UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE ESTUDOS DA LINGUAGEM

RAFAEL LUIS BERALDO

COMPUTATIONAL MODELS OF LEXICAL ACQUISITION: SURVEYING THE STATE OF THE ART

MODELOS COMPUTACIONAIS DE AQUISIÇÃO LEXICAL: EXPLORANDO O ESTADO DA ARTE

CAMPINAS
2020

RAFAEL LUIS BERALDO

COMPUTATIONAL MODELS OF LEXICAL ACQUISITION: SURVEYING THE STATE OF THE ART

MODELOS COMPUTACIONAIS DE AQUISIÇÃO LEXICAL: EXPLORANDO O ESTADO DA ARTE

Master's Thesis presented to the Institute of Language Studies of the University of Campinas in partial fulfillment of the requirements for the degree of Master, in the area of Linguistics.

Dissertação de mestrado apresentada ao Instituto de Estudos da Linguagem da Universidade Estadual de Campinas para obtenção do título de Mestre em Linguística.

Advisor: Prof. Dr. Pablo Picasso Feliciano de Faria

Este arquivo digital corresponde à versão final da dissertação defendida pelo aluno Rafael Luis Beraldo e orientada pelo Prof. Dr. Pablo Picasso Feliciano de Faria.

CAMPINAS
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Estudos da Linguagem
Leandro dos Santos Nascimento - CRB 8/8343

B45c

Beraldo, Rafael Luis, 1990-
Computational models of lexical acquisition : surveying the state of the art / Rafael Luis Beraldo. – Campinas, SP : [s.n.], 2020.

Orientador: Pablo Picasso Feliciano de Faria.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Estudos da Linguagem.

1. Aquisição de linguagem. 2. Linguística computacional. I. Faria, Pablo, 1978-. II. Universidade Estadual de Campinas. Instituto de Estudos da Linguagem. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Modelos computacionais de aquisição lexical : explorando o estado da arte
Palavras-chave em inglês:
Language acquisition
Computational linguistics
Área de concentração: Linguística
Titulação: Mestre em Linguística
Banca examinadora:
Pablo Picasso Feliciano de Faria [Orientador]
Elaine Bicudo Grolla
Daniel Yurovsky
Data de defesa: 09-04-2020
Programa de Pós-Graduação: Linguística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-6751-2420
- Currículo Lattes do autor: http://lattes.cnpq.br/9566706168712642

BANCA EXAMINADORA:

Pablo Picasso Feliciano de Faria

Elaine Bicudo Grolla

Daniel Yurovsky

IEL/UNICAMP
2020

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação do IEL.

Agradecimentos/Acknowledgements

Aos meus pais, Márcia e Oswaldo, sem os quais nada teria sido possível. Agradeço sempre o apoio de vocês, que se realiza de tantas maneiras.

Ao Pablo Faria, meu orientador, mas também companheiro de conversa nas salas do IEL, nos almoços e cafés da tarde. Agradeço principalmente, é claro, a tua orientação, fundamental para que este trabalho pudesse continuar em tantos momentos.

Aos membros da banca de qualificação, Ruth Lopes e Marcelo Barra Ferreira, e defesa, Elaine Grolla e Daniel Yurovsky, pelas importantes contribuições e discussões que permeiam a estrutura e conteúdo deste trabalho.

Aos meus amigos de hoje e outrora. Marcos e Marina, algumas das primeiras pessoas que conheci por aqui e com quem firmei amizade tão profunda. Yuri, pelas garrafas de vinho, as dicas de cozinha, as conversas urgentes. Igor, pelos desafios fotográficos, pelas muitas trilhas por vir. Ao Diego, com quem há muito infelizmente já não convivo, mas que na época de república em Araraquara dividiu momentos e conversas formadores. Ao Joka, um dos meus amigos mais antigos e fã, como eu, das mensagens de áudio com mais de cinco minutos e que explorem as questões mais esotéricas o possível. Àqueles que compartilharam refeições comigo no bandejão.

A todos vocês, anônimos ou não, que passaram pela minha vida e me impeliram a esse caminho, agradeço pelas influências, mais ou menos evidentes ou conscientes.

Sinto que devo algo a um livro que tenho desde pelo menos cinco ou seis anos de idade. *O pálido ponto azul*, de Carl Sagan, cujas magníficas ilustrações me assombravam quando ainda não sabia ler e, mais tarde, cujas ideias fizeram nascer em mim a vontade de fazer ciência.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa concedida para a realização desta pesquisa (processo 130415/2018-9).

First and foremost, I thank my parents, Márcia and Oswaldo, and their incredible support. Without you none of this would have been possible.

I owe a lot to Pablo Faria, my supervisor and beyond that, a friend. Your company and conversations were much appreciated. Above all, I thank you for all the help, essential to get this research going.

I also thank the qualifying committee, Ruth Lopes and Marcelo Barra Ferreira, as well as the defense committee, Elaine Grolla and Daniel Yurovsky, whose important contributions are imbued in this work's structure and content.

To new as well as old friends. Marcos and Marina, you were among the first people I met here and your friendship has become very important to me. Yuri, I will always remember each and every bottle of wine, kitchen tip and urgent conversation. Igor, our one-shot photography contests were a lot of fun. May there be many hikes ahead of us. Diego, I may not see you as much as when we shared a house in Araraquara, but know that our conversations and the moments we shared were fundamental. Joka, you are one of my oldest friends and I have always enjoyed sending and receiving 5-minute-long audio messages about the craziest of subjects. And thanks to all who shared a meal with me at the university restaurant.

To all of you, anonymous or not, who were part of my life and influenced me to follow the paths I ended up taking.

For some reason, I feel I should thank a book I've had since I was little. *Pale Blue Dot*, by Carl Sagan, had such beautiful illustrations which inspired awe in me before I could even read. And after I had learned to read, the ideas in this book got me interested in doing science.

And a final thanks to CNPq (the National Council for Scientific and Technological Development) for the financial support, grant 130415/2018-9.

Abstract

Children learn words despite great difficulties, one being referential uncertainty. Before having acquired her mother tongue's vocabulary, she has to somehow be able to come up with the correct hypotheses for the meanings involved in linguistic interactions and then map words to meanings. Cross-situational learning has been proposed as a viable mechanism to solve this mapping problem, one that has led to a number of computational studies. Almost three decades of modeling work has yielded different interpretations of this cognitive mechanism, each touted as a plausible implementation given some evaluative methodology. We present an in-depth and extensive exploration of the state of the art of these computational models of lexical acquisition, raising the fundamental question of how much these models have explained so far. In order to answer this, we ask whether input data is actually representative of real-world learning; whether the evaluation methods are sane; and whether their assumptions and simplifications do more harm than good. Each model is reviewed and then compared, in aggregate, to the current theoretical and empirical knowledge of word learning. Further, we carry out novel computational experiments that seem to hint at the fact that some results found by newer models could be replicated by previous computational approaches. All in all, we argue that the current state of the art is primitive, and point to some directions for improvement.

Keywords: lexical acquisition; computational modeling; cross-situational learning

Resumo

A criança aprende palavras apesar de grandes desafios, dentre eles a incerteza referencial. Antes mesmo de ter adquirido o vocabulário de sua língua materna, ela deverá, de alguma maneira, ser capaz de criar hipóteses corretas sobre os significados envolvidos nas interações linguísticas das quais participa para então mapear as palavras a esses significados. A aprendizagem transituacional tem sido proposta como um mecanismo viável que dê conta desse problema de mapeamento, e como tal é objeto de muitos estudos computacionais. Quase três décadas de modelagem computacional nos trouxeram diferentes interpretações desse mecanismo cognitivo, com cada modelo se apresentando como uma implementação plausível de acordo com alguma metodologia avaliativa. Neste trabalho, apresentamos uma exploração extensa e aprofundada do estado da arte desses modelos computacionais da aquisição lexical, levantando a questão fundamental sobre quanto esses modelos já foram capazes de explicar. Para responder a essa questão, nos perguntamos se os dados de entrada são de fato representativos do problema como ele se dá no mundo real; se os métodos avaliativos são razoáveis; e se as assunções e simplificações embutidas nesses modelos mais atrapalham que ajudam. Os modelos são revistos individualmente e então comparados, como um todo, ao conhecimento teórico e empírico atual sobre aprendizagem de palavras. Realizamos ainda experimentos computacionais originais que parecem indicar que alguns dos resultados encontrados por modelos mais novos podem ser replicados por abordagens computacionais anteriores. De um modo geral, defendemos que o atual estado da arte da modelagem da aquisição lexical é primitivo, além de apontar alguns caminhos para o avanço desse tipo de estudo.

Palavras-chave: aquisição lexical; modelagem computacional; aprendizagem transituacional

Contents

Overview

Scientists in many different areas of knowledge have appreciated, for a while now, the usefulness of writing computer programs to better understand the processes under scrutiny. These computational models try to capture enough aspects of the phenomenon so that it is simulated appropriately. From these simulations, different questions can then be asked: what predictions can we make about Earth's future climate? What variables best explain the progressive pollution of our rivers? What were the conditions that led our solar system to its current arrangement? The map, of course, is not the territory; however, if well employed, models can help us put theory to the test and hopefully tell the future as well as unveil the past.

