic:** Lithuanian, Latvian; **Indo-Aryan:** Fiji Hindi, Marathi, Urdu, Bosnian, Croatian, Punjabi, Serbian; **East Slavic:** Belarusian, Russian, Ukrainian; **Frisian:** West Frisian

**Table 1:** Table of Indo-European languages used, language families in bold.

**Austronesian:** Minang, Amharic, Indonesian, Malay, Sundanese, Cebuano, Tagalog, Waray-Waray, Buginese, Javanese; **Altaic:** Mongolian, Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Turkish, Tatar, Uzbek; **creole:** Haitian; **Austroasiatic:** Vietnamese; **Kartvelian:** Georgian; **Niger-Congo:** Swahili, Yoruba; **Vasonic:** Basque; **Afro-Asiatic:** Malagasy; **Quechuan:** Quechua; **Semitic:** Arabic, Egyptian Arabic, Hebrew; **Korean:** Korean; **Uralic:** Estonian, Finnish, Hungarian; **Tai:** Thai; **constructed:** Esperanto, Interlingua, Ido, Volap

**Table 2:** Table of non-Indo-European languages used, language families in bold.

more similar to one another.

## 2 Method

### 101 orthographic lexicons:

We used the lexicons of 101 languages extracted from Wikipedia. The details on these lexicons, including the typological details and our corpus cleaning procedure, are explained in Appendix A. The languages analyzed included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, 12 language families are represented as well as a Creole and 4 constructed languages (Esperanto, Interlingua, Ido, Volap) that have some speakers. (The analysis is qualitatively the same after excluding constructed languages.) The languages analyzed are shown in Tables 1 and 2.

For this analysis, we defined a lexicon as the set of the 20,000 most frequent unique orthographic wordforms (word types) in a given language. We used only orthographic wordforms here, which are a good proxy for phonological forms (an assumption tested in Section 3.2 for a small number of languages).

### 3 phonemic lexicons:

To assess whether the Wikipedia corpus (which uses orthographic forms and contains morphologically complex words) is a good proxy for a more controlled corpus that uses phonemic representations and is restricted to monomorphemic words, we also analyzed phonemic lexicons derived from CELEX for Dutch, English and German (Baayen et al., 1995) and Lexique for French (New et al., 2004). The lexicons were

restricted to include only monomorphemic lemmas (coded as "M" in CELEX; I.D. (a French native speaker) identified mono-morphemes by hand for French). That is, they contained neither inflectional affixes (like plural *-s*) nor derivational affixes like *-ness*. In order to focus on the most used parts of the lexicon, we selected only words whose frequency is greater than 0. (The CELEX database includes some rare words listed as having 0 frequency, which were not in the original CELEX sample.) Since we used the surface phonemic form, when several words shared the same phonemic form (e.g., 'bat') we included this form only once.

All three CELEX dictionaries were transformed to make diphthongs into 2-character strings. In each lexicon, we removed a small set of words containing foreign characters. This resulted in a lexicon of 5459 words for Dutch, 6512 words for English, 4219 words for German and 6782 words for French.

**Variables under consideration:**

For each word in each language we computed the word's:

- Token frequency: for orthographic lexicons: across all the Wikipedia corpus of the language; for phonemic lexicons: using the frequency in CELEX
- Orthographic probability (as a proxy for phonotactic probability): We trained an ngram model on characters ( $n = 3$  with a Laplace smoothing of 0.01 and with Katz backoff in order to account for unseen but possible sound sequences) on each lexicon and used the resulting model to find the probability of each word string under the model. Table 3 shows examples of high and low probability English words under the English language model.

word	log probability
shed	-3.75
reed	-3.69
mention	- 4.63
comment	-4.68
tsar	-8.64
Iowa	-9.47
tsunami	- 12.90
kremlin	-11.53

**Table 3:** Phonotactically **likely** and **unlikely** words in English with their log probabilities

- Orthographic neighborhood density (as a proxy for phonological neighborhood density): PND is defined for each word as the number of other words in the lexicon that are one edit (an insertion, deletion, or substitution) away in phonological space (Luce, 1986; Luce & Pisoni, 1998). For instance, 'cat' and 'bat' are phonological neighbors, as well as minimal pairs since they have the same number of letters and differ by 1. 'Cat' and 'cast' are neighbors but not minimal pairs. We will focus on minimal pairs, as opposed to neighbors, in order to avoid confounds from languages having different distributions of word lengths.

word length	mean correlation	proportion showing positive correlation	proportion showing significant correlation
3 letters	.27 (.26)	1.00 (1.00)	.97 (.92)
4 letters	.24 (.24)	.98 (.98)	.97 (.96)
5 letters	.23 (.22)	.99 (.99)	.98 (.96)
6 letters	.21 (.20)	1.00 (1.00)	.97 (.98)
7 letters	.19 (.19)	1.00 (1.00)	.98 (.99)

**Table 4:** Separated by length, (a) the mean correlation across languages for the relationship between orthographic probability and frequency, (b) the proportion of languages that show a positive correlation between orthographic probability and frequency, and (c) the proportion of languages for which this relationship is significantly different from 0 at  $p < .05$ . In parentheses, we include each value for the subset of the lexicons that do not appear in the English Subtlex subtitles corpus.

word length	mean correlation	proportion showing positive correlation	proportion showing significant correlation
3 letters	.19 (.19)	1.00 (.99)	.97 (.84)
4 letters	.17 (.16)	.98 (.99)	.97 (.93)
5 letters	.18 (.18)	.98 (.98)	.98 (.96)
6 letters	.19 (.18)	1.00 (1.00)	.97 (.97)
7 letters	.18 (.18)	.99 (.99)	.98 (.96)

**Table 5:** Separated by length, (a) the mean correlation across languages for the relationship between number of minimal pairs and frequency, (b) the proportion of languages that show a positive correlation between number of minimal pairs and frequency, and (c) the proportion of languages for which this relationship is significantly different from 0 at  $p < .05$ . In parentheses, we include each value for the subset of the lexicons that do not appear in the English Subtlex subtitles corpus.

### 3 Results

#### 3.1 Large-scale effects of frequency on 101 languages

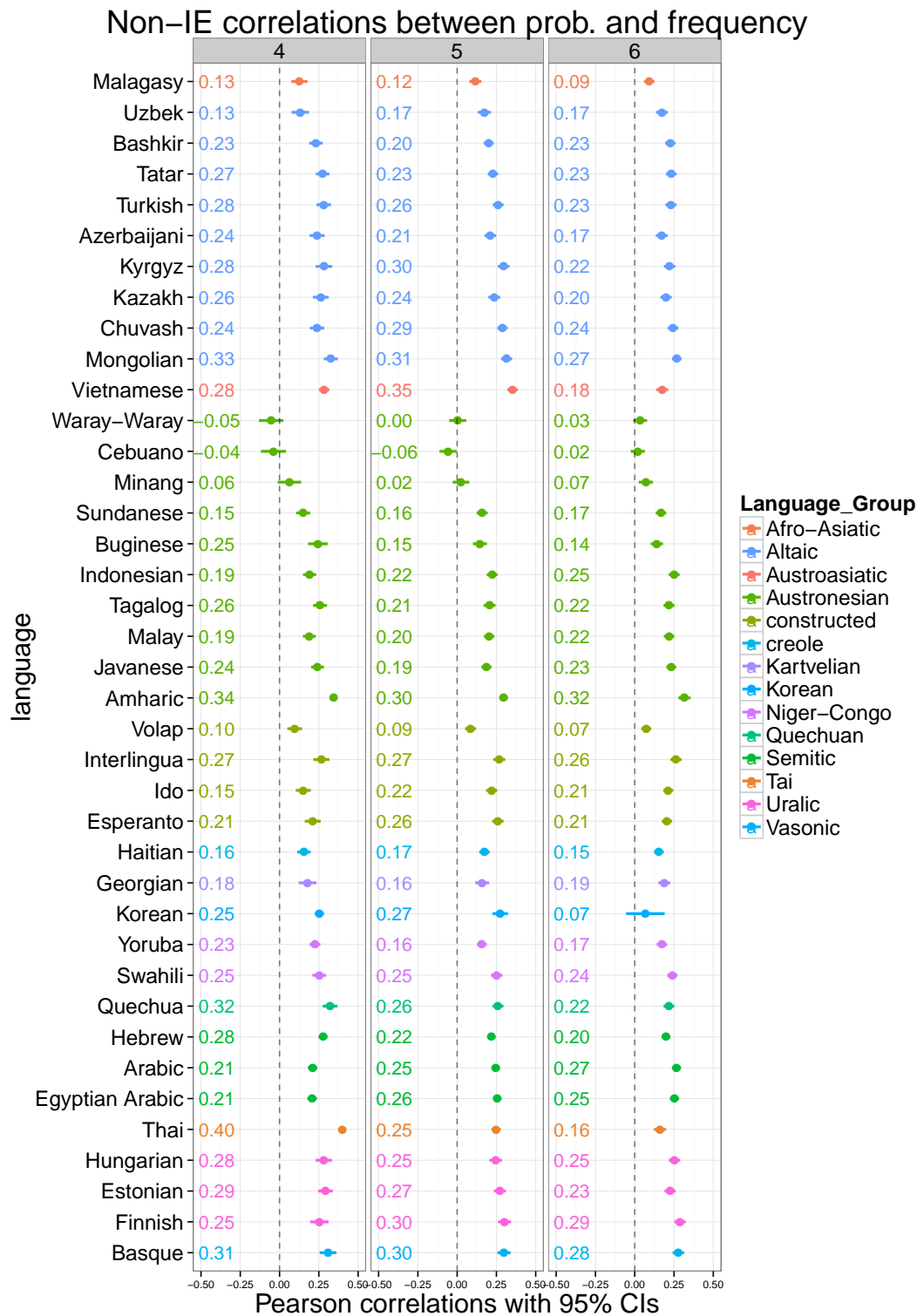
##### Correlational analysis

Figures 1 and 2 plot the correlations for each language and length (from 4 to 6 letters) separately, between orthographic probability and frequency for non-Indo-European and Indo-European language respectively. Points to the right of the dotted line at 0 show a positive correlation. Almost all languages indeed show a positive correlation.

Figures 3 and 4 are similar plots for the correlation between minimal pairs and frequency. Once again, almost all languages show a positive correlation.

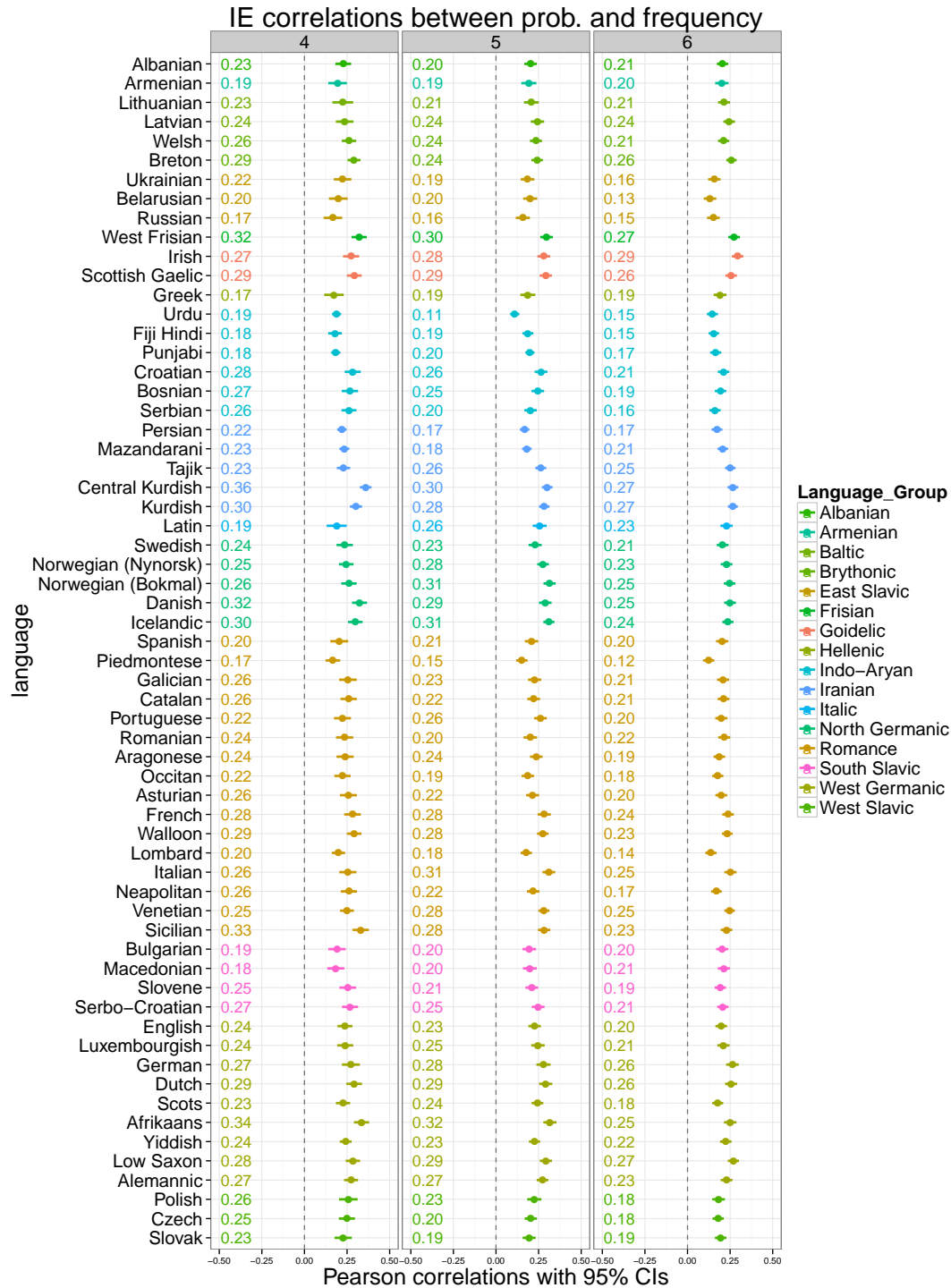
Analyzing each length separately and focusing on words of 3 to 7 letters, we found a significant correlation between log frequency and orthographic probability in most languages (see Table 4). For instance for the 4-letter words, 99 of 101 languages showed a positive correlation and 98 out of the 101 correlations were significantly positive at  $p < .05$ .