Cognition is one such area being explored computationally. Under the assumption that our mind is the product of the operation of several complex systems, computational models can help us understand what mechanisms underlie human thought and mental abilities. Language – this evergreen source of amazement – is a major cognitive ability that can be and has been modeled. It lends itself to the problem quite well. First, several levels of analysis can be distinguished, which means language can be seen as a complex system itself. Both observable aspects, such as articulation, and unobservable ones, such as concept formation, are necessary for language to exist. Second, it is a task (nearly) every human is capable of accomplishing seemingly without effort, thus being a universal phenomenon. Finally, the input and output is, to some extent at least, well understood, such that expectations for our models can be clearly set.

Whenever input and output can be identified and represented computationally, problems of language acquisition can benefit from being modeled (PEARL, 2010). In general, the starting point is some theory of language learning that is formally specified. The goal is testing whether that theoretical mechanism can learn from the data (and sometimes through biases) thought to be available to children. Since theories of acquisition are written from the high level of natural language, many gaps and assumptions become clear in the course of modeling work. This is one of the major difficulties but also contributions of computational models: writing algorithms demands clear instructions, leading to explicit assumptions (BROEDER; MURRE, 2002; PEARL, 2010). Likewise, modeling also allows us to manipulate variables, compare the performance of different learning mechanisms on the same input, and make well-defined, reproducible predictions (POIBEAU et al., 2013).

Among the aspects of language acquisition modeled computationally, Pearl (2010) lists studies concerning the sound system, words and syntax. For example, she mentions a statistical model of word segmentation by Gambell and Yang (2005) that tests whether transitional probability (the probability of y occurring after x, composing the syllable xy) is a good predictor of word boundaries. They find that this strategy does not fit children's

word segmentation abilities. However, if the same mechanism is given the explicit bias that each word has only one primary stress, then its performance becomes much better. This illustrates how a variable can be shown to be important (at least in the confines of the problem as it is computationally represented), which might lead to further empirical studies in children, seeking to confirm if this hypothesis does pan out.

For the past three decades, models have been proposed to explain how children learn words from their mother tongues. A survey of these models will be the main focus of this work. As we shall see, most set off from a fairly straightforward account of the problem: children are born into a world of symbolic communication in which adults and young users of language employ words to talk about objects, states, events and relations between objects. Two cognitive domains, that of words and that of concepts, are assumed to play a part in this learning. What children seek to do is tease out words from the linguistic input, hone in on which concepts are relevant to a given situation and establish a mapping between these words and concepts. For nearly thirty years, since Siskind's (1996) pioneering work, this has been the problem being tackled.

One would expect, perhaps be sure, that the mystery would by now be a closed case. Nothing would be farther from the truth. Whereas progress has been made, it is often in testing new approaches to solving this mapping problem instead of integrating these models to other levels of linguistic analysis. For example, most work has been limited to studying how nouns – which indeed hold an early position in the course of lexical acquisition – are learned. This may be well and good, yet it is only part of the puzzle. Since the 1990s, research on other classes of argument-taking words, such as adjectives and verbs, has shown that they are acquired differently than nouns (WAXMAN; LIDZ, 2006; TOMASELLO; MERRIMAN, 2014). Consequently, while there have been advances, mostly in the form of new learning mechanisms, one could hardly propose an account for the emergence of a child's lexicon based solely on the models available. They cannot be patchworked together to achieve a broader explanation. As we stand, there is a lack of a robust model (or body of communicating models) that allows for making predictions and testing ideas.

This work is an extensive survey of the state of the art of models of lexical acquisition. I will try to answer the following question: have we left the rudimentary, prehistoric[1] account proposed by early research on the matter (SISKIND, 1990, 1996), or are we still bound to what could be called "cave art?" And if so, what could be holding these models back?

In order to answer these questions, this text is organized in four chapters. Chapter 1 aims to be a summary of the problems of lexical acquisition as put by the different computational models to be reviewed later. It has the dual purpose of being an introduction to how

[1]Inspiration for the title was given by Prof. Marcelo Barra Ferreira in his speech as member of the qualifying committee and evaluator of an earlier version of this work. I do not mean to say Siskind's model – or any other model for that matter – is fundamentally flawed, but merely that since it pioneered the studies in the area, one would naturally assume it would represent an early stage of understanding.

linguistic phenomena can be represented computationally, and being a first explanation of what is involved in learning a lexicon, to be expanded later.

A series of representative models of lexical acquisition are then presented and contrasted, somewhat chronologically, in Chapter 2. These have been published over almost three decades and as a result the expectation is that an evolution in explanatory power can be clearly outlined. Instead, I try to hint at an impending conclusion that for the most part, they have been chasing tail by making simplifying, albeit initially justifiable, assumptions. The chapter also features our own computational studies performed to investigate a hypothesis that Siskind's (1996) early contributions are still relevant, which helps to at least put in check the notion that newer models have departed completely from their origins.

Chapter 3 takes a step back into theoretical and empirical aspects of what is known about lexical acquisition. Here, an attempt is made to explore how far modeling work has veered off from the actual phenomenon. Have they ignored essential aspects of how children learn words and of the difficulties found in natural settings? Or, conversely, have theory and descriptions left so many gaps that the researcher modeling lexical acquisition is forced to simplify the matter right and left? I try to show that instances of both faults can be found.

Chapter 4 wraps up with a summary and discussion of this work's main findings, comparing the current empirical and theoretical understanding of lexical acquisition with a general overview of the models presented here. Limitations of this work are then discussed. To conclude, I ponder on future possibilities for modeling word learning.

Before moving on, a word of warning to researchers of language acquisition: in some ways, the modeling analyzed ahead can be frustratingly simple-minded. Simulating cognition or even input is not a trivial task, which means none of these models come even close to capturing the full theoretical breadth of the field of word learning. For instance, categorization is surely a fundamental aspect of learning words and how to relate them syntactically and semantically. However, since categorization falls outside the problem as it is described computationally, this matter (and many others) are not even discussed, although there are models dedicated to it, such as Redington, Chater, and Finch (1998). This is the double-edgedness of modeling: it requires very explicit instructions and data, meaning simplifications are inevitable, but on the other hand the aspects actually being captured can be specified in detail.

1 Lexical acquisition as construed computationally

A computational model of some phenomenon that takes place in the world is, to some extent, a simulation of that actual event. In order to model a mechanism that might explain children's word learning abilities, the input children receive as well as their cognitive processes have to be simulated. These two aspects of word learning – the cognitive mechanism and the input – have of course been object of extensive theoretical debate. However, since building a theory of lexical (and language at large) acquisition is still an ongoing effort, computational approaches have had to resort to simplifications and assumptions. In this chapter, I try to recount the story of how children go about learning words, and what that means, as if such simplifications and assumptions were the whole truth. This is done by piecing together from the views implied by the models studied ahead. Once the computational account is summarized here and further explored in Chapter 2, we will be better positioned to understand where it is in touch with empirical observations and theoretical explanations of lexical acquisition, and where it departs from these.

Turning back to simplifications, these seem to come in two flavors. A bitter but necessary pill every modeler has to swallow is that which abstracts away from complicating factors of word learning. For example, as I alluded to in the previous chapter, most models actually only try to explain early concrete noun learning. This is due to the difficulties of representing more complex meaning. Verbs might, for example, express a relation between two or more nouns (GLEITMAN; GLEITMAN, 1992), and closed-class words such as prepositions might only be understood in the context of syntactic relations, leaving the purely conceptual domain. There is no clear way to satisfactorily represent these semantic properties compatible with computational modeling. However, there is a second, sweeter side to simplifying the problem: generalizations might be found that cut to its core. Even if different word classes are associated with wildly different meanings, the same mechanism could in principle explain all learning. Modeling work tries to strike a balance between abstracting while also keeping what is essential about the phenomenon.

For the moment though, we will take the view of word learning purported by our models of lexical acquisition as the gospel. This view is broken down by following Bertolo's (2001) five questions on characterizing learning problems (p. 2), reproduced below and developed in the five following sections. I outline what it would take and what it would mean to learn a vocabulary if the input data, mechanisms and end goals encoded in these models were the actual word learning experience children go through. The resulting description will later be brought into question. I also discuss what aspects of word learning can be adequately reproduced computationally in section 1.6.

1. What is being learned, exactly?
2. What kind of hypotheses is the learner capable of entertaining?
3. How are the data from the target language presented to the learner?
4. What are the restrictions that govern how the learner updates her conjectures in response to the data?
5. Under what conditions, exactly, do we say that a leaner has been successful in the language learning task?

1.1 The object of lexical acquisition

It is commonly stated that children start acquiring words very slowly by the end of their first year, but ultimately amass an extensive vocabulary of around 60,000 words (BLOOM, 2000). However, not much is said about the internal structure that the word *vocabulary* implies. Likewise, lexical acquisition is generally modeled as the building of a list of word-to-meaning mappings. More will be said about the nature of "word" and "meaning" in due time. For now, we accept the simplification that the learner has to infer that a sequence of lowercase characters, say *kettle*, is associated (or *maps*) to the sequence of uppercase characters KETTLE. Eventually, a list of such mappings will build a lexicon which hopefully mirrors the one employed in generating the utterances in the input corpus.

Due to the different learning mechanisms employed by each different model (discussed ahead), the shape of these mappings may vary. To some (YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010), words are mapped to meanings via a probability distribution. Since a word is normally seen in different extra-linguistic contexts, the learner assigns some strength to each cooccurring meaning. Given enough data, higher probability will eventually be assigned to the correct meaning, following some learning mechanism. Under this approach, a word is considered acquired when the strength of one of these alignments exceeds a given threshold.

Another way mappings might be represented is by a strict 1:1 alignment, that is, each word maps to only one meaning. Consequently, of course, homonyms cannot be distinguished. This can be solved by admitting another level to the lexicon (SISKIND, 1996), such that each word maps into one or more *senses* which individually map to a single meaning. This is the only model in which the resulting vocabulary is structured.

To conclude, under the current computational view, the goal of lexical acquisition is building a vocabulary. Normally, a vocabulary (or lexicon) is a list of sequences of lowercase characters called *words*, each mapped to a single sequence of uppercase characters, a *meaning*, or to a *distribution of probabilities* for different meanings. In some cases, where a 1:1 mapping is assumed, an intermediate level may be added. This introduces structure into the lexicon and allows the learner to distinguish between homonyms – words which look the same, however should be associated to different meanings.

1.2 Hypotheses the learner makes

Quine's (1960) gavagai problem is a staple question of lexical acquisition. Simply put, it asks the question: how can one correctly infer what speakers of an unknown language are talking about, given extra-linguistic context can be conceptualized in virtually infinite ways through language? Children are born just in that circumstance: they have no prior knowledge of the words used by their interlocutors and yet have to hypothesize what the speakers might be conveying through their utterances. Linguistic and cognitive theory say children naturally restrict their hypothesis space in a number of ways: by having lexical biases (MARKMAN, 1989, 1990) that restrict possible word meanings, by paying attention to social clues such as eye gaze and pointing (BALDWIN, 1993, 1995), by developing a theory of mind or "mind reading" capabilities (BLOOM, 1997, 2000), or by considering cues from syntax (LANDAU; GLEITMAN, 1985; GLEITMAN; GLEITMAN, 1992). There are boundless sources of information for the child to tap into, thus considerably reducing the problem of referential uncertainty. Although some models do encode aspects of these sources of information, as we shall see in the next chapter, most embed lexical biases in the input data.

Virtually all models considered here only study the learning of concrete nouns. In this scenario, learners hear utterances and see the world around them. The objects they can see and represent in their mind constitute all hypotheses available. Therefore, if the learner hears the utterance "The coffee doesn't have sugar" at the breakfast table, the word *coffee* might refer to anything from the TABLE to the KETTLE on the table, including even COFFEE itself. However, since on this particular table there is no sugar to be seen, in this situation the learner could not postulate the correct hypothesis and learn the association *sugar* → SUGAR. Notice that verbs, adjectives, adverbs, closed-class words etc. are competing for association with these referents, making it harder to learn the noun-meaning mappings.

A notable exception is the view (SISKIND, 1996) that for each utterance, a number of complex meaning representations will be generated by the learner's mind. These representations are supposed to capture the *full* meaning of that utterance. Some utterances are paired with a correct as well as some incorrect meanings, thus the learner has a chance of learning (at least partially) the individual meanings of each word. Other utterances, however, are not paired with correct meanings at all. This is meant to capture the fact that sometimes learners will only make incorrect conjectures. Further, using more complex meanings allows the model to represent verb-like as well as noun-like meanings.

In summary, learners are capable of generating hypotheses for the meanings of words or utterances. These meanings are taken from context and thus hypothesizing is prone to some error: it might be that the utterance does not mention anything that can be seen around. Also, because there are usually many objects in sight, there is (normally) always more than one possible hypothesis, which constitutes referential uncertainty.

1.3 The input data

Let us assume that our learner has two sensory channels: auditory and visual perception. The data from these channels are further refined by cognitive abilities that extract words and concepts. In a number of computational models, as has been said, these words are sequences of lowercase characters (*omelet*), while concepts are sequences of uppercase characters (OMELET). These are, of course, meant to allow a representation of the mapping problem and subscribe to a symbolic view of cognition.

What is the nature of these symbolic representations? The input data these models receive generally come from corpora of child-directed spoken utterances[1]. These corpora have been transcribed and each word is defined as being the sequence of characters delimited by spaces and punctuation. However, there are no corpora annotated with some semantic representation of the intended meaning, let alone all viable interpretations of the extra-linguistic context. Some works on computational modeling employ human annotators that observe the scene when an utterance is spoken and manually list all available referents following some methodology. Others use a programmatic approach where each word is pre-mapped to some concept. Then, when an utterance is presented, the corresponding concepts come together with some extra concepts, modeling referential uncertainty.

This means, under the computational view, a learner receives as input a list of utterances paired with meanings. Utterances are made of words as defined above and meanings, which are an atomic symbol standing for an object in the world or a mental concept.

There is an alternative view (SISKIND, 1996) where semantic representations are more complex. Each utterance is paired with a (number of) decomposable tree of meaning symbols. This representation is meant to capture verb-like meaning (argument-taking semantic "functions") and noun-like meaning (symbols or structures of symbols found in argument positions). Together, these two types of meaning contribute to expressing the meaning of the whole utterance. Each utterance is paired with one or more of these representations, each an interpretation of what could possibly have been said. This representation is explored in detail in section 2.1.1.

1.4 Learning mechanisms

The two-channel input described above coordinates synchronized pieces of information: spoken words from utterances and meanings available in the extra-linguistic context. Even though there is temporal coincidence, the problem of word learning as construed computationally is such that mere cooccurrence does not entail association. One problem is that more than one word can be spoken in the presence of more than one perceived object in the real world. Thus, the learner has to find a way of aligning words to things.

[1] There is no reason why spoken utterances could not be replaced by signed utterances, thus modeling (an aspect of) sign language vocabulary acquisition.

Another problem is that words might refer to different things than that conceptualized by the learner's mind at any particular time. Consider the utterance "This is my favorite breakfast dish" while the mind is entertaining a rather delicious looking OMELET. This situation would allow the inference of the mapping *breakfast dish* → OMELET, which is not completely accurate.

A unifying idea of all models discussed here is that a learner could solve such problems by employing a *cross-situational learning strategy* (PINKER, 1989; GLEITMAN, 1990; FISHER et al., 1994), discussed in detail in the next chapter. The rationale is that children are somehow capable of observing several uses of a word, store meaning hypotheses licensed by each extra-linguistic context, and then select a winner by intersecting which meaning is consistent *across* these contexts. Since specifying how this mechanism could be cognitively instantiated is the point of each computational study, this is also where they diverge.

There are two conflicting dimensions in the computational view of cross-situational learning. Some (YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010) assume that the learner is capable of tracking the probability a word will cooccur with a given meaning by means of an associative mechanism. The more frequently a pairing is seen, the higher that pairing's association. Another view (SISKIND, 1996) is that the learner could store and then eliminate or promote meanings compatible with context by following some heuristics. These two views constitute the first dimension, which I call *probabilistic vs. deterministic*. The second dimension, *global vs. local* (nomenclature introduced by Stevens et al., 2017) further divides probabilistic approaches into two modes of operation. Global learners are those which spread the association strength of a given word to all meanings having cooccurred with it. Local models, on the other hand, select a single meaning candidate for association and verify for consistency with context whenever possible, strengthening or weakening this association accordingly. The proponents of this second approach (TRUESWELL et al., 2013; STEVENS et al., 2017) argue that it yields results comparable to its global counterparts, while being a more parsimonious explanation.

A number of these proposals view the learner as having serial, *online* access and processing of these utterances. After each utterance, its knowledge of the words seen so far will have been updated. Others (YU, 2008; FRANK; GOODMAN; TENENBAUM, 2009; YUROVSKY; FRANK, 2015) take a different approach: the learner considers all utterances in a chunk of the corpus and try to infer the best lexicon given all that data. These are called *batch* learners.

In very broad strokes, learners are thus cross-situational and update their hypotheses by either following rules in order to infer word-to-meaning mappings, or by increasing or decreasing association strengths. The specific mechanisms for this link building will be different depending on the particular implementation of the learner. These mechanisms generally keep either global or local hypotheses and process the data either in an online or batch fashion.