We also found a robust correlation between log frequency and number of minimal pairs (mean  $r = .18$  across lengths and languages) for almost all languages, as shown in Table 5.



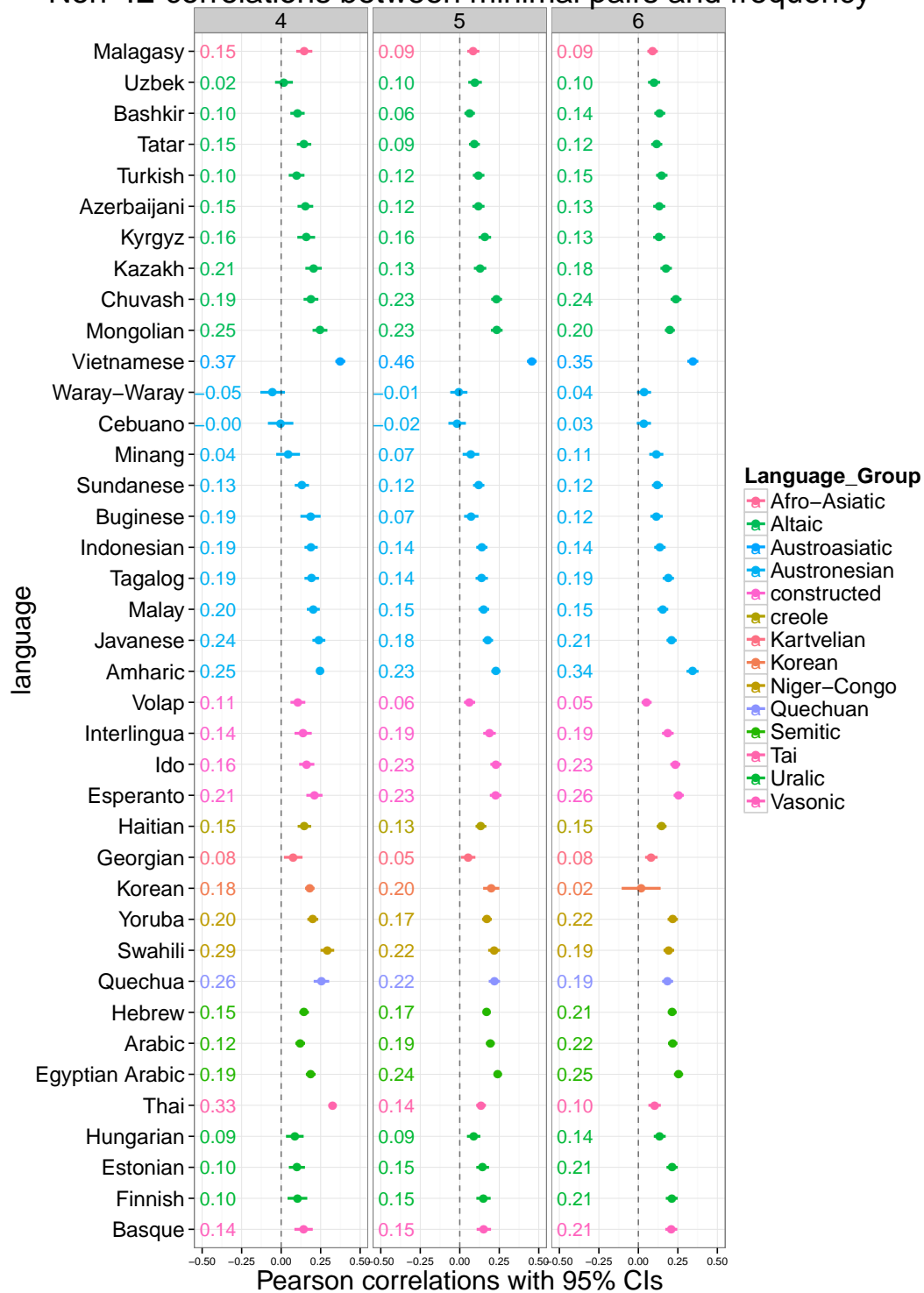
**Figure 1:** Correlation coefficients between orthographic probability and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.



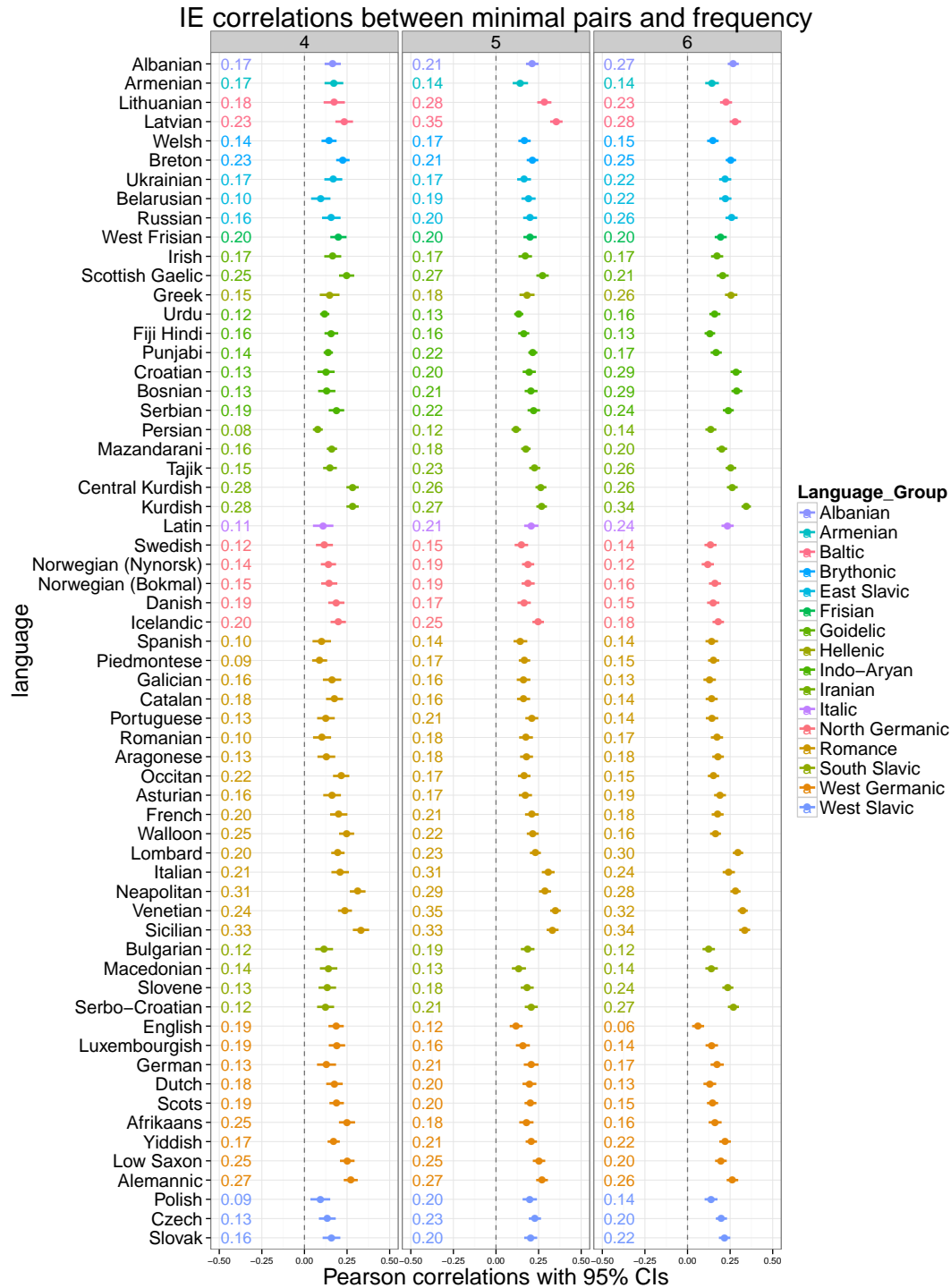


**Figure 2:** Correlation coefficients between orthographic probability and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.

### Non-IE correlations between minimal pairs and frequency



**Figure 3:** Correlation coefficients between number of minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for non-Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.



**Figure 4:** Correlation coefficients between number of minimal pairs and frequency, by language and length, with 95% confidence intervals based on Fisher transforms for words of length 4 to 6 for Indo-European languages. Dots to the right of the dotted line at 0 show a positive correlation. The numbers along the y-axis are the Pearson correlations. Text and points are colored by language family.

---

In order to ensure that any observed effects are not the product of English overlap, we ran the same analyses on the full lexicons as well as on subsets of lexicons that exclude any word that also appears in the English Subtlex subtitles database (Brysbaert & New, 2009). This excludes English intrusions but also excludes perfectly good words like *die* in German (which means “the” and is unrelated to English “die”) and French *dire* (meaning “to say” and unrelated to the English adjective *dire*). Note that, for all lengths, the results obtained when excluding all English words are similar in terms of overall correlation. Because most of the English words excluded are actually not intrusions but are native words that just happen to also be English forms, we include them in all subsequent analyses.

Additionally, we find a robust correlation between orthographic probability and number of minimal pairs (mean  $r = .49$  across all lengths considered). This result holds for all lengths across the vast majority of languages and is consistent with the idea that words with high orthographic probability are more likely to have neighbors since the orthographic probabilities of their neighbors will be on average high too. For example, a word like ‘set’ is more likely to have more minimal pairs in English than the word ‘quiz’ simply because the letters in ‘set’ are more common and so, probabilistically, there are more opportunities for a word to be orthographically close to ‘set’ than to ‘quiz.’

It follows that the correlations between frequency and phonological similarity uncovered previously should be (partly) due to both frequency and orthographic probability being correlated with phonological similarity. Thus, the question becomes (a) whether the correlation between frequency and phonological similarity remains after factoring out the effect of orthographic probability and (b) whether the correlation between frequency and orthographic probability remains after factoring out the effect of phonological similarity. Moreover, many of the languages in this study are highly related, so we need an analysis that generalizes across families and languages to make sure that the effect is not just lineage-specific.

### Mixed effect analysis

We ran a mixed effect regression predicting (scaled) frequency for each word from orthographic probability and number of minimal pairs, where both predictors were normalized for each language and length. We used a maximal random effect structure with random intercepts for each language, language sub-family, and language family and slopes for orthographic probability and number of minimal pairs for each of those random intercepts. In effect, this random effect structure allows for the possibility that some languages or language families show the predicted effect whereas others do not. It allows us to test whether the effect exists beyond just language-specific trends. Because of the complex random effect structure and the large number of data points, we fit each length separately and focused on words of length 3 through length 7.