1.5 Successful word learning

The final goal of word-learning is building a vocabulary or lexicon, as was stated above. In general, this is a list of *word* → MEANING associations. Once the learner has processed all of its input data, the resulting lexicon can be compared to the gold-standard, the computational equivalent of the shared vocabulary of the linguistic community a child is born into.

Two measures are sometimes used to assess a learner's performance. These are precision and recall. Precision is the proportion of pairings learned which are correct. Thus, if precision is fifty percent, that means the learner was capable of correctly inferring half of all word meanings, although the other half will be incorrect mappings i.e. false positives. The second measure, recall, is taken looking only at the correct mappings learned and comparing these to all the gold-standard. In other words, it measures how much was learned from the expected grand total.

More frequently though, success is measured by looking at the learning path. It is thought that children exhibit a distinct curve marked by early slowness and followed by an explosion in the rate of word learning. Further, children are also known to fast map (CAREY, 1978; CAREY; BARTLETT, 1978), that is, quickly learn words with a single exposure when the context is telling enough, a task dependent on prior knowledge of words. Another common way of measuring success is by comparing the learner's behavior to adults (or sometimes children) solving an experimental word-learning task. Such tasks are simplified experiments meant to capture an aspect of lexical acquisition. Thus, if a computational learner shows a curve fitting children's reported learning path or if it behaves similarly to humans in simulations of experimental tasks, it is considered a candidate explanation of lexical acquisition.

1.6 The computational mapping problem

So far, we have answered Bertolo's questions to characterize the problem of lexical acquisition as viewed through the computational lens. However, we still have to understand what aspects of word learning are captured by the problem of associating two streams of data. As foreshadowed elsewhere, these associations are normally called *mappings* (FISHER et al., 1994) and the overall task of word learning is equated to the so-called *mapping problem*. An interesting property of this characterization of the problem is that by assuming these two streams of input, a number of difficulties of actual word learning can be represented.

The first computational instantiation of the mapping problem, as far as could be found in the cognitive modeling literature, comes from Siskind (1996). This depiction (p. 54) lays out the common features assumed by subsequent models. Roughly, Siskind describes it as follows: a word learner is presented with a corpus of utterances paired with semantic

representations of these utterances. Utterances may be lists of words, whereas semantics may be represented by a list of referents, a structured description of a visual scene, or some kind of mental conceptualization. The learner further assumes that each utterance in the corpus is generated from a lexicon shared by the surrounding linguistic community. The task is to infer the correct mappings from word to meaning.

A number of properties are shared between this problem and what seems to be required of the child learning words. First and foremost, *referential uncertainty* is achieved by pairing multiple semantic representations with each utterance – just as it may be reasonable for the child to postulate a number of possible interpretations, sanctioned by the surrounding environment, to any utterance. A first approximation at representing the child's experience may look like this[2]:

Utterance	"Daddy has bought a new toy!"
Meaning	{DAD, BUY, A, NEW, TOY, SHOP, DOLL, GIVE}

Here, the atomic meaning representations DAD, BUY, A, NEW and TOY are mapped, as expected, to the words "daddy," "has bought," "toy" and so forth. However, there are so to speak "extra" meanings – SHOP, DOLL, GIVE – which model other possibly salient referents available in the extra-linguistic context. This ambiguity can also be taken to an extreme by having no correct meaning representations paired with an utterance: this constitutes *noise*. Noisy pairs are an obstacle for cross-situational approaches because they provide evidence towards wrong associations between words and meanings. It should also be said that this atomic semantics is by no means the only option for representing meaning, as we shall see.

Apart from referential uncertainty, which is the major difficulty cross-situational strategies try to overcome, *alignment ambiguity* (FAZLY; ALISHAHI; STEVENSON, 2010) is another aspect of the mapping problem. Siskind (1996) shows that only about 9% of child-directed speech is actually composed of single-word utterances (p. 48). That means a majority of the input a child receives has some degree of alignment ambiguity, that is, *a priori* any word in the utterance can be reasonably mapped to any concept. How each model deals with this or ignores it is going to be discussed in detail in the next chapter.

Synonymy and *homonymy* are two other phenomena easily captured by this depiction of word learning. $1 \rightarrow 1$ mappings happen when a word has only one sense. $1 \rightarrow n$ mappings represent homonymous words, that is, a situation when a single word has n senses. Finally, the last possibility are $n \rightarrow 1$ mappings, when n words map to the same single meaning, such as in the case of synonyms. By virtue of these representations, many of the issues in lexical acquisition can then be explored in a computational setting. However, there is one issue which cannot be captured by simple mappings like the ones above. Polysemous words are a challenge since their different senses are closely related rather than being completely

[2]This scheme closely resembles that of Fazly, Alishahi, and Stevenson (2010), to be discussed later.

arbitrary. Even then, some of their aspects can be represented by more nuanced semantics, as we will see for one of the models.

Homonymous words, if not assumed by the learner, will be the same as noise. Consider a naive learner who has acquired the mapping "tail" → TAIL. She might later on hear the homophone "tale," which roughly maps to STORY, and think her previous hypothesis is wrong. She would then revise her hypothesis by either ruling out the previous mapping, or weakening her confidence in it when in fact the correct conclusion would be that /teɪl/ has two senses. Therefore, the learner has to somehow be able to identify homonyms in her experience if lexical acquisition is to be successful while accounting for homonymy. Of course, the researcher might decide to build into the model an assumption that the data contains homonymous words. Another approach is building a model capable of inferring this property from the data. Either way, an explanation for the ability children exhibit of postulating a new word sense is needed. If the researcher has reason to believe that a priory biases cue the child in on the fact that lexicons have homonymy, then the first choice would naturally follow. If however awareness of homonymy is a corollary effect of how the model operates, then it can be underdetermined in this way. This is a desirable feature when trying to explain word learning by appealing only to more general cognitive processes. Nevertheless, in both cases the researcher's assumptions will have been made explicit. Similar modeling decisions will indeed be a source of disagreement when we discuss the computational models below.

Synonyms, on the other hand, might not pose the same problems, unless the model assumes some kind of absolute constraint that objects only have one label, as has been proposed (MARKMAN, 1990; CLARK, 1987; GOLINKOFF; HIRSH-PASEK, et al., 1992). This could be a way of achieving fast mapping (CAREY; BARTLETT, 1978), the ability children have of learning a word within a single trial. If that assumption is made, this would be equivalent to a naive learner hearing "I got you a puppy" and, having already acquired "dog" → DOG, barring the mapping "puppy" → DOG. On the other hand, if the model does not have a strong theory of no-synonymy, these words should not be a problem unless they cooccur in the same sentence. For example, the learner hears the utterance "this puppy is such a cute dog" paired with only one occurrence of DOG in their interpretation of the world. If the model expects that every word contributes to the meanings available, then it could be derailed by this piece of data. Again, different solutions have been proposed and are discussed in the upcoming chapter.

1.7 Word learning as a computational problem

We are now in a position to understand the problem of lexical acquisition computationally. Lexical learners are cross-situational mechanisms which employ deterministic or probabilistic rules in order to build a lexicon, that is, a list of associations of words and

meanings. Words are defined as being sequences of lowercase characters found in corpora of child-directed speech. Meanings, generally objects available in the extra-linguistic context, are synchronized to utterances from these corpora by manual or automatic annotation.

In fact, these meanings represent hypotheses the learner has to entertain, since it is a property of the data that mere coincidence does not necessarily lead to a word-to-meaning mapping. Thus, learners have to exploit other regularities to solve this otherwise misleading mapping problem. The mapping problem features difficulties such as referential uncertainty, noise, and alignment ambiguity. The underlying vocabulary generating the input utterances may include synonyms and homonyms.

Turning back to the cross-situational strategies, besides being deterministic or probabilistic, these can also assume larger (global) or more restrict (local) memory of past meaning hypotheses. A global learner will remember and use as much information as it can from observed cooccurrences, while local approaches use narrower hypotheses and discard or confirm them as they go. This greatly impacts the behavior of the learner. Similarly, some views give the learner serial access to the data, while others give it time to process data in a batch, allowing for more ways of building associations.

Successful acquisition is defined as yielding a lexicon that is as close as possible to the gold-standard. However, other ways of assessing a learner are comparing its behavior to reports of empirical studies of child acquisition (e.g. learning rate, fast mapping) or to experimental results found in experimental tasks meant to simulate word learning.

Note that such generalizations are only meant as a thirty-thousand-foot view of lexical acquisition as it is construed computationally. As we discuss each model, we will also consider its peculiarities. In general, though, this is a good approximation which will guide our evaluation of the models against the current empirical and theoretical account of lexical acquisition.

2 Computational models of lexical acquisition

In the previous chapters, I tried to establish that computational models are tools that can shed some light on cognitive and linguistic issues. I also discussed which aspects and strategies of lexical acquisition have been relevantly modeled. In this chapter, I will present a detailed survey of a number of cross-situational models of word learning which have attempted to solve the mapping problem. Each implements a different, hopefully plausible cognitive mechanism and then tries to show how it copes with referential uncertainty when learning a lexicon. Thus, there are a number of perspectives from which to analyze the models: their specific depiction of the mapping problem, what input data and representation was chosen, as well as their performance.

Although most models are hard to compare directly, a table with their main characteristics will be provided and updated after each model is presented. This will be a useful tool to give unity to this analysis and help understand not only each model's contributions, but also where they overlap or diverge. Table 12 summarizes the contributions brought by each model, allowing for easy comparison of their properties.

At the end of the main discussion (section 2.1), some issues regarding memory and word learning will hopefully become self-evident. These issues spurred brand new computational investigations which I present in section 2.2. Since I am trying to get at the state of the art of the models of lexical acquisition, these investigations are presented as a way to compare the very first model proposed with the latest results in the literature. This way, the question of how much has advanced may be answered in a more direct way.

The main discussion is then extended in the final section (2.3), devoted to a few other studies in lexical acquisition which, for whatever reason, do not try to solve referential uncertainty, but are instead concerned with related issues. They include, for example, how semantic categories are formed or the interaction between words and categorization.

2.1 Main trends in a sample of models of lexical acquisition

Even in the recent lexical acquisition literature (YANG, 2019), the case continues to be made that the associations between language and world are not trivially available to the word learner. This has been called "Gleitman's Problem" and harks back to Quine's (1960) musings over linguists trying to figure out the meanings of words from unknown languages. If it is a property of language that it does not exclusively refer to the present or the immediate context, then it follows that inferring the meanings of words might not be an

easy task. Referential uncertainty, then, emerges as the learner's biggest obstacle – one that the computational models reviewed in this section all try to overcome.

Cross-situational strategies, as discussed in Chapter 1, have been proposed (GLEITMAN, 1990; YU; SMITH, 2007; PINKER, 1989) to explain how children solve referential uncertainty and related issues. However, this is not enough to provide a full account of word learning under such a cognitive mechanism. What does it mean, exactly, to say that children are able to observe words being used across situations and then reliably find what is common between them? Take nouns for example: children could store in memory all word-object cooccurrences (SISKIND, 1996; YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010) or else make a guess and store only one hypothesis at a time while keeping an eye out for counterexamples (TRUESWELL et al., 2013; STEVENS et al., 2017). Both have been shown to yield similar end results while making different predictions about lexical development. Modeling allows for studying which implementations of cross-situational learning solve referential uncertainty, as well as which display behavior compatible with actual lexical development.

It will soon become evident, however, that it is not obvious in any way how to directly compare these models. Not only do they implement different strategies of word learning, but also differ in what their input data is like. Some assume a richer semantic representation and a wider array of syntactic categories, while most barely have any semantics at all and only model concrete noun learning. Some use corpora of child-directed speech, whereas others generate their own synthetic corpora. Similarly, even their results and metrics of success are not the same. While some researchers have opted for measuring how good a model is at learning a lexicon (usually compared to other models), others argue that their proposals mirror child behavior. This could mean showing a learning curve which resembles what children seem to go through: vocabulary bursts, difficulty learning synonyms, initial confusion of words etc. Yet another approach is giving models a simulation of a psycholinguistic experiment and comparing their behavior to that of human subjects. In the absence of a common testbed, a more qualitative evaluation is needed.

This section is organized in an almost chronological order. Instead of attempting to crown one of the following models as king, I will sing their praises and point out their shortcomings. In the end, however, one question remains: altogether, what have these models achieved and what are they lacking? My coming attempt at an answer might point future work into new explorations.

2.1.1 Siskind's early work[1]

The first cognitive models to explore word learning come from Siskind's (1990; 1996) work. Although some other researchers had already computationally explored the acquisition of word meaning (SALVETER, 1979; PUSTEJOVSKY, 1988), Siskind's (1996) model was the first to implement a cross-situational strategy designed specifically to solve referential uncertainty in a cognitively plausible way. Many of the models which followed echo in some way the decisions made by this early work, thus it is invaluable to spend some time to understand its inner workings, its assumptions about lexical acquisition, what data was used and its results. This will form a basis from which to understand subsequent modeling efforts and developments.

Siskind's (1996) lexical acquisition model learns word meanings under referential uncertainty, noise, multi-word utterances, and homonymy, thus simulating many of the issues associated with the mapping problem. The task is simplified to be the interaction of a lexical learning faculty with two other cognitive faculties (see Figure 1). The speech perception faculty represents children's ability to segment words from the speech stream. It generates word symbols which correspond to some mental representation of the spoken word. Thus, upon hearing the utterance "Mary lifted the chair" this faculty generates the list *Mary, lifted, the, chair*. The second input comes from the conceptual/perceptual faculty, which represents children's ability to infer what might have been said given some extra-linguistic setting. This faculty generates Jackendovian (1983) conceptual expressions that represent meaning. For example, upon seeing Mary lift a chair, it generates CAUSE(**Mary**, GO(**chair**, UP)). Referential uncertainty in this model is the measure of how many conceptual expressions come paired with an utterance. Noise happens when an utterance is paired only with expressions which do not correctly represent its meaning.

As the model observes new data, it will eventually determine that the word symbols *Mary, chair* and *lift* map to the conceptual expressions **Mary, chair** and CAUSE(x, GO(y, UP)). As can be seen, there are two types of expressions: those which have argument positions (x and y) and those which do not. These respectively represent verbs, which in this model always have argument positions to be filled, and nouns, which do not (p. 52). Further, each word symbol may be homonymous, since they can be associated with one or more senses. As for functions words, such as the determiner *the*, they are modeled as elements lacking any semantics. This is a significant limitation and although Siskind argues (p. 46) that "adopting a richer conceptual-symbol inventory would allow the algorithm to represent and learn the meanings of determiners," research suggests there may be unforeseen consequences (FARIA, 2015); also see section 2.3.2 of this work.

The model follows four principles required for successful learning. The first is an assumption that children can gather and use *partial knowledge* of word meanings. This

[1]Much of the text in this section was translated and adapted from a paper I submitted to *Revista do SETA* (BERALDO, 2019).

Figure 1 – Siskind's (1996) model's architecture: the lexical learning faculty receives input from two other faculties: one that segments words from speech and another that generates conceptual expressions from the conceptual/perceptual data available to the learner. No claim is made that such faculties really do exist, this model being a simplification of the actual problem. Figure adapted from p. 44.

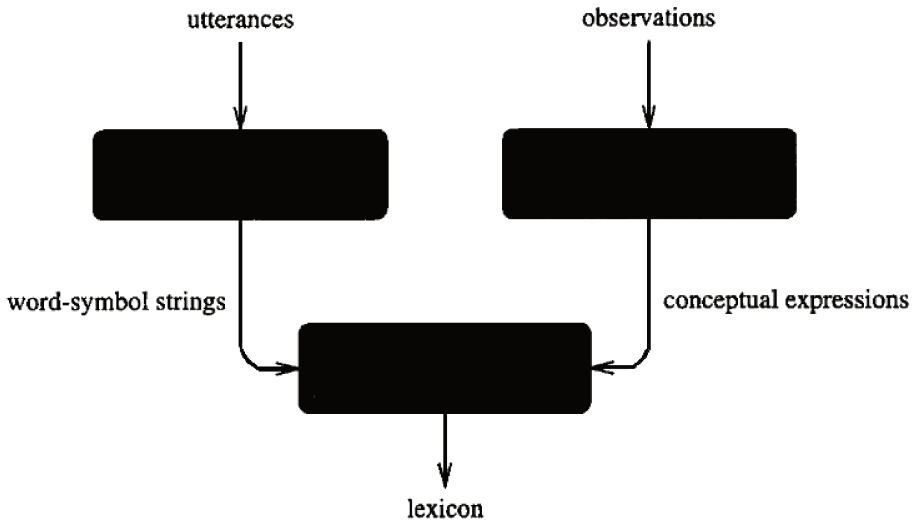


Figure 2 – The use of partial or full knowledge of a word (*John*, in this case) allows the learner to rule out meanings licensed by the context but incompatible with the utterance.

knowledge is then used to check if a given meaning paired with an utterance is possible or not. For example, given the utterance "John fell" and the meaning hypotheses FALL(**John**) and EXIST(**banana-peel**, ON(**floor**)), sufficient partial knowledge of either *John* or *fell* is enough to rule out the second hypothesis. See Figure 2 for an illustrated example.

Next, the second assumption is that children employ *cross-situational inference* to learn words (Figure 3), allowing them to observe what meanings are common across different uses of that word. That can be seen as forming meaning sets for each extra-linguistic situation and then intersecting those sets. For example, upon hearing "John fell" correctly paired with FALL(**John**) and then "Mary fell" correctly paired with FALL(**Mary**), the intersection of the available meaning sets {FALL, **John**} and {FALL, **Mary**} is {FALL}. Applying certain rules of semantic composition allows the learner to eventually acquire the conceptual expression[2] FALL(x) as the full meaning of *fall* under this approach.