For 4-letter words (a representative length), a 1 standard deviation increase in orthographic probability was predictive of a .20 standard deviation increase in frequency; a 1 standard deviation increase in number of minimal pairs was predictive of a .06 standard deviation increase in frequency. To assess the significance of orthographic probability above and beyond the number of minimal pairs, we performed a likelihood ratio test comparing the full model to an identical model without a fixed effect for orthographic probability (but the same random effect structure). The full model was significantly better by a chi-squared test for goodness of fit ( $\chi^2(1) = 30.9, p < .0001$ ). To assess the significance of the number of minimal pairs above and beyond the effect of orthographic probability, we compared the full model to an identical model without a fixed effect for the number of minimal pairs using a likelihood ratio test. Once again, the full model explained the data significantly better ( $\chi^2(1) = 10.6, p < .001$ ). Thus, both the number of minimal pairs and orthographic probability appear to make independent contributions in explaining word frequency. This effect holds above and beyond effects of language family or sub-family, which are included in the model

word length	orthographic probability	number of minimal pairs
3 letters	.23**	.08**
4 letters	.20***	.06***
5 letters	.19***	.07**
6 letters	.15***	.11***
7 letters	.13***	.11***

**Table 6:** Separated by length, the model coefficient from the full model including random intercepts and slopes for language, sub-family, and family for orthographic probability and number of minimal pairs. Two asterisks means that by a likelihood test, the predictor significantly improves model fit at  $p < .01$ . Three asterisks means  $p < .001$ .

as random effects. Note that the effect size is larger for orthographic probability than it is for number of minimal pairs and that a model including a fixed effect of probability but not minimal pairs has a better model fit (AIC = 520310) than one that includes minimal pairs but not probability as a fixed effect (AIC = 520330). We find a similar pattern of results for all other lengths examined, as summarized in Table 6. Overall, these results suggest that both the number of minimal pairs and the orthographic probability independently predict frequency but that the effect of orthographic probability is stronger and is likely, in part, driving the neighborhood effect.

In Appendix B, we show the results of a lasso regression (Tibshirani, 1996), for each length and language, predicting scaled frequency from scaled orthographic probability and scaled number of minimal pairs. As with our other analyses, this analysis suggests that more frequent words have higher orthographic probability and more minimal pairs but that the minimal pairs result is driven, at least in part, by orthographic probability.

### 3.2 Testing correlation generalizability to phonemic representations

We used orthographic lexicons because they could be easily extracted for a large number of languages. However, a better measure of phonotactics could be calculated on the phonemic transcription of words, and a better measure of phonological similarity should exclude morphological similarity by focusing only on monomorphemes. To assess whether the correlation between frequency and phonological similarity and between frequency and phonotactic probability hold in a set of monomorphemic words with phonemic representations, we performed the same analysis using the four phonemic lexicons from Dutch, English, French, and German.

As before, we tested whether the token frequency could be predicted by phonotactic probability (here approximated by *phonemic* probability using a ngram model operating over triphones) and/or number of minimal pairs. The correlations obtained in these four phonemic lexicons replicated previous correlations with the orthographic lexicons for these languages: all four languages still showed positive correlations for the relationship between phonotactic probability and frequency and between number of minimal pairs and frequency.

That said, the correlations were slightly lower in the more controlled set for the 4 languages than when using the same measures in the larger data set: the correlation between minimal pairs and frequencies (across the 4 languages and word lengths 3-7) is, on average, .03 lower for the correlation between minimal pairs and frequency and .10 lower for the correlation between orthographic probability and frequency. This

---

suggests that part of the effect could be driven by morphology—which is absent in the controlled phonemic lexicons but present in the Wikipedia corpus.

## 4 Discussion

We found that frequent wordforms are more likely to be similar to other wordforms and composed of more likely sequence of phonemes than infrequent ones. These correlations were robustly present across a large number and wide variety of typologically different languages. Just as the Zipfian word frequency distribution allows for functional optimization of word lengths (Piantadosi, 2014; Piantadosi et al., 2011) and word forms (Piantadosi et al., 2012), this work shows that the frequency profile of even words of the same length is structured in a non-arbitrary way so as to maximize the use of “good” word forms.

Note that, from a noisy channel perspective, there is a tradeoff in structuring the lexicon this way. In a language where phonotactically probable strings in dense neighborhoods are the most frequent words, there may be an increased chance of perceptual confusion. On the other hand, it is possible that in everyday speech, noise conditions may not be extreme enough that perceptual confusion would be an issue. Or maybe the cost of perceptual confusion—asking an interlocutor to repeat a phrase or getting the information some other way—is typically not high enough to offset the cost that would come with structuring the lexicon such that the most frequent word forms were phonetically odd.

We do not believe that the main result of this paper is purely a result of morphological regularity since the same analyses run on monomorphemic words in a subset of languages show the same pattern of results. Moreover, although phonotactic constraints are an obvious and major source of regularity in the lexicon, it is important to note that these results are not likely just the result of phonotactic constraints since the results hold even after controlling for the influence of phonotactic probability, at least in the analyses reflecting the influence of word usage. In a companion study (Mahowald et al., *submitted*) in which we constructed a phonotactically-controlled baseline for lexicons, we provided compelling evidence that natural lexicons are more tightly clustered in phonetic space than would be expected by chance, over and above the constraints imposed by phonotactics. This, taken together with the present results, suggests that languages tend to favor wordform similarity in the lexicon.

In this study, we addressed the issue of *why* wordform similarity in lexicons diverges from what can be expected by chance alone, but we leave it to future work to investigate how it got to be that way. One promising body of work in that vein concerns language evolution. Indeed, there has been much experimental work studying the evolution of language showing that language users will preferentially discard forms and structures that are disadvantageous in favor of other, fitter words and phrases (Fedzechkina et al., 2012; K. Smith et al., 2003). Thus, one plausible mechanism for the effects described here is that generations of learners improve on the lexicon, honing it over time by avoiding words that are too strange, complex, or that otherwise don’t fit with the rest of the words in the lexicon.

## References

- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31–56.
- Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2) [cd-rom]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

- Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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(4), 977–990.
- Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, 119(2), 417.
- Coady, J. A., & Aslin, R. N. (2004). Young children's sensitivity to probabilistic phonotactics in the developing lexicon. *Journal of Experimental Child Psychology*, 89(3), 183–213.
- Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 6, 9.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44), 17897–17902.
- Frauenfelder, U., Baayen, R., & Hellwig, F. (1993). Neighborhood density and frequency across languages and modalities. *Journal of Memory and Language*, 32(6), 781–804.
- Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806.
- Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*.
- Jusczyk, P., & Luce, P. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630–645.
- Kawasaki, H., & Ohala, J. J. (1980). Acoustic basis for universal constraints on sound sequences. *The Journal of the Acoustical Society of America*, 68(S1), S33–S33.
- Landauer, T., & Streeter, L. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12(2), 119–131.
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing* (p. 234–243).
- Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In H. T. Schalkopf & B. Platt (Eds.), (pp. 849–856). Cambridge, MA: MIT Press.
- Lindblom, B. (1983). *Economy of speech gestures*. Springer.
- Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling* (pp. 403–439). Springer.
- Lindblom, B. (1992). Phonological units as adaptive emergents of lexical development. *Phonological development: Models, research, implications*, 131–163.

- 
- Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on Speech Perception. Technical Report No. 6.*
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and hearing, 19*(1), 1.
- Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*(1), 133–156.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516–524.
- Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science, 16*(1), 24–34.
- Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130.
- Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526.
- Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*(3), 280–291.
- Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive psychology, 68*, 33–58.
- Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 623–656.
- Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life, 9*(4), 371–386.
- Smith, P. (1970). Communication over noisy channels: Applications to the statistical structure of english. *British Journal of Psychology, 61*(2), 197–206.
- Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language, 90*(1), 413–422.
- Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25*(02).
- Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language, 36*(02), 291.
- Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49*(6), 1175–1192.
- Storkel, H. L., & Hoover, J. R. (2010). An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken american english. *Behavior Research Methods, 42*(2), 497–506.



- Storkel, H. L., & Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2), 191–211.
- Swingle, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive psychology*, 54(2), 99.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- van Son, R., & Pols, L. (2003). How efficient is speech? *Proceedings Institute of Phonetic Sciences, University of Amsterdam*, 25, 171–184.
- Vitevitch, M. S. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2), 306–311.
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, 9(4), 325–329.
- Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4), 491–504.
- Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and cognitive processes*, 21(6), 760–770.
- Zipf, G. (1935). *The psychology of language*. NY Houghton-Mifflin.
- Zipf, G. (1949). *Human behavior and the principle of least effort*.

---

## A Appendix: Dataset of 101 lexicons from Wikipedia

We started with lexicons of 115 languages from their Wikipedia databases (<https://dumps.wikimedia.org>). We then excluded languages for which a spot-check for non-native (usually English) words in the top 100 most frequent words in the lexicon between 3 and 7 characters revealed more than 80% of words were not native. In this way, languages that used non-alphabetic scripts (like Chinese) were generally excluded since the 3-7 letter words in Chinese Wikipedia are often English. However, we included languages like Korean in which words generally consist of several characters. After these exclusions, 101 languages remained.<sup>2</sup> We analyzed the data both with and without these exclusions, and the exclusions do not significantly affect the overall direction or magnitude of the results. The final languages included 62 natural Indo-European languages and 39 non-Indo-European languages. Of the non-Indo-European languages, there are 12 language families represented as well as a Creole and 4 constructed languages (Esperanto, Interlingua, Ido, Volap) that have some speakers. (The analysis is qualitatively the same after excluding constructed languages.)

To get a sense of how clean these Wikipedia lexicons are, we randomly sampled 10 languages for which we then inspected the 100 most frequent words and an additional 100 random words to look for intrusion of English words, HTML characters, or other undesirable properties.

For the top 100 words in the lexicons of the 10 sampled languages, we found at most 3 erroneous words. For the same languages, we also inspected a randomly selected 100 words and found that the mean number of apparently non-intrusive words was 93.5 (with a range from 85 to 99). The most common intrusion in these languages was English words.

## B Appendix: Lasso regression analysis

### Lasso analysis

We fit separate lasso regressions (Tibshirani, 1996) for each length and language predicting scaled frequency from scaled orthographic probability and scaled number of minimal pairs. The lasso regression puts a constraint on the sum of the absolute value of the regression coefficients (L1-regularization) and thus effectively pushes some coefficients to 0 if they are not needed. We set the value of the lasso parameter using cross-validation to minimize the out-of-sample error and used the `lars` software package (Efron et al., 2004) in R.

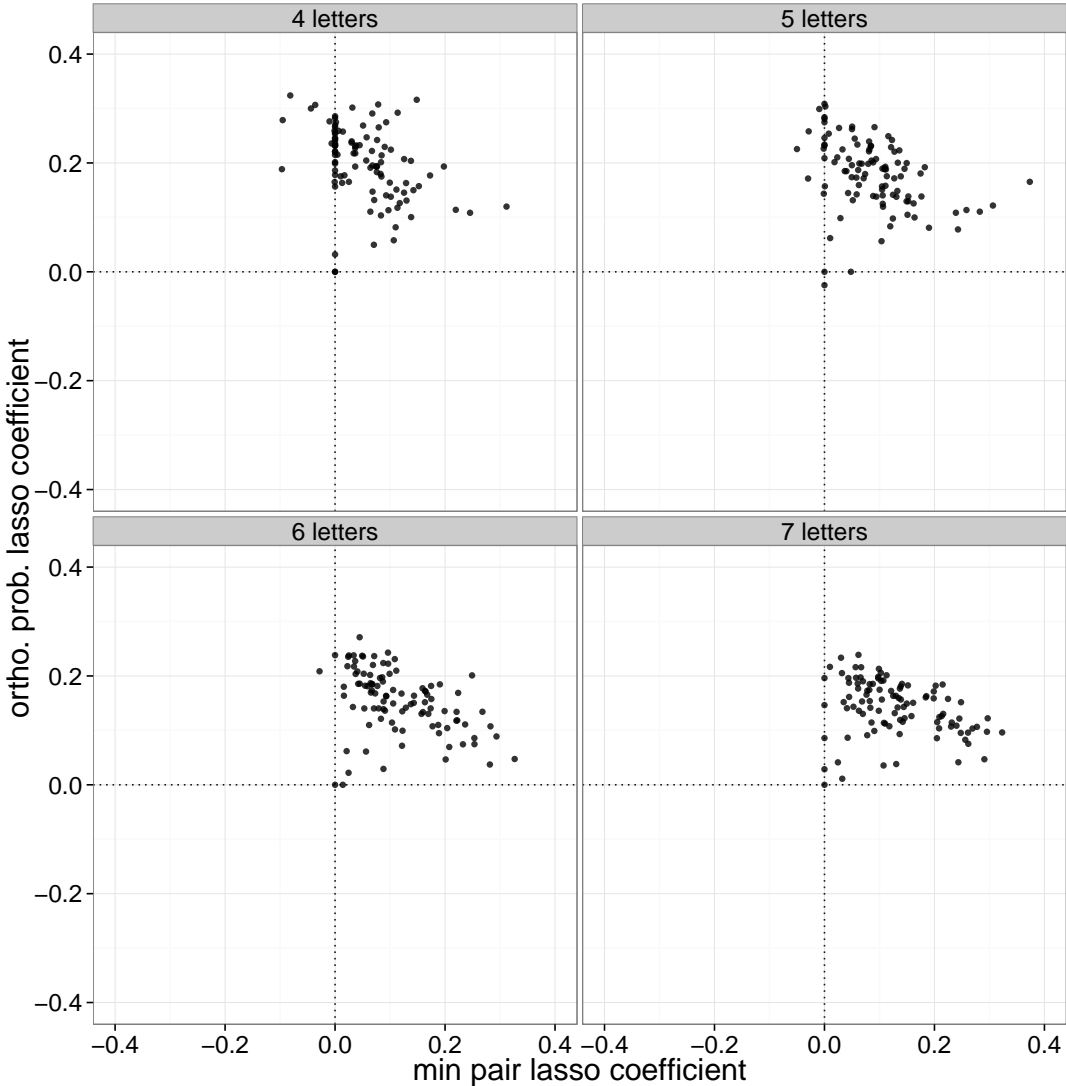
In Figure 5, for each language, we plot the coefficient for scaled number of minimal pairs on the x-axis against the coefficient for scaled orthographic probability on the y-axis. If both predictors show robust effects, we would predict the points to cluster in the upper right quadrant. If both predictors showed little effect, the points would cluster around 0.

Although the lasso regression drives some of these coefficients to 0, the plot clearly shows effects of both minimal pairs and orthographic probability (with larger coefficients in general for orthographic probability). Specifically, for 4 and 5 letter words, only 2 languages show negative coefficients for orthographic probability and none for 6 and 7 letter words.

Thus, it appears that more frequent words also have higher orthographic probability as well as more minimal pairs, although the minimal pairs effect is driven in part by orthographic probability.

---

<sup>2</sup>We excluded: Gujarati, Telugu, Tamil, Bishnupriya Manipuri, Cantonese, Newar, Bengali, Japanese, Hindi, Malayalam, Marathi, Burmese, Nepali, Kannada



**Figure 5:** This plot shows the lasso regression coefficients (predicting scaled frequency) for scaled number of minimal pairs and scaled orthographic probability.

## RESEARCH REPORT

# Cross-Situational Word Learning in the Right Situations

Isabelle Dautriche and Emmanuel Chemla

Laboratoire de Sciences Cognitives et Psycholinguistique, DEC-ENS/EHESS/CNRS, Paris, France

Upon hearing a novel word, language learners must identify its correct meaning from a diverse set of situationally relevant options. Such referential ambiguity could be reduced through *repetitive* exposure to the novel word across diverging learning situations, a learning mechanism referred to as *cross-situational learning*. Previous research has focused on the amount of information learners carry over from 1 learning instance to the next. In the present article, we investigate how *context* can modulate the learning strategy and its efficiency. Results from 4 cross-situational learning experiments with adults suggest the following: (a) Learners encode more than the specific hypotheses they form about the meaning of a word, providing evidence against the recent view referred to as *single hypothesis testing*. (b) Learning is faster when learning situations consistently contain members from a given group, regardless of whether this group is a semantically coherent group (e.g., animals) or induced through repetition (objects being presented together repetitively, just like a fork and a door may occur together repetitively in a kitchen). (c) Learners are subject to memory illusions, in a way that suggests that the learning situation itself appears to be encoded in memory during learning. Overall, our findings demonstrate that *realistic* contexts (such as the situation in which a given word has occurred; e.g., in the zoo or in the kitchen) help learners retrieve or discard potential referents for a word, because such contexts can be memorized and associated with a to-be-learned word.

*Keywords:* word learning, hypothesis testing, language acquisition, memory, lexical representation

Children observe their environment and learn the associations between word forms and their world referents. Yet, the signal is noisy: A word is not uttered in the sole presence of its referent but in a complex visual environment where multiple word-to-meaning mappings are available (Quine 1964). One possible mechanism that may reduce the referential ambiguity is *cross-situational learning*, or the aggregation of information across several exposures to a given word (Akhtar & Montague, 1999; Pinker, 1989; Siskind, 1996).

Cross-situational learning has been studied experimentally with adults and infants (K. Smith, Smith, & Blythe, 2011; L. Smith &

Yu, 2008; Trueswell, Medina, Hafri, & Gleitman, 2013; Vouloumanos & Werker, 2009; Yu & Smith, 2007). Typically, participants are asked to learn the meaning of several (up to 18) new words in situations simulating the ambiguity of the real world. For example, Yu and Smith (2007) exposed adults to a series of learning trials containing  $n$  words and a set of  $n$  possible referents. Each trial separately was thus underinformative, but toward the end of the study, participants selected the correct referent at greater-than-chance levels. Participants' success in these paradigms has been taken as evidence for an *accumulative account* of word learning (K. Smith et al., 2011; L. Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu & Smith, 2007). According to this view, each time a new word is uttered, children entertain a whole set of situationally plausible meanings and learning entails pruning the potential referential candidates as new instances of the word are made implausible by the situation. The word-meaning mapping thus starts as a one-to-many association.

Such accumulative account of word learning has recently been challenged by an alternative *hypothesis-testing account* (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell et al., 2013). Unlike the accumulative account, the hypothesis-testing strategy does not require learners to remember multiple referents for a given word. Instead, based on a single exposure to a given word, a learner selects the most plausible interpretation of this word (a process referred to as *fast-mapping*). As new information becomes available in subsequent word usages, this hypothesis may be confirmed or falsified. In the case of falsification, the old referential candidate is promptly replaced by a new one. Thus, according

---

This article was published Online First January 13, 2014.

Isabelle Dautriche and Emmanuel Chemla, Laboratoire de Sciences Cognitives et Psycholinguistique, DEC-ENS/EHESS/CNRS, Paris, France.

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n.313610 and was supported by ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC and a doctoral fellowship from the Direction Générale de l'Armement (DGA; France) supported by the doctoral program Frontières du Vivant (FdV) to Isabelle Dautriche. We thank Anne Christophe, Marieke van Heugten, Benjamin Spector, Judith Koehne, and Lila R. Gleitman for stimulating and helpful contributions and discussions.

Correspondence concerning this article should be addressed to Isabelle Dautriche, Laboratoire de Sciences Cognitives et Psycholinguistique, Ecole Normale Supérieure, 29 rue d'Ulm, Paris, France. E-mail: [isabelle.dautriche@gmail.com](mailto:isabelle.dautriche@gmail.com)

to this view, word-meaning mapping involves a one-to-one association, which continues to be updated until it reaches a stable (adult) stage. Support for such an account comes from the observation of the sequence of hypotheses learners formulate during the course of word learning. In a modification of the original experiment of Yu and Smith (2007), Trueswell and colleagues (2013) presented adults with a series of learning trials containing *one* word and *n* candidate referents and asked subjects to select the word meaning at each trial. In line with previous work, participants learned the meaning of words over the course of the study. However, contrary to previous experiments in which analyses focused on participants' final performance, Trueswell and colleagues examined participants' trial-by-trial accuracy. Crucially, they found that (a) participants persisted in their choices (e.g., if they picked dog as the meaning for the word *blicket*, they would maintain this hypothesis as long as it is confirmed by the learning situation), and (b) participants picked a new meaning hypothesis *at chance* among the available candidates otherwise (we propose a refinement of this measure below). This was taken as evidence that participants had no memory for previously seen referents beyond the one they entertained as a possible meaning, as predicted by an hypothesis-testing account.

Work on cross-situational learning has typically focused on the nature of the word-meaning mappings during the learning process. On the one hand, a complete one-to-many word-meaning mapping (following the accumulative account) seems implausible given the memory cost this presupposes. On the other hand, one-to-one word-meaning mappings (following the hypothesis-testing account) imply that a vast amount of potentially useful information is lost along the way.

In this study, we investigate one potential source of information left out by these two extreme views, the broader *context* of the learning situation, and we examine its role in constraining word learning strategies. Although naturalistic word learning environments introduce a potentially more complicated set of referent candidates that are typically eliminated in lab-based settings, this richer context may in fact contain more structure and could, as a result, help learning. That is, the set of possible referents for a word in a real learning situation is not a pseudo-random set of unrelated objects; they co-occur in the real world, and this could play an important role in cross-situational learning.

Our reasoning is best introduced with an example. In a zoo, people naturally talk about animals, whose name children may or may not know ("Do you see the *blicket* there?"; "The *dax* seems hungry today!"). An accumulative word learner would encode the full one-to-many word-meaning mapping as constrained by the situations for each occurrence of a new word (a "*blicket*" could mean lion, elephant, or monkey, and so could "*dax*," as this word has been heard in the same situation). By contrast, a hypothesis-testing learner would bind each word to one chosen referent (a "*blicket*" could mean a lion, whereas a "*dax*" could mean a monkey). In both cases, however, subsequent learning could be constrained at a different level if the learner encodes that these words were encountered in a zoo. Hence, the information that a *zoo*-word refers to an animal may persist beyond the specific situation in which it was uttered and on top of the currently entertained hypotheses. In other words, learners may encode higher order properties of situations and use it to deduce meaning

across situations ("I heard *blicket* in the zoo, it must be one of these animals . . .").

We thus propose to investigate to what extent cross-situational learning relies on context to develop word-meaning mappings. To this end, we first replicate the results of previous word learning experiments using a paradigm similar to Trueswell et al. (2013) (Experiment 1) and introduce a novel measure that quantifies the amount of information stored and retrieved across trials in such a paradigm. Second, we investigate whether introducing more ecologically valid situations would further boost memory retrieval of previously encountered referents. Specifically, we manipulate higher order properties of a word-learning situation—the semantic relation among the possible referents (Experiment 2) and context consistency (Experiment 3)—and test their effects on participants' learning strategy using the measure developed in Experiment 1. Finally, we demonstrate that if context can improve word learning, this improvement is subject to memory illusions, in a way that suggests that the learning situation itself is memorized and associated to novel words during cross-situational learning (Experiment 4).

## Experiment 1

We conducted a classical word-learning experiment using a paradigm similar to that used by Trueswell et al. (2013). Participants were exposed to a sequence of learning instances. In each instance, participants saw four images and a sentence featuring a to-be-learned word (e.g., "There is a *blicket* here"). At each learning instance, participants were asked to select a plausible referent for the word (based on the current and past information they received). The correct word referent was present in all learning instances for that word.

Our goal was to develop a measure suitable to quantify the amount of information that participants store and retrieve from a previous learning instance. Our measures differed from the one used in Trueswell et al. (2013) in two ways. First, we did not base our measure on the actual accuracy of answers but solely on their compatibility with previous learning instances. Second, we focused on learning instances of a word *W* where the referent selected in the previous learning instance for *W* is absent (and not on all cases in which this previous choice was incorrect, as Trueswell et al., 2013, did). According to the hypothesis-testing view, if participants remember only their conjecture for *W*, *these* are the cases in which they should randomly pick a novel referent among the current candidates since they cannot confirm their previous hypothesis. By contrast, if participants remember more than their single previous hypothesis for the word, their choice of a new referent should be informed by the set of referents that was present in previous learning instances.

## Method

**Participants.** Fifty adults were recruited through Amazon Mechanical Turk (22 females,  $M = 34$  years of age, 48 native speakers of English—as per voluntary answers given on a questionnaire at the end of the experiment). The experiment lasted between 5 and 10 min, and participants were paid \$0.85.

**Stimuli and design.** Twelve phonotactically legal English non-words were selected from <http://lexicon.wustl.edu/> (*blicket*,

*dax, smirk, zorg, leep, moop, tupa, krad, slique, vash, gaddle, and clup*)<sup>1</sup> as well as 12 objects representing these non-words (*cat, dog, cow, rabbit, pants, hat, socks, shirt, pan, knife, bowl, and glass*). For each of these 12 objects, five different photographs were selected. The one-to-one pairing between the 12 non-words and the 12 objects was fully randomized and differed for each participant.