[2]Siskind calls meaning fragments *conceptual symbols*, which can be arranged into a *conceptual expression*, that is, a nested structure of conceptual symbols. Verbs are modeled as conceptual expressions

Figure 3 – A cross-situational strategy allows the learner to intersect the meanings in common between situations (1) and (2), thus inferring that *fell* means FALL.

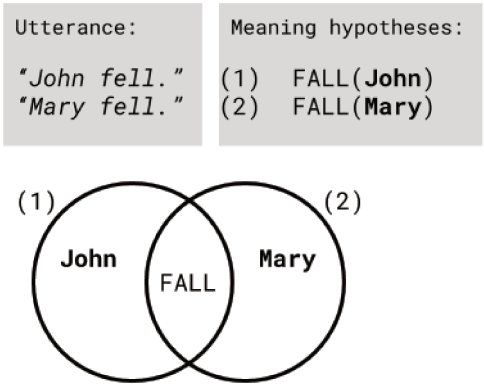
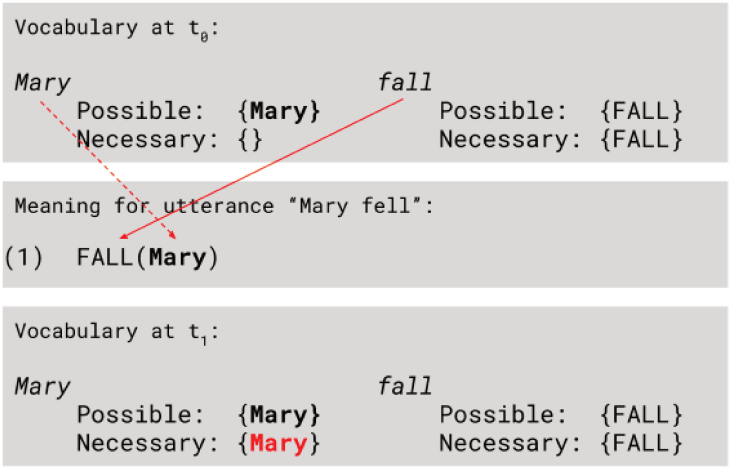


Figure 4 – In order to determine if a concept is *necessary* to the meaning of a given word, the learner assumes that all meanings are contributed by the words in the utterance. Thus, since only *Mary* contributes the meaning **Mary**, the learner determines (dashed arrow) that **Mary** must be a necessary meaning of the word *Mary*.



Although cross-situational inference allows the learner to identify which meaning fragments cannot be part of the meaning of a given word, it is not sufficient to determine those which *must* be. Thus, the third principle assumes the learner operates under some *covering constraints* (Figure 4) which say that the meaning of a given utterance is strictly derived from its words. Then if no other word contributes with FALL in "Mary fell," it follows that *fell* must contain the meaning FALL. Whereas cross-situational inferences progressively rule out possible meanings, the covering constraints allow a learner to determine necessary meanings. When both possible and necessary meaning sets have been established and are identical, the meaning fragments of a word have been found: it is said that the word "converged" on its meaning fragments.

The fourth and final *exclusivity principle* states that words in an utterance do not contribute overlapping meanings. Guided by this principle, a learner can determine whether the meaning of *fall* in "John fell" is FALL(**John**) or FALL(x) by knowing that **John** is

taking arguments or *variables*, such as x, y and z. Nouns are decomposable expressions having no variables.

already contributed by another word. Notice that this poses a problem to learning if agreement is modeled, since some features such as [±PLURAL] are contributed by more than one word in languages like Portuguese (FARIA, 2015). For example, the utterance "os cachorros correram atrás do João" (meaning "the dogs chased after John") has three instances of that feature. This problem is set aside in Siskind's original study, since it does not model functional items; however, I will come back to this point later in the discussion.

Lexical acquisition under this model occurs in two interwoven stages, the first to determine meaning fragments and the second their arrangement. Utterances are processed in an online fashion, that is, they are observed one by one and then discarded. Each utterance is fed into a set of heuristics[3] which encodes the four principles just discussed and was designed to extract as much information from each learning instance as possible. As alluded to above, words may be in one of two learning stages. In the first stage, the possible and necessary meaning fragments are determined; e.g. for *enter*, the intersection of these two sets will eventually contain {GO, TO, IN}. Once these fragments or *conceptual symbols* have been found, the algorithm will then assemble them into all possible *conceptual expressions* to be tested until the winner is found. *Enter* can in principle be represented by GO(x, TO(IN(y)), GO(TO(IN)), IN(GO(x), TO(x, y)) etc.

For a word w, its lexical entry[4] is represented as three tables, $P(w), N(w)$ and $D(w)$. The first two are used during the first stage and contain, respectively, the set of possible conceptual symbols of w and the set of necessary conceptual symbols of w. When entering the first stage, a word w has its $P(w)$ initialized to be the universal set, which represents all conceptual symbols in the target vocabulary. The table $N(w)$, on the other hand, is initialized to be empty. Symbols are gradually removed from the possible set or added from it to the necessary set. Once the contents of $P(w)$ and $N(w)$ are identical, the correct conceptual symbols of w have been found. The second stage then begins. All possible combinations of these symbols are built and added in $D(w)$ and removed one by one whenever possible until a single one is left. When that has happened, the model has "converged on the conceptual expression" (p. 56) which represents w's meaning. Both stages coexist as the algorithm runs, since different words can be at different stages of acquisition.

An exhaustive discussion of Siskind's six heuristics is beyond the scope of this review. However, it is worth to briefly consider the first four rules the model follows when processing

[3] Farrell and Apter (1972), discussing artificial intelligence, define heuristics in the following way: "An algorithm is, as we have seen, a set of rules which, if followed literally, will achieve a desired solution. A heuristic, on the other hand, is a set of rules which, if followed, may achieve a solution but cannot guarantee doing so." (p. 83). Heuristics are often referred to as "rules of thumb," that is, useful ways of solving a problem but which lack a proof or are not completely reliable.

[4] Siskind also introduces a more advanced mechanism to learn homonymous words and resist noise. This is achieved by adding another level of abstraction to the lexical organization of his model. Each word w maps to a sense set. Each sense s in the sense set represents a word's sense; thus *ball* would map into the sense set {**round-toy**, **formal-dance-party**}. The actual details are not important for the purposes of this review. However, I will discuss their implication when comparing Siskind's to other models.

an utterance-meanings pair. These will be fleshed out and more formally stated when I present the computational simulations forming the body of my main experimental contributions in section 2.2, in which a direct adaptation of Siskind's model is presented. These four rules are designed to, whenever possible, 1) reduce referential uncertainty by removing utterance meanings inconsistent with partial or full knowledge of the words in that utterance; 2) remove meaning fragments not appearing in the remaining utterance meanings from $P(w)$; 3) add fragments from $P(w)$ to $N(w)$ if they are uniquely contributed by w; and 4) remove meanings from $P(w)$ if they appear only once in some $N(w')$. When working together processing data with referential uncertainty but lacking noise or homonymy, the four first rules are able to converge on the conceptual symbols representing the meaning of some w. The two last rules are designed to determine the final conceptual expression stored in $D(w)$. For an example of the first four rules in operation, I refer the reader to Beraldo (2019); for a complete explanation and example, see Siskind (1996).

The full version of the model includes a mechanism to acquire a lexicon when input data has homonymous words as well as noisy utterance-meanings pairs. These represent a problem for the heuristics discussed above, since a homonymous word may look like noise, and noise itself may corrupt word meanings. Suppose a word is under acquisition, say *duck* meaning **domestic-bird**, and now an utterance containing the homonymous verb is presented to the model, paired with a correct meaning representation. Since the current meaning hypothesis set of *duck* contains **domestic-bird**, and the current utterance has the verb *duck* and thus lacks that meaning, the heuristics will rule out all meaning representations and discard this learning instance. This is one problem posed by homonyms. Noise, on the other hand, may introduce wrong meanings into a word's meaning hypothesis set. Imagine an utterance is presented to the system, but all paired meaning hypothesis are wrong. If the utterance contains only new words, for example, the heuristics have no way of knowing all meanings are inconsistent with the utterance's real meaning. Then, it will assign wrong meanings to each word. This corrupts the lexicon and can spread to other words (p. 62). For those reasons, Siskind's final model is able to propose new senses for a word when certain conditions are met and has a sort of "garbage collector" that periodically forgets meanings which have not been up to snuff, following metrics discussed later on in section 2.2.

This homonymy-and-noise-ready version of the model is then tested in four simulations. The first three were designed to study 1) how sensitive the model was to different variables, 2) vocabulary growth as a function of the number of utterances processed, 3) and the rate of new word acquisition. Siskind reports (p. 72) that the model was sensitive to variations in the homonymy and noise rate, but not to vocabulary size, conceptual symbol inventory size or rate of referential uncertainty; that it showed behavior akin to what is observed in children, first acquiring words slowly but then exhibiting a spurt (NELSON, 1973; DROMI, 1999) in the rate of new word acquisition; and that, in keeping with fast mapping (CAREY; BARTLETT, 1978; GOLINKOFF; JACQUET, et al., 1996), after having

processed about 4,000 utterances, the model is able to learn new words with only one or two observations. These results are taken to suggest that the model is indeed compatible with the behavior observed in children and that, assuming it is a fair characterization of the lexical learning task and of the strategy children employ, homonymy and noise are predicted to be bigger hurdles to clear than referential uncertainty.

The fourth and last simulation was intended to approximate the lexical acquisition task children face. The parameters were adjusted as follows: target vocabulary of 10,000 words, referential uncertainty of 10 hypotheses per utterance, a conceptual symbol inventory containing 250 symbols, homonymy rate at 1.68 and noisy utterance-meanings pairs at 5%. Of these parameters, Siskind only justifies the choice of the homonymy rate, which was adjusted following the number of homonyms found in WordNet. After having processed 1.5 million pairs, the model had learned 80.7% of words. The remaining 19.2% were false negatives, that is, items missing in the vocabulary. False positives, wrong items added to the vocabulary, were at 12.2%. According to the author, although the previous simulations had been run until converging on 95% of the total vocabulary, this last experiment had to be aborted due to computational limitations.

Since there are no corpora of naturally occurring speech annotated semantically compatible with the needs of this model, input data came from a synthetic corpus generated on the fly. This allowed for varying the parameters described above, as well as mean length of utterance (MLU). Thus, a portion of the algorithm is responsible for generating utterances, which are simply symbols representing words, such as $w_1, w_2 \ldots w_n$, paired with conceptual expressions representing meaning hypotheses generated by the perceptual/conceptual faculty, such as $f_1(x, f_2(f_3(y)))$. Word distribution in the corpus was controlled following Zipf's Law[5], so that a small number of words occurs much more frequently than most words, constituting a long-tail distribution. The use of artificial corpora is criticized by other researchers, such as Fazly, Alishahi, and Stevenson (2010), and I will comment on this in due time.

Siskind's pioneering model of lexical acquisition is a stepping stone in how to represent the issue computationally. If its artificial "corpus" of "utterance-meaning pairs" is accepted as a fair representation of children's experience, the model is then able to simulate lexical acquisition, albeit in an abstract way. It also shows that a cross-situational learning strategy, paired with some common-sense principles or intuitions, can produce similar behaviors to children learning their early vocabulary. The model's main characteristics are summarized in Table 1 below, which will be expanded to include each new model as we explore them.

Still, the model has been criticized for two reasons. First, it has been argued (YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010) that the use of synthetic corpora may misrepresent the task of lexical learning. Second, its heuristic rules have been considered too constraining

[5]Zipf's Law is given in some detail on page 73.

Table 1 – Summary of Siskind's (1996) model. Its main characteristics are being a deterministic online system for acquiring word-meaning pairings given a synthetic corpus of utterances paired with some meanings. The semantic representation is based on Jackendoff's (1983; 1990) conceptual semantics, thus being compositional and verb-dependent. Siskind's proposal models referential uncertainty, noise, and homonymy, as well as two different lexical categories: nouns and verbs. It is evaluated against the behavior children exhibit during their first years learning their language's lexicon.

	Siskind (1996)
Type	deterministic, online
Models	ref. uncertainty, noise, homonymy, lex. categories
Evaluation	behavior
Input	synthetic corpus of utterance-meaning pairs
Semantics	compositional (Jackendovian)

(FAZLY; ALISHAHI; STEVENSON, 2010). It may be the case that simpler, probabilistic mechanisms could yield comparable results. Next, I review two probabilistic models, each important in their own right, and then move on to new developments in the area challenging the very foundations on which those models have been built upon.

2.1.2 Two probabilistic models

Results in Yu and Smith (2007) have been taken as evidence for the idea that humans employ a cross-situational strategy in learning words; and based on a subsequent study (SMITH; YU, 2008), that children as young as 12 months exhibit the same ability. Building on this, Yu (2008) presents a computational model of word learning to attempt to quantitatively characterize the previous findings.

Yu's model's main difference is using a probabilistic mechanism to process a corpus of child-directed speech. Instead of building one-to-one associations between words and their referents, the model builds a system of associations in which a word-referent pair is correlated with other pairs sharing the same word or referent. Yu is especially concerned with the cumulative effects brought about by recruiting partial lexical knowledge, reacting to Bloom's (2000) criticism that probabilistic systems could not account for the child's few false positives when learning a lexicon. The researcher also points out that this is the first model to make use of naturalistic data, collected in the lab and processed manually.

Data collection started with the selection of six picture books which had their text removed. Caretakers were then instructed to tell a story to their 20-month-old children from the illustrations. Each book was considered an independent learning episode, adding up to six in total. The audio was then transcribed and paired to cooccurring referents, as available in the illustrations. Each utterance was defined to be the space between two pauses in speech. Input data was thus of the form {utterance, referents}. Notice that only

Figure 5 – Simplified example input to Yu's (2008) model. The utterance is paired with referents available on the left page. The only target word in this utterance is *fox*, and since the system has not had time to accumulate enough evidence, referents are equally likely for pairing. Furthermore, words which are not target for learning – including *late* – are also competing for pairing. Photo by Lina Kivaka (linakivaka.com).

the nouns which had an annotated referent were targeted for learning, however nontarget words in an utterance are also competing to be associated with the available referents. That is, each utterance was presented in full to the model and thus any of the words could conceivably map to the annotated referents; however, only a fraction of those mappings would actually be correct (see Figure 5 for an example). Further, an utterance may be referring to one, more than one or no available referents. In this way, the data present referential uncertainty and noise.

Yu also built into the model an assumption that the child identifies function words. For that reason, a referent `NON` was always made available with all utterances, such that associations between function words and `NON` could be built over time. A further assumption was that the learner is able to create an association matrix representing partial knowledge of possible word-referent pairings. This matrix is combined to the one previously generated as the result of processing a learning episode. In this way, the model is able to capture evolving knowledge about word meanings. Notice, however, that each episode is processed in a *batch*. This is a major difference from Siskind's *online* learner. In fact, most models update their lexicon after processing each utterance, a departure from Yu's, which must find the most likely lexicon given a whole episode.

Three pre-processing steps are applied to each utterance in order to reduce the size of the problem. They make use of partial knowledge accumulated up to that point. First, function words are removed by rules designed to identify them; next, pairs already learned are also removed; finally, association scores are ascribed based on the partial knowledge matrix. The model gets better at each of these steps as it processes more learning episodes.

Two conditions were tested. In the first, the six learning episodes are treated independently, whereas in the second, the model makes use of partial knowledge accumulated as it processes each episode. Yu reports that the model shows better performance under the second, cumulative condition, producing stronger correct as well as weaker incorrect

associations in comparison to the first condition. Most importantly, learning performance gets better from episode to episode, capturing the fact that children become better lexical learners with accumulated experience. The use of partial knowledge is accredited as responsible for this improvement. It guides the underlying learning mechanism, causing changes in its behavior without actually changing the mechanism's inner workings. Yu argues that this could explain vocabulary spurt and although the author suggests this is the first time such a claim is made, it seems that Siskind (1996) came to the same experimental conclusion in his study, just reviewed above. Yu's proposal is also capable of learning synonymous words due to an assumption that an object may be associated with multiple words within a single learning episode. Homonymous words can also be learned without any explicit rules other than an assumption that there is no homonymy within a single learning episode.

The use of partial knowledge is of paramount importance in both Yu and Siskind to reduce the space of hypothesis; this is the purpose of the three pre-processing rules. The only difference in Yu is there being a specific mechanism for dealing with function words. He observes that they can be singled out for having a strong association with `NON` and for occurring in several non-overlapping contexts. This last point means a word like "patient" might be expected to occur in the same contexts as "hospital." However, words like "the" and "for" occur in many more contexts. Siskind's model, on the other hand, does not assume anything about function words except that they are meaningless. This is, the model still has to learn that these function words are not to contain any meaning. Nevertheless, they are not leveraged in any special way to reduce referential uncertainty.

A property of its learning mechanism allows Yu's model to acquire homonymous and synonymous words. Synonyms can be learned by assuming that within a single learning episode multiple words may map into the same object. This makes sense when looking at the data, since adults may choose to talk about the same referent with different words: a dog in the story may be called "puppy," then "doggy," then "dog" and so forth. Assuming that synonyms occur temporally close may capture a cognitive bias or children's ability to notice this fact. Homonymous words, however, are assumed *not* to occur within the same learning episode. Again, this seems reasonable from the outside, since a story may refer to an "iron" for ironing clothes but is unlikely to then also mention the chemical element. Notice, however, that learning success in both conditions hinges on the fact that the learning mechanism processes a batch of utterances called a learning episode. This may not be cognitively plausible, since it is equivalent to saying that the child first accumulates in memory a number of utterances *and* corresponding extra-linguistic contexts, to then extract information about the words.