The trial design follows the same constraints as that in Experiment 1 of Trueswell et al. (2013), with the exception that each learning instance contained four possible referents in our study but five possible referents in theirs. As represented in Figure 1, each trial was a learning instance for a given word, for example, *blicket*, consisting of four pictures aligned horizontally on a white background along with a written prompt “There is a *blicket* there.” The pictures were selected pseudo-randomly such that (1) the correct referent was always represented, (2) no incorrect referent occurred with a word more than twice in the experiment, (3) each object appeared the same number of times (5 times as the correct referent and 15 times as a distractor), and (4) all pictures occurred the same number of time in the experiment. There were five learning instances per word during the experiment, resulting in a total of 60 trials. The experiment consisted of five blocks each of which contained 12 trials, one for each to-be-learned word. The list of 12 words occurred in the same order in each of the blocks.

**Procedure.** Participants were tested online. They were instructed that they were to learn words by associating them with images displayed on the screen. Prior to test, participants were given a screenshot of a learning instance involving a word and a set of pictures that were not used at test. No information about the number of to-be-learned words or the number of learning instances was given. For each trial, participants were asked to click on the image they believed could represent the meaning of the word. Once they responded, the test continued with the next trial. We recorded participants’ answers at each trial as well as their response times.

**Data processing.** Five participants were excluded from our analysis for obvious violations of the instructions (two always selected the left image, three had reaction time [RT] patterns indicating that they were 5–10 times faster in the last block than in the first and second block—including these participants in the analyses does, however, not impact the pattern of results). We also removed five responses out of 3,000 for being implausibly fast (below 1 s) or slow (above 30 s; following K. Smith et al., 2011). Participants who provided 50 or fewer responses out of 60 were discarded (but this criterion did not eliminate any participants in this first experiment).

**Data analysis.** Participants’ responses were coded as 0 (incorrect) or 1 (correct) for each trial. Since we analyzed categorical responses, we modeled them using logit models as recommended by Jaeger (2008). We ran mixed model analyses using R 2.15 and the lme4 package (Bates & Sarkar, 2007); plots have been realized using the ggplot2 package (Wickham, 2009). Beta estimates are given in log-odds (the space in which the logit models are fitted), with the odds of an event defined as the ratio of the number of occurrences where the event took place to the number of occurrences where the event did not take place. Significant positive beta estimates indicate an increase in the log-odds, and hence an increase in the likelihood of occurrence of the dependent variable with the predictor considered (calculated using the inverse logit

function [ $\text{logit}^{-1}$ ]). We computed two tests of significance: (a) the Wald’s  $Z$  statistic, testing whether the estimates are significantly different from 0, and (b) the  $\chi^2$  over the change in likelihood between models with and without the considered predictor. Since the results did not change between the two tests, we report the  $Z$  statistic only.

The random effect structure chosen for each model is the maximal random effect structure justified by model comparison and supported by the data. We followed the procedure outlined in Baayen, Davidson, and Bates (2008), starting with the full random effect structure and reducing the structure on a step-by-step basis until excluding a random term resulted in a significant decrease of the log-likelihood compared to the model including it. For the sake of clarity, the chi-square comparisons between models are not reported.

## Results and Discussion

We report three analyses looking at (1) the learning curve; (2) accuracy as a function of the previous response, following Trueswell et al. (2013); and (3) a novel measure characterizing information retrieval from prior experience.

**1. Learning curve: A replication.** Figure 2 presents participants’ accuracy in each block. We modeled the accuracy with a mixed logit model using a predictor Block (1–5) with subjects and words as random effects on intercepts plus a random slope for the effect of Block with subjects. We found a significant effect of Block on accuracy ( $\beta = 0.36, z = 10.25, p < .001$ ). The beta coefficient indicates that for every new block, participants were 59% ( $\text{logit}^{-1}[0.36]$ ) more likely to be accurate than in the previous block. We thus replicate previous findings showing that participants gradually learned word-meaning mappings across learning instances (Yu & Smith, 2007; Trueswell et al., 2013).

**2. Trial-by-trial analysis: Accuracy dependent responses.** Using Trueswell et al.’s (2013) analysis on participants’ responses, we compared the average proportion of correct responses in Blocks 2–5 depending on whether the previous referent selection for that particular word was correct or incorrect (see Figure 3). We modeled the proportion of correct responses using a predictor Previous Response Accuracy (Correct vs. Incorrect) with subjects and words as random effects on intercepts and a random slope for the effect of Previous Response Accuracy with subjects. We applied an offset corresponding to the logit of the chance level to the model (i.e., .25, the probability of being correct in a trial) to compare the intercept against chance level. We found a main effect of Previous Response Accuracy ( $\beta = 1.40, z = 10.94, p < .001$ ), showing that participants were 80% ( $\text{logit}^{-1}[1.40]$ ) more likely to be accurate when they were correct on the previous learning instance than when they were incorrect.

We then compared participants’ average accuracy against chance level separately depending on whether their previous response was correct or incorrect. We found that (a) participants’ accuracy was significantly above chance when they had

<sup>1</sup> As one reviewer pointed out, three of these words are actually real words: *smirk, leep, and slique* (although the latter two are spelled differently). However, because accuracy was not predicted by word type (non-word vs. real words:  $z < 1, p > .4$ ), it is unlikely that *only* the small group of real (but infrequent) words induced the observed results.



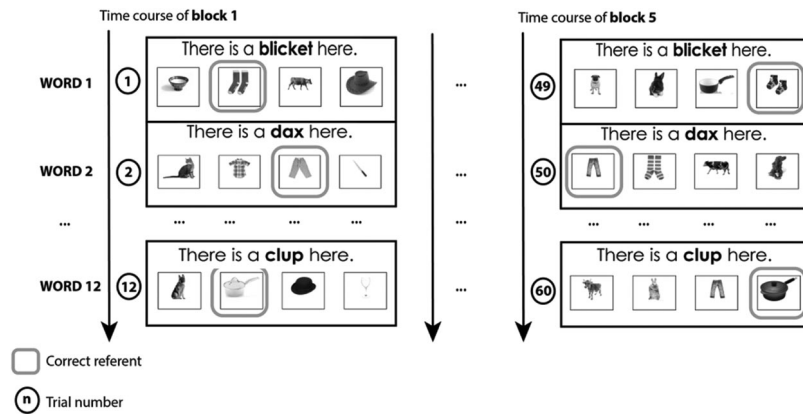


Figure 1. Experimental design. A learning trial of a to-be-learned word is a set of four candidate referents presented with the word in a simple declarative sentence. The five learning instances for each word are distributed in five blocks such that there is exactly one learning instance for a given word per block—hence, 12 trials per block. As depicted, each block is an ordered list of 12 trials, such that there are exactly 11 intervening trials between two learning instances of the same word. This resulted in a total number of 60 trials. The word–referent pairings were randomly assigned for each participant.

been correct in the previous learning instance for that word (789 data points;  $\beta = 3.13$ ,  $z = 12.10$ ,  $p < .001$ ), and (b) accuracy also exceeded chance after being incorrect in the previous trial (1,339 data points;  $\beta = 0.33$ ,  $z = 3.48$ ,  $p < .001$ ).

While (a) aligns nicely with the results from Trueswell et al. (2013), (b) does not. Instead, Trueswell et al. found that after an incorrect response, participants were at chance in the next learning instance.

The apparent difference between our results and Trueswell et al.’s (2013) results could be explained when one takes into account that the current analysis collapses two situations for which the hypothesis-testing strategy predicts different behaviors: (I) if the participant’s previous selection is present, participants should repeat their incorrect previous hypothesis, and (II) if it is not present, participants should be at chance in selecting the correct referent. Hence, the outcome of this analysis is dependent on the proportion of instances of Types I and II.

Both Trueswell et al.’s (2013) first experiment and the present experiment are constrained in the same way: No object can be repeated more than twice as a distractor for a given word. However, in Trueswell et al., each trial displayed five possible referents (in contrast to the four referents displayed here); hence, objects had to be repeated more often as distractors to account for the additional fifth picture on each trial. While both occurrences for a given distractor are not necessarily in two subsequent trials for a given word, there should be a higher proportion of instances of Type I in Trueswell et al.’s study than in the present experiment (12% of Type I instances on the total number of trials where the previous choice is incorrect). Since Type I trials lead to incorrect responses, this difference could explain why the analysis reveals better results for the current experiment.

**3. New analysis: A measure of information retrieval.** To distinguish learning strategies based on one-to-one and one-to-many word-meaning mappings, we need to quantify the amount of information stored and retrieved at each learning occasion

during cross-situational learning. In the following, we propose such a measure.

We selected from Block 2 all learning instances of Type II, that is, learning instances for a word  $x$  in which the participant’s choice for  $x$  from Block 1 is not present. Figure 4 represents a measure of selecting a response that is informed by previously seen referents. Specifically, for each trial, we computed the set  $S$  of referents that were also present in the first block for this word. Figure 4 represents the proportion of responses that belong to  $S$  minus the expected proportion of falling in  $S$  by chance (cardinal of  $S$  divided by 4). We modeled the proportion of responses that belong to  $S$  with subjects and words as random effects on intercepts and applied an offset corresponding to chance to the model. Note that chance level of selecting a referent present in the previous learning instance is now trial dependent (1, 2, or 3 images could be repeated from the previous trial); hence, the offset applied to each trial was the logit of the corresponding chance level of selecting a previously seen referent (.25, .50, or .75).

The resulting measure significantly exceeds zero (336 data points;  $\beta = 0.27$ ,  $z = 2.28$ ,  $p < .05$ ); that is, participants were more likely than chance to select a previously seen referent. This result is not specific to Block 2: While considering all learning instances where participants’ previous choice was not present, the measure also significantly exceeded zero (1,172 data points;  $\beta = 0.27$ ,  $z = 3.77$ ,  $p < .001$ ). This analysis shows that in our paradigm, participants store more than a single hypothesis for the meaning of a word. Specifically, we show that participants resorted to previously encountered, but not chosen, referents in cases where their previous hypothesis is irrelevant.

Although Trueswell et al. (2013) did not employ this analysis, one would expect to find the same result in one of their experiments: their Experiment 3. In this experiment, participants were presented with only two objects on the screen at a time, and, crucially, no single object was used twice as a

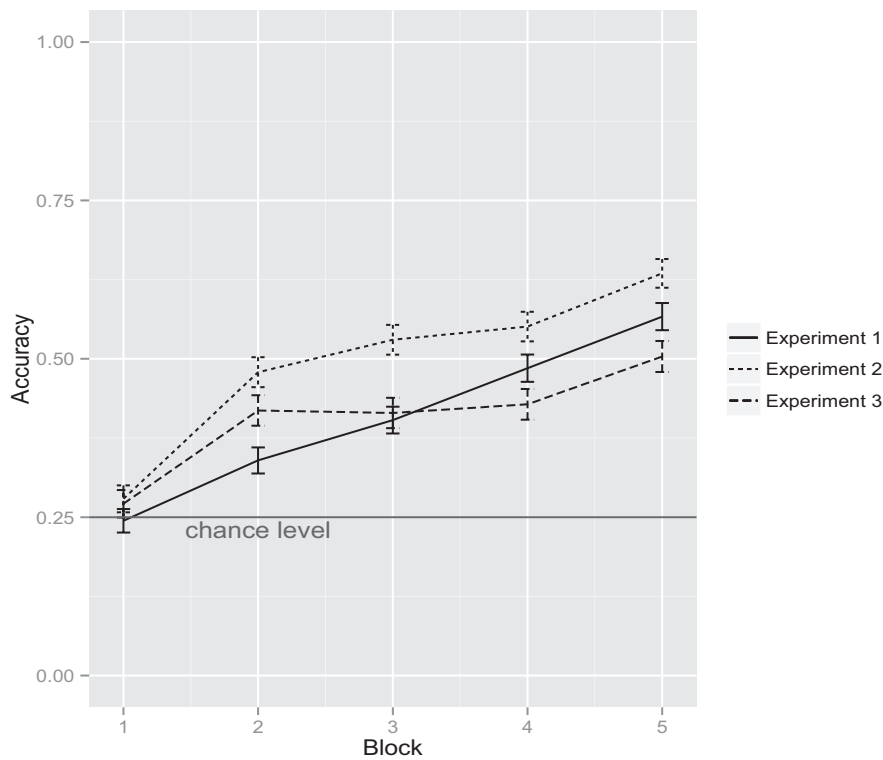


Figure 2. Learning curves. Average accuracy aggregated by subject for each block in Experiments 1–3. Error bars indicate the standard error of the mean.

distractor for a given word. Hence, the accurate answer (not selecting the distractor) corresponds to the only response that is fully coherent with previous learning instances (since the distractor was never previously presented with this word). The two measures are thus merged here. Yet, Trueswell et al. did not find improved accuracy following an incorrect selection. To explain the absence of evidence for accumulative learning in Trueswell et al.'s Experiment 3, one could think about reasons why two-object and four-object trials trigger different strategies. It is possible that participants' strategy depends on a tradeoff between the cost and the incentive to remember more than a single conjecture for a word in a given experimental situation. While memorizing two possible referents is easier than memorizing four possible referents, it is not clear that there is a real advantage of doing so to succeed in the task. Remembering only the object guessed means remembering 50% of the whole scene in the two-object trial—hence, an already quite high probability of success in the next trial (where chance is already at 50%). While the cost of remembering the objects may be higher in the four-object trial, there would also be more incentive to do so given the higher ambiguity following the lower probability of success (chance is at 25%, so it may be worth investing resources into enhancing this probability). Albeit speculative, superficial aspects of the experimental situation could thus in principle alter participants' strategy. We leave the exploration of this issue for future research. Our current goal is to investigate the effect of context on prior

experience retrieval, and we will do so with the novel, more restrictive measure we proposed.

**4. Control analysis: Participants' strategy in online versus in lab experiments.** So far, the discussion has not considered the possibility that there could be more fundamental differences between Trueswell et al.'s (2013) paradigm and ours. For instance, our participants were not present and monitored in the lab. It is thus possible that they completed the task in a different way (e.g., taking notes) and that their performance would therefore not reflect the natural learning ability. To assess this possibility, we analyzed participants' response times, and we gathered more information about our population in a replication of Experiment 1.

1. In Experiment 1, participants took on average 5,323 ms ( $SE = 68$ ) to associate a meaning to a word, making it unlikely they took notes. More objectively, a linear regression on the participants' accuracy in the final block using average RT throughout the experiment as a predictor did not reveal any effect of RT on accuracy ( $z = -1.24, p > .2$ ). This suggests that there is no division within the population between participants who would have taken notes (thus being slow and accurate) and those who would not have taken notes (thus being relatively fast and inaccurate).

2. The same experiment was administered to 30 new participants recruited in exactly the same way from the same population. The crucial difference was the addition of a question at the end of the final questionnaire: "Did you take notes during



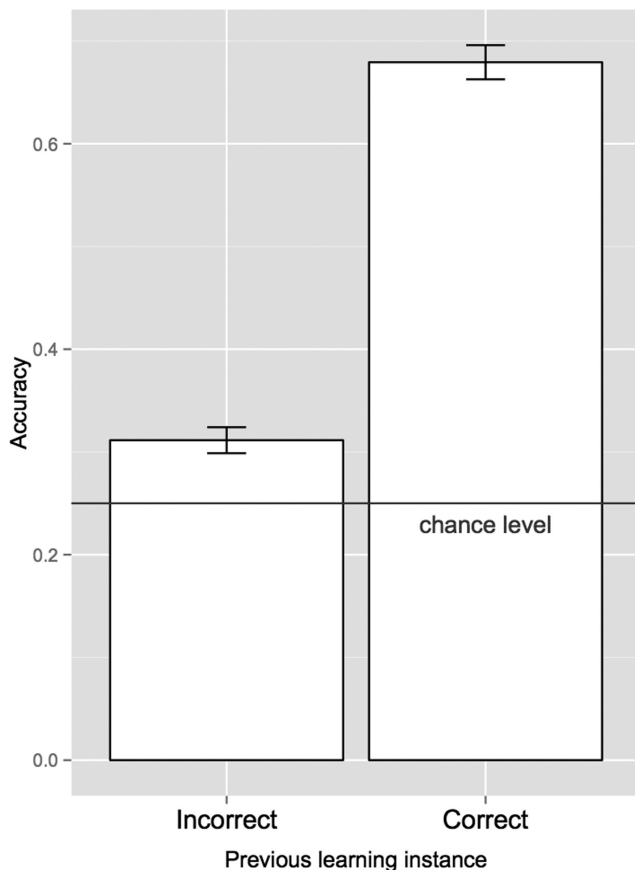


Figure 3. Accuracy dependent measure. Accuracy in Blocks 2–5 for previously correct or incorrect words in Experiment 1. Error bars indicate the standard error of the mean.

the task?” Among the 28 participants who finished the task, none of them reported taking notes, suggesting that the new participants performed the task in the appropriate way. The results of this control experiment patterned with those of Experiment 1 on the three analyses that were conducted.<sup>2</sup> This suggests that the methodology used in Experiment 1 corresponds to the type of cross-situational learning exercise we are interested in.

**Summary.** Our results provide evidence that participants store more than simple one-to-one word-meaning mappings. In the next experiments, we investigate whether external constraints on simultaneously presented referents for a word can alter prior information retrieval.

### Experiment 2: Encoding Semantic Relation

We adapted Experiment 1 to evaluate one of such contextual constraint: the semantic relation among the possible referents. We modified the first block such that all four pictures on each trial corresponded to one of the following natural categories: animals (dog, cat, rabbit, cow), dishes (pan, bowl, knife, glass), clothes (pants, socks, shirt, hat). For instance, if *blicket* referred to a dog, the three other distractor images it co-occurred with were *all* possible animal referents, mimicking a zoo-context. Furthermore,

words belonging to a given category were presented on consecutive trials (allowing the learner to first learn words related to the zoo, and then words related to a bedroom and so on). By imposing these constraints on the situation of the first learning instance, we hope to reduce the overall memory cost for encoding the situation and thus improve cross-situational learning. As a consequence, we expect an increase in performance in the second learning instance for Experiment 2 compared to Experiment 1.

### Method

**Participants.** Forty adults were recruited from Amazon Mechanical Turk (25 females,  $M = 40$  years of age, 37 native speakers of English). Two participants were excluded from our analysis because over 20% of their responses fell outside the 1–30-s response time window (see Experiment 1—Analysis).

**Stimuli and design.** The stimuli and the design were the same as in Experiment 1 except for new constraints on the first block of learning instances (see Figure 5 for a schematic description): (1) on all trials of the first block, each word was presented along with distractors from the target object category (*animals*, *clothes*, or *dishes*), and (2) the words from a given category were presented in consecutive trials.

**Procedure and analysis.** The procedure and analysis are identical to those in Experiment 1.

### Results

We replicated the two main results of Experiment 1. First, we modeled the accuracy with a mixed logit model using a predictor Block (1–5) with subjects and words as random effects on intercepts and a random slope for the effect of Block with subjects (Model 1). Participants demonstrated a gradual learning of word-referent pairs across learning instances, as evidenced by a significant effect of Block on accuracy (see Figure 2;  $\beta = 0.39$ ,  $z = 7.12$ ,  $p < .001$ ). Participants were 60% ( $\text{logit}^{-1} [0.39]$ ) more likely to be accurate than in the previous block. Second, we modeled the measure defined in Experiment 1 with subjects and words as random effects on intercepts (Model 2). Participants stored more information during the first exposure of the word than expected by chance (see Figure 4; 223 data points;  $\beta = 1.20$ ,  $z = 6.37$ ,  $p < .001$ ).

We compared Experiment 1 and Experiment 2 along these two dimensions. First, we modeled participants’ accuracy in Blocks 1

<sup>2</sup> Regarding the learning curve, we modeled the accuracy with a predictor Block (1–5) and a predictor Experiment (Experiment 1, Control) with subjects and words as random effects on intercepts. There was no effect of the predictor Experiment ( $z < 1$ ,  $p > .4$ ), showing that the learning curves were similar.

Furthermore, accuracy was modeled after an incorrect response with a predictor Experiment (Experiment 1, Control) with subjects and words as random effects on intercepts and an offset of the chance level. There was no effect of the predictor Experiment ( $z = -1.5$ ,  $p > .1$ ), showing that control participants’ accuracy after an incorrect response was not different from earlier participants’ ( $M_{\text{Control}} = 0.32$ ,  $SE_{\text{Control}} = 0.02$ ;  $M_{\text{Exp.1}} = 0.31$ ,  $SE_{\text{Exp.1}} = 0.02$ ).

Finally, we modeled our measure of information retrieval with a predictor Experiment (Experiment 1, Control) with subjects and words as random effects on intercepts and an offset of the chance level, and we found no difference between the Control and Experiment 1 ( $z = 0.2$ ,  $p > .8$ ;  $M_{\text{Control}} = 0.06$ ,  $SE_{\text{Control}} = 0.02$ ;  $M_{\text{Exp.1}} = 0.06$ ,  $SE_{\text{Exp.1}} = 0.01$ ).

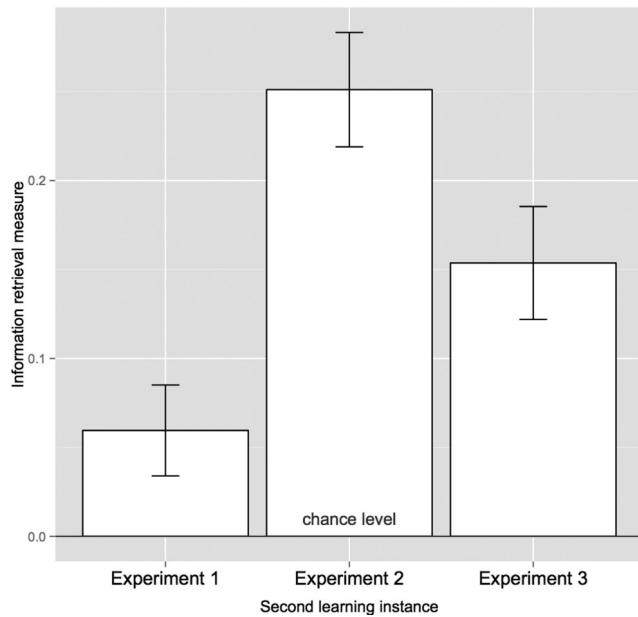


Figure 4. Information retrieval measure. Corrected tendency to select a previously seen referent: average for all second learning instances of 1 or 0 (depending on whether the answer was in the previous learning instance) minus the chance of selecting a referent present in the previous learning instance. Error bars indicate the standard error of the mean.

and 2 for these two experiments similarly to Model 1 but applied to the results of both experiments at once and with an additional predictor Experimental condition (Experiment 1, Experiment 2) and its interaction with Block (1 vs. 2). We restricted the comparison to Blocks 1 and 2 to ensure that distance or performance at or near ceiling would not mask the effect of Block 1. As discussed above, we observed a significant effect of Block on accuracy. In addition, we also observed a significant interaction between Block and Experimental condition (see Figure 2;  $\beta = 0.43$ ,  $z = 2.11$ ,  $p = .03$ ). Second, we modeled our measure of information retrieval for Experiments 1 and 2 similarly to Model 2 with a predictor Experimental condition (Experiment 1, Experiment 2). Our information retrieval measure shows that participants in Experiment 2 were significantly more likely than participants in Experiment 1 to resort to previously encountered, but not selected, referents (see Figure 4;  $\beta = 0.90$ ,  $z = 4.32$ ,  $p < .001$ ). The probability of choosing a previously encountered referent increased by 71% ( $\logit^{-1}[0.90]$ ) in Experiment 2 compared to Experiment 1.

## Discussion

The comparison between Experiment 1 and Experiment 2 shows that providing learners with an opportunity to rely on higher-order properties of situations allowed them to resort to previously encountered experience more efficiently than participants who were exposed to artificial, randomly assembled situations (Experiment 1).

As expected, richer contextual information boosted participants' use of a cross-situational learning strategy. There are three possible interpretations for this result. (1) *Context consistency and memory*: Participants used contextual information to inform their word learning strategy. We come back to this issue in Experiment 4, but

it is important to note that there are two possible explanations for such an effect. First, in a one-to-many mapping approach, temporary lexical entries may be easier to memorize if the multiple potential referents for a word are semantically coherent. Second, it could be that contextual information is stored as an independently accessible source of information: Participants may memorize associations between a word *and* situations in which it was uttered, and these situations could directly inform word-meaning mappings in subsequent learning instances. (2) *A closest-match strategy*: Participants follow a hypothesis-testing strategy, but when their current hypothesis is absent from the picture display, they resort to the closest match. Concretely, if their current hypothesis is that *blicket* means *dog*, but no dog is present in the display, learners would not randomly select any other possible meaning but would rather select the closest match, which in this experiment will be another animal. (3) *Partial representations*: Participants entertain partial representations: They may encode the semantic category of a word (e.g., *animal*) in the same way as they may encode the grammatical features of this word (e.g., syntactic category, gender, animacy, etc.) without encoding any meaning hypotheses. In following learning instances, participants would then (randomly) select one member of the encoded category.

Hypotheses 2 and 3 contrast with Hypothesis 1, as for these two options, participants would thus select a distractor from the correct category more often than chance, but this would not be mediated by memory for the previous learning situation itself. Experiment 3 disentangles between these possible interpretations of the improvement observed between Experiments 1 and 2.

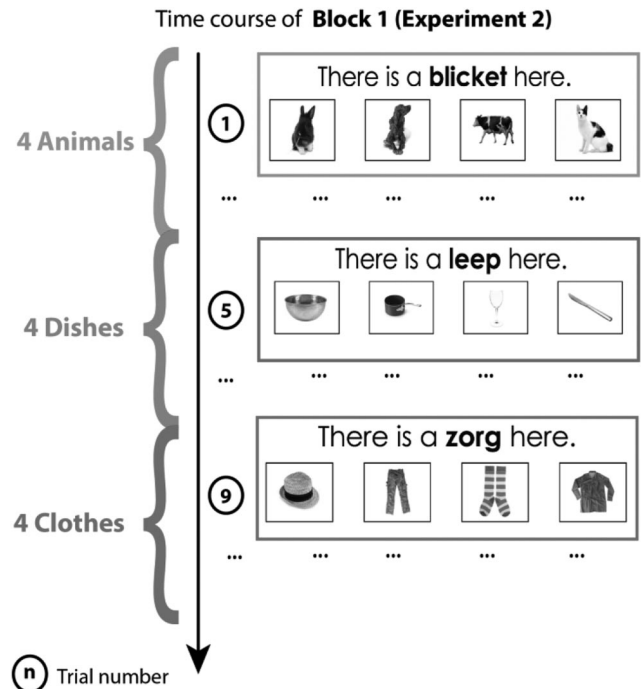


Figure 5. An example of the trial presentation in Block 1 for Experiment 2. Adults saw 12 trials, one for each to-be-learned word, such that all objects in one trial were from the same natural category of the referent (animal, cloth, dish). All words referring to objects from the same natural category appeared in succession.

### Experiment 3: Encoding Context Consistency

Experiment 3 was set up to replicate Experiment 2 with three *artificial* categories of objects with no a priori coherence (e.g., {apple, dog, flower, hat}) instead of “natural” categories. Note that despite the lack of semantic coherence among these objects, categories could nonetheless emerge here due to the repeated and consecutive co-occurrence of the four objects that constitute each of them.

Although these categories are clearly artificially induced, the process of category induction may in fact not be unnatural. Specifically, under the right circumstances, many sets of objects, however unrelated they may appear to be, can co-occur. For example, in the kitchen, you may simultaneously see an apple, a dog, a vase with flowers, and a hat hung on the wall. These items are not transparently related, but all of them may be simultaneously found in the kitchen, possibly for quite different reasons. Thus, in the absence of semantic relations between the objects of an *artificial* category in Experiment 3, the coherence may be induced by their co-occurrence on four consecutive trials (once for each of the word that refers to them). The consistent display here plays the role of the kitchen in the example above.

If participants fail to use this contextual consistency, then they should behave like participants in Experiment 1, where none of the objects within a trial was semantically related to the other. This would favor Hypotheses 2 and 3, which attribute the improvement observed in Experiment 2 to semantic consistency. By contrast, if the artificial categories improve cross-situational learning compared to Experiment 1, this would favor Hypothesis 1, which relies on consistency in general and not on a tendency to resort to a semantically close selection (as of Hypothesis 2) or on partial representations (remembering “*animal*” instead of specific animals; as per Hypothesis 3).

### Method

**Participants.** Forty adults were recruited from Amazon Mechanical Turk (12 females,  $M = 34$  years of age, 39 native speakers of English). Four participants were excluded from our analysis because over 20% of their responses fell outside the 1–30-s response time window (see Experiment 1—Analysis) ( $n = 2$ ) or because they participated in previous experiments ( $n = 2$ ).

**Stimuli and design.** The design was similar to Experiment 2. We used a novel set of objects in order to minimize the potential semantic associations among them within each of the three artificial categories. Categories were defined as follows: {apple, dog, flower, hat}, {pants, chair, pan, teddy bear}, and {leaf, snake, watch, book}.

The first block follows the same design as in Experiment 2, but the position on the screen for each object within the trials of the same category was fixed. For example, considering the set {apple, dog, flower, hat}, these objects appeared in the same position (albeit with different images) on the screen in all four learning instances for the four target words associated with them. This should raise the awareness that the situation is constant. Thus, a dog might be the left-most object for four consecutive trials, but the image used on each trial will change.

**Procedure and analysis.** The procedure and analysis are identical to those in Experiments 1 and 2.

### Results

We replicated the two main results of Experiments 1 and 2. First, we modeled participants’ accuracy in Experiment 3 with a mixed logit model using a predictor Block (1–5) with subjects and words as random effects on intercepts and a random slope for the effect of Block with subjects (Model 1). Participants demonstrated a gradual learning of word-referent pairs across learning instances as evidenced by a significant effect of Block on accuracy (see Figure 2;  $\beta = 0.20$ ,  $z = 3.48$ ,  $p < .001$ ). Second, we modeled our measure of information retrieval in Block 2 with subjects and words as random effects on intercepts (Model 2). Participants retrieved more information from the first exposure to a word than expected by chance (see Figure 4; 231 data points;  $\beta = 0.68$ ,  $z = 4.68$ ,  $p < .001$ ).

We compared the three experiments along these two dimensions. First, we modeled participants’ accuracy in Blocks 1 and 2 for the three experiments with the predictor Block (1 vs. 2) used in Model 1 and an additional predictor Experimental condition (Experiment 1, Experiment 2, Experiment 3) and its interaction with Block. There was no significant interaction between Block and Experimental condition (Experiment 3 vs. Experiment 1:  $\beta = -0.20$ ,  $z = -1.02$ ,  $p = .3$ ; Experiment 3 vs. Experiment 2:  $\beta = 0.21$ ,  $z = 1.02$ ,  $p = .3$ ; see Figure 2). Second, we modeled our information retrieval measure in Block 2 for the three experiments similarly to Model 2 with a predictor Experimental condition (Experiment 1, Experiment 2, Experiment 3). Participants in Experiment 3 were significantly less likely to choose a previously seen, but not selected, referent than participants in Experiment 2 ( $\beta = 0.47$ ,  $z = 2.14$ ,  $p < .05$ ), but they were significantly more likely to do so than participants in Experiment 1 ( $\beta = -0.42$ ,  $z = -2.11$ ,  $p < .05$ ; see Figure 4).

Overall, these results demonstrate that participants in this experiment retrieved the systematic co-occurrence of seemingly unrelated objects to degree intermediate between participants in Experiments 1 and 2. This shows that participants use contextual information from consistent contexts to inform word learning, and they do so to a greater extent if contexts furthermore share a semantic relation.

### Discussion

Participants used the artificial categories presented in the first learning instance to guide their choice of the word’s referent in subsequent instances. Crucially, this effect was preserved even though none of the objects presented in the first learning instance shared a “natural” property. This rules out the possibility that the results from the previous experiment could be due entirely to an under-specification of a selection (e.g., *animal* instead of *dog*) or to a tendency to resort to a semantically close choice (e.g., from *dog* to *cat*) when the previously hypothesized referent was not available. Instead, our results favor the hypothesis that contextual consistency helps encoding situations in both Experiments 2 and 3.

Nonetheless, participants in Experiment 3 were less likely to resort to previously encountered referents than participants in Experiment 2. One reasonable explanation may be that encoding an artificial relation is more demanding than encoding a natural relation: While participants in Experiment 2 could remember a label readily available to characterize the relation among objects (“animal,” “clothes,” or “dishes”), participants in Experiment 3

had to encode the category as a plain list of objects. Hence, learners may have encoded contextual information in both experiments, but the format of the relevant information varies from one experiment to the other, and this could recruit different memory resources.

Experiment 3 showed that contextual consistency, and not only semantic consistency, helped learners resort to possible word meaning hypotheses. However this effect could be explained by two possible representations of context in memory. (a) *Internal to word-meaning mappings*: One-to-many word-meaning mappings may be more or less easier to remember, and a coherence between the possible meanings may indirectly boost an active memory for these mappings. As a result, multiple hypotheses for a word are better remembered when these hypotheses form a coherent group, but context is not necessarily stored in memory as such. (b) *External to word-meaning mappings*: Contextual information could be directly accessible as an independent source of information, that is, learners could remember the situation in which they heard a word *in addition* to the single or multiple hypotheses they entertain for this word. In this case, contextual information can be used actively to constrain subsequent learning instances.

Experiment 3 did not distinguish between an internal versus an external representation of context since contextual representation was confounded with word-meaning representations. In Experiment 4, we propose to disentangle these two possibilities and assess whether context is represented *per se* in memory.

#### Experiment 4: Context Representation in Memory

Experiment 4 investigates whether the effect of context observed in Experiments 2 and 3 is the result of an internal or an external representation of context. Much like Experiments 2 and 3, objects in the first block of this experiment were grouped into three sets. Two of these sets contained objects from a single natural category (animals and clothes) as in Experiment 2, henceforth, “natural sets.” By contrast, the third set was hybrid: It contained two (new) animals and two (new) pieces of clothes. For a word whose referent belongs to a natural set, participants could encode a natural category (e.g., animal), as in Experiment 2. However, some objects from this natural category occurred in the hybrid set and should not be considered possible referents for this word after the first learning instance (contrary to Experiment 2).

We propose to reproduce a memory illusion effect identified in earlier work (Roediger & McDermott, 1995) showing that participants asked to remember a list of words are likely to *mis-report* a word as being part of this list if there is a natural relation between the word and the list. For example, participants incorrectly recall the word *sleep* as a member of a list such as *bed, pillow, night*. Applied to our word learning task, lists can be thought of as sets of objects seen in the first block (e.g., *dog, cat, snake, cow*). If context is encoded as an additional source of information (Hypothesis b), participants are in the same situation as in the memory illusion experiment, and we expect to reproduce the same illusion. Participants should be more likely to map a target word in the natural sets onto a distractor object from the appropriate natural category than from the other category. However, crucially, this bias for the appropriate category should be observed even when we compare *only* distractors from the hybrid set, which had never appeared with the target word before. If context is not encoded

independently and the effect occurs at the level of the lexicon (Hypothesis a), then there is no immediate expectation with respect to this illusion.

#### Method

**Participants.** A total of 119 adults were recruited from Amazon Mechanical Turk (47 females,  $M = 36$  years of age, 116 native speakers of English). Twenty-four participants were excluded from our analysis because they participated in previous experiments ( $n = 14$ ), because they indicated that they took notes during the task ( $n = 4$ ), or because their RT patterns were highly irregular, in a fashion similar to participants who indicated that they took notes (e.g., 5–10 times faster from one block to another;  $n = 6$ ).

**Stimuli and design.** The design was similar to Experiment 2. We formed two natural sets—animals {*cat, cow, snake, rabbit*}, and clothes {*pants, tie, hat, socks*}—and one hybrid set of images mixing objects from each natural category: {*dog, rat, shirt, shoe*}. The hybrid set served as a reservoir of objects that could reveal the illusion when used as distractors. The hybrid set was always presented first.

We generated the learning trials following the constraints described in Experiment 1. However, our planned analysis focused on responses in the second block such that the target would be from one of the natural sets but responses would be a distractor from the hybrid set H. Hence, in order to have more data points of interest, we assigned the learning instance with the maximal number of distractors belonging to H (among the four learning instances otherwise distributed randomly in Blocks 2–5) to the second block. To limit the frequency of objects from H in Block 2, we did this for target objects in natural sets, but the opposite for target objects in H (which trials were not of interest). As a result, participants saw on average five instances of the objects in the hybrid set during Block 2 (instead of four instances before).

**Procedure and analysis.** The procedure and analysis are identical to those in Experiment 1, 2, and 3.

#### Results

We selected learning instances from Block 2 for words belonging to the two natural sets of objects. We looked at the artificial set of objects to compare the proportion of responses that belong to the set S of distractors from the same category and to the set D of distractors from a different category. Figure 6 shows the proportion of responses in S and D minus the probability of selecting them by chance (the cardinal of S and D divided by 4). Note that we selected trials where neither set S nor set D were empty (482 data points; chance level for S or D was either .25 or .50).

We modeled the proportion of responses in the artificial set of objects by a predictor Distractor type (Same category vs. Different category). Observations of the results led us to add a predictor Semantic category (Animals vs. Clothes) to the model, as well as its interaction with Distractor type. The random structure included subjects and words as random effects on intercepts and no random slope was justified. We applied an offset corresponding to chance to the model.

We observe a main effect of Distractor type ( $\beta = 0.45$ ,  $z = 2.7$ ,  $p < .01$ ), showing that participants were 61% ( $\text{logit}^{-1} [0.45]$ )



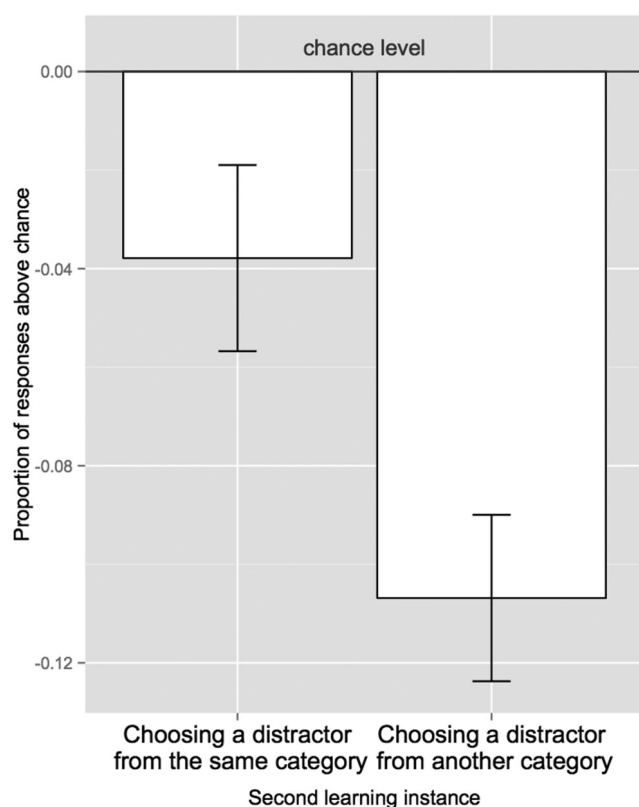


Figure 6. Experiment 4. Proportion of responses falling in the hybrid set as whether responses are from the semantic category of the target word (left bar) or from another category (right bar) minus the probability of selecting them by chance. Error bars indicate the standard error of the mean.

more likely to choose a distractor object from the same category as the target than from another category, even if this object did not co-occur with the word in the previous trial.<sup>3</sup>

## Discussion

Participants are more likely to select a distractor from the semantic category of the target than a possible referent from another category. Crucially, this effect occurs even though none of the distractors were present in the first learning instance for that word. This illusion is consistent with previous findings both in word learning task (Koehne & Crocker, 2011) and in memory tasks (Roediger & McDermott, 1995) and, thus, provides an indirect argument for the fact that learning situations are stored in memory per se.

Our results thus suggest that a situation in which a novel word occurs can be stored and bound to this word during word learning. Others have argued that, even in adults, the information that is retrieved about a word is the accumulation of all the situations in which that word has been encountered (Perfetti & Hart, 2002). Although our results are compatible with such a proposal, they are at present restricted to cross-situational word learning stages and provide no evidence that early word representation is the set of learning contexts in which this word was encountered. In the

General Discussion, we discuss the broader implications of our results for the role and representation of contextual information during the development of lexical representations.

## General Discussion

The present article examined the impact of the context on word learning mechanisms. In four experiments, we showed that learners can simultaneously retrieve multiple candidates for the meaning of a word and that manipulating the contextual properties of the set of plausible candidates could boost the amount of information retrieved. Specifically, our results show that cross-situational learning benefits from higher-order properties of a word-learning situation: the semantic relation between the possible referents (Experiment 2) as well as contextual consistency (Experiment 3). Moreover, this effect is subject to memory illusions, in a way that suggests that the effect of context found above is the result of an attempt to store contextual information directly in memory (Experiment 4).

## Learning Strategies

Most of the accounts of cross-situational learning have concentrated on the amount of information the learner stores for each learning instance. We introduced two learning strategies at the opposite end of the continuum: an accumulative learning account, in which the learner encodes one-to-many word-meaning mappings, and a hypothesis-testing account, in which the learner remembers a single word-meaning association. While computational models have emphasized the importance of defining the number of hypotheses entertained at each point in time (Yu & Smith, 2012), we add a new parameter showing that learners could also encode a different kind of information, context, to increase the amount of prior experience they could retrieve. Our results argue against an extreme version of the hypothesis-testing account where learning operates *only* through a single hypothesis for each word. Instead, we suggest that cross-situational learning is informed by the type of learning context.

One may imagine other learning strategies in more intermediate continuum positions to accommodate the finding that learners encode more than a single meaning hypothesis. For instance, Koehne, Trueswell, and Gleitman (2013) proposed multiple-hypothesis tracking strategy, according to which learners may memorize not only one hypothesis, but all past hypotheses for a given word. Previous research on cross-situational learning has also suggested that learners do not attend equally to all possible meanings for a word and use several additional strategies to prune the set of possible meanings (mutual exclusivity: Yurovsky & Yu, 2008; attention to stronger associations: Yu & Smith, 2012).

Overall, investigations about word learning strategies concentrated on the possible forms of relations learners could entertain

<sup>3</sup> Additionally there is a significant interaction between Distractor type and Semantic category ( $\beta = 1.07$ ,  $z = 3.19$ ,  $p < .01$ ). Participants were significantly more likely to choose a distractor from the same type as the target for words referring to clothes than for words referring to animals. This could be due to the fact that in this task the memory illusion may be stronger for one category than for the other (e.g., because the animal category may be more salient than the cloth category, making it more subject to illusions).

between a word and possible referents. Here, we propose that some contextual information is memorized and can boost word learning in realistic situations.

### Implications for Learning Words in the Real World

Learners relied on previously experienced information more efficiently when this information was packaged conveniently. That is, cross-situational learning was improved not only by a natural relation between possible referents (Experiment 2) but also by an artificial relation between objects solely induced by their repetitively joint presentation (Experiment 3). Of course, real-life situations are much more complex learning environments than the situations in the word-learning paradigm we used in Experiment 1 (Medina et al., 2011): Here, the level of referential ambiguity is relatively low (four possible referents), only one word is presented at a time, and the true referent is always present in all word occurrences. Further simplification of the task may hence seem inappropriate. However, the specific simplifications we introduced in Experiments 2 and 3 in fact make the task more ecologically valid. In daily life, learners navigate through situations they may be interested in and find coherent. This could help them remember various properties of these situations (a kitchen, a zoo, a pantry, etc.). In Experiments 2 and 3, we introduced such coherence and showed that it has a specific impact on their strategy and performance for learning new words.

Interestingly, a recent computational approach looking at environment regularities showed that coherent activity contexts such as eating, bathing or other regular activities could help simplify the learning problem (Roy, Frank, & Roy, 2012). Our results align with this view, showing that more complex information from the broader context in which a word has been uttered is part of the learning problem faced by the child. The role of the learning environment on word learning requires attention in future research.

### The Representation of Lexical Meaning During Learning

One important issue in the acquisition of word meaning involves the kind of representations children form about words. In other words, what do learners encode about a word when they first hear it? The full understanding of a word requires that learners not only know its word form, its meaning, and its syntactic properties but also information about contexts in which this word may occur. Recent evidence has shown that even infants in the first year of life have already acquired some knowledge for basics words (Bergelson & Swingley, 2012, 2013). However, there is growing evidence that children do not fast-map a dictionary-like definition at the first encounter of the word. Instead, word learning, including verb learning, seems to be a slow process gradually emerging through the accumulation of syntactic, semantic, and pragmatic fragmental evidence (Bion, Borovsky, & Fernald, 2013; Gelman & Brandone, 2010; Yuan & Fisher, 2009). However it is currently unclear what this partial knowledge might be. The present results suggest that, alongside linguistic features (e.g., phonological form, syntactic category), non-linguistic features such as semantic category (Experiment 2) and situations in which the word occurred (Experiments 3 and 4) may be encoded and part of an early word representation. Non-linguistic relations between words are a cru-

cial component of the organization of the lexicon. Work on lexical priming has evidenced that young 21-month-olds already possess a structured knowledge of familiar words based on non-linguistic information such as semantic and associative relations (Arias-Trejo & Plunkett, 2013). As models of lexical development suggest (Steyvers & Tenenbaum, 2005), such a semantic organization of the lexicon is the product of the mechanisms by which word-meaning associations are constructed throughout learning. This suggests that semantic and contextual relations may be encoded from the earliest step of lexical acquisition (see Wojcik & Saffran, 2013, for evidence that toddlers can encode similarities among referents when learning words).

However, to our knowledge, no cross-situational study investigated the role of the learning context in word learning. Such studies could not only shape our understanding of early word representation but also shed light on the content and structure of adults' mature lexical entries.

### Summary

Overall, our findings suggest that learners store in memory the learning situation in which they hear a novel word and use this information to constrain their word-meaning hypotheses. We first proposed a new way to analyze classical word learning experiments through an information retrieval measure. We then modified the classical word learning paradigm to evaluate whether realistic features of the world could inform word learning strategies. Our results show that prior experience is better used when it consists of coherent contexts, and real-world situations may well be coherent contexts in the relevant sense. We conclude that such paradigms, however simple, could and should be used to further study the structure, richness, and poverty of the representations that constitute the early developing lexicon.

### References

- Akhtar, N., & Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. *First Language, 19*, 347–358. doi:10.1177/014272379901905703
- Arias-Trejo, N., & Plunkett, K. (2013). What's in a link: Associative and taxonomic priming effects in the infant lexicon. *Cognition, 128*, 214–227. doi:10.1016/j.cognition.2013.03.008
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. doi:10.1016/j.jml.2007.12.005
- Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using Eigen and S4 classes (R package Version 0.99875-6) [Computer software]. Available at <http://cran.r-project.org/web/packages/lme4/index.html>
- Bergelson, E., & Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, USA, 109*, 3253–3258. doi:10.1073/pnas.1113380109
- Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition, 127*, 391–397. doi:10.1016/j.cognition.2013.02.011
- Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition, 126*, 39–53. doi:10.1016/j.cognition.2012.08.008
- Gelman, S. A., & Brandone, A. C. (2010). Fast-mapping placeholders: Using words to talk about kinds. *Language Learning and Development, 6*, 223–240. doi:10.1080/15475441.2010.484413

- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. doi:10.1016/j.jml.2007.11.007
- Koehne, J., & Crocker, M. W. (2011). The interplay of multiple mechanisms in word learning. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1930–1936). Austin, TX: Cognitive Science Society.
- Koehne, J., Trueswell, J. C., & Gleitman, L. R. (2013). Multiple proposal memory in observational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 805–810). Austin, TX: Cognitive Science Society.
- Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences, USA*, *108*, 9014–9019. doi:10.1073/pnas.1105040108
- Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Vehoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam, the Netherlands: John Benjamins.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Quine, W. V. O. (1964). *Word and object* (Vol. 4). Cambridge, MA: MIT Press.
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. doi:10.1037/0278-7393.21.4.803
- Roy, B. C., Frank, M. C., & Roy, D. (2012). Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91. doi:10.1016/S0010-0277(96)00728-7
- Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*, 480–498. doi:10.1111/j.1551-6709.2010.01158.x
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568. doi:10.1016/j.cognition.2007.06.010
- Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78. doi:10.1207/s15516709cog2901\_3
- Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*, 126–156. doi:10.1016/j.cogpsych.2012.10.001
- Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, *45*, 1611–1617. doi:10.1037/a0016134
- Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.
- Wojcik, E. H., & Saffran, J. R. (2013, August 12). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*. Advance online publication. doi:10.1177/0956797613478198
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420. doi:10.1111/j.1467-9280.2007.01915.x
- Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*, 21–39. doi:10.1037/a0026182
- Yuan, S., & Fisher, C. (2009). “Really? She blicked the baby?” Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*, 619–626. doi:10.1111/j.1467-9280.2009.02341.x
- Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 715–720). Austin, TX: Cognitive Science Society.

Received May 26, 2013

Revision received November 20, 2013

Accepted November 25, 2013 ■