Another glaring difference between these two models is the fact that Yu's only considers noun-to-object mappings, or strict reference. Verb acquisition, for instance, is not accounted for. Although children do know and produce proportionally more nouns early on in their lexical development, "action" and "personal-social" words, among others, are unequivocally

Table 2 – Yu's (2008) model in a nutshell. Differently from Siskind's, it is a probabilistic system which processes the data in batch rather than online (utterance by utterance). Although it simulates synonym as well as homonym acquisition, it does not model noise (completely incorrect utterance-meaning associations). Rather, each utterance comes paired with a number of referents available when it was spoken. Some utterances will naturally come associated with referents for which there are no words available. As a result, it only captures concrete noun learning. Another shortcoming is that the meaning of each noun is represented by an atomic (i.e. non-compositional) symbol, such as DOG.

	Siskind (1996)	Yu (2008)
Type	deterministic, online	probabilistic, batch
Models	ref. uncertainty, noise, homonymy, lex. categories	ref. uncertainty, homonymy, synonymy
Evaluation	behavior	behavior
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs
Semantics	compositional (Jackendovian)	atomic

present (NELSON, 1973; BENEDICT, 1979). In fact, focusing on concrete noun acquisition is a defining trend for most of the models of lexical acquisition.

In some respects, Yu's model seems to have improved over its predecessor. Its probabilistic mechanism is arguably simpler than Siskind's heuristics, even though it does some pre-processing in order to reduce referential uncertainty by using accrued partial and full word knowledge. In a way, this phase of reducing uncertainty is comparable to Siskind's heuristics. The input is at once more realistic, since it was taken from child-directed (although controlled) speech, but also less realistic, since only concrete noun learning is investigated, whereas Siskind's synthetic corpus brings much more complex meaning representations. In fact, its simpler mechanism seems to stem from the fact that referent meanings, not full utterance meanings, are represented. Finally, perhaps its biggest shortcoming is the batch processing of each learning episode, as opposed to a more natural utterance-by-utterance processing. It does bring the advantage of making synonym and homonym learning possible without specific learning rules. However, this is at the cost of assuming less plausible constraints about the linguistic experience, such as that no homonyms will occur within the same learning episode. Table 2 adds a summary of this model before moving on to our next model, which addresses some of the issues just discussed.

Also employing a probabilistic strategy, but this time in an online fashion, Fazly, Alishahi, and Stevenson (2010) propose a model which learns a lexicon of word-referent mappings extracted from a corpus from the CHILDES project (MACWHINNEY, 2014). Input data looks very similar to what we saw towards the end of Chapter 1:

Utterance	"Joe is quickly eating an apple."
Scene	{JOE, QUICKLY, EAT, A, BIG, RED, APPLE, HAND}

Each scene is a set of atomic meaning symbols, representing concepts generated by the learner's mind. Referential uncertainty is modeled by including symbols not in the utterance. Half of all utterance-scene pairs contain some degree of referential uncertainty. That means 50% of the utterances are pristine learning opportunities, while the other 50% will contain a variable number of incorrect symbols. Across all utterances, there is an average of 78% extra symbols. They are not random but taken from the next utterance in the corpus, which is, for this reason, not included in the input to the model. Further, some pairs are noisy: a single random necessary symbol is removed in 20% of all utterance-scene pairs. Notice that noise in this proposal is not a complete lack of correct pairings available, like in Siskind's case, but rather the exclusion of one correct word-symbol mapping. A noisy pair is still comparatively highly informative.

The task is to find word-symbol alignments[6], considering partial knowledge gathered from previous observations. A word's meaning is represented by a probability distribution of associations between that word and all symbols with which it has cooccurred. The association score of a word and a meaning symbol varies from 0–1 to represent how sure the learner is of that mapping. As the model observes more utterance-scene pairs, some alignments are penalized while others are promoted. When this score reaches a threshold θ of 0.7, the word is said to have been acquired. This optimal threshold was defined empirically by testing various different values.

The model's performance is characterized in two groups of experiments. In the first, it processes 20,000 utterance-scene pairs, in a simulation of the child's task. The authors find that referential uncertainty negatively impacts performance more than noise: 70% of all words are learned with referential uncertainty at 78%, against 90% when referential uncertainty was null. No such effect was found in the presence or absence of noise.

The observation that children undergo a vocabulary spurt has gone through some scrutiny (BLOOM, 2000; GANGER; BRENT, 2004). The notion of a "spurt" or "explosion" implies that children move from a phase where learning new words happens at a slow pace to a phase of faster rate of learning. However, these authors argue that the definitions

[6] Although the utterances considered in this model include all word categories, the semantic representations associated to each utterance do not make any distinction between different types of meaning. That is, the concrete noun *dog* is associated to DOG just as *jump* is to JUMP. So, among other semantic phenomena, the model does not capture the fact that verbs express relations among nouns (i.e. are in general argument-taking). For this reason, I argue that this model only simulates "noun-like" acquisition.

of a spurt are based on thresholds such as "having acquired x number of words in y weeks." Since children must increase their rate of learning in order to reach the mark of 60,000 words by adulthood, they will inevitably go through an acceleration that may fit whatever threshold is defined. Ganger and Brent (2004) advance a new definition of spurt, based on fitting the data of cumulative vocabulary over time into two types of functions, logistic and quadratic. If a logistic function is a better fit for the data, then a mathematical spurt has been identified. Their results show that only 5 out of 20 children studied presented a spurt, indicating that it is not a ubiquitous phase in word learning.

Following this line of reasoning, Fazly et al. argue that this idiosyncratic behavior could be explained by how conservative a child is before employing a word. In other words, some children might have a bias against using words if they are unsure of their meanings, whereas others might be more lax. According to the researchers, this can be captured by their model's θ value. When $\theta = 0.7$, the model exhibits a curve of acquisition similar to a spurt. However, if θ is lowered to 0.5 – a less conservative threshold – the learning curve is more gradual, without a well-defined point of explosion.

Next, the second group of experiments was run. These consisted of exposing the model to 1,000 pairs, then giving it a task to solve. Three main findings are stressed. First, under conditions conducive to fast mapping, the model is capable of learning words with only one exposition. Second, learning homonyms is not possible under the current design, since the same word will have to split its association score with the two (or more) correct meanings. The authors recognize this limitation, blaming their definition of learning as θ being greater than 0.7. They argue this problem could be overcome by replacing the "current threshold-comparing mechanism with one that instead detects significant peaks in the probability distribution" (p. 1051). A last finding is that synonyms could be learned, since the model does not use information about other probable alignments when acquiring a new word. That is, the fact that *dog* is strongly associated with DOG says nothing about the pair *puppy*-DOG. However, since both words are present, the model has some initial difficulty in establishing the mapping, which mirrors child behavior according to studies cited by the authors.

Notice that Fazly et al.'s finding that referential uncertainty impacts learning more than noise is at odds with both Siskind's result and reasonable expectations. Noise is the equivalent of a child attending to a scene while the utterance she hears is talking about something else completely. While she might be playing with a toy in the back of the car, her mother might say "isn't gramma nice?," as they come back from tea break. Assuming a naive learner, like Fazly et al.'s, such pairings would be particularly devilish to learn. However, their modeling of noise is not realistic, erasing *only one* important word from the available meanings. It stands to reason that completely false pairings should produce a bigger effect. A computational study could establish whether this would be the case.

Fazly et al. also criticize (p. 1021) Siskind's need to include a mechanism specifically designed to learn under a homonymous corpus. They argue that their model could learn

homonyms if modified with a mechanism "that instead detects significant peaks in the probability distribution" (p. 1051) to consider a word learned. However, it should be noted that both are decisions arising from the difficulties imposed by homonyms. All things being equal, the researcher authoring a model has to consider this property of the input data and add provisions to it, or else propose a mechanism that can identify this fact independently. Hence, it might be a property of the mental lexicon – and thus an expectation from the learning strategy – that words might have several senses. On the other hand, Siskind's model says nothing about synonyms, whereas Fazly et al.'s can accommodate them quite naturally.

How have these two probabilistic models advanced from Siskind's original idea? It is tempting to say that both Yu's and Fazly et al.'s are simpler while also being able to learn homonyms (in Yu's case) and synonyms (in both cases). However, although probabilistic mechanisms are compatible with human cognitive capabilities, it should be pointed out that in Yu's case, experience is processed in a batch and that is what allows for learning homonyms and synonyms alike; and that in Fazly et al.'s case, the distribution of association scores which represents the meaning of a given word means that *all* possible alignments are kept in memory. Both problems do not exist in Siskind's proposal, which at the same time processes utterances one at a time and keeps a reduced number of meaning fragments under consideration for each word. Furthermore, Fazly et al.'s criticism that Siskind had to include a rule to specifically deal with homonyms seems to fall apart when we consider their own findings. Under their model, a homonym's association score becomes split between the two (or more) consistent meanings. Some kind of mechanism is thus necessary to posit a new word sense, ultimately leading to a multilevel lexical organization where words map into senses which in turn map into meanings. It is almost as if the hierarchical properties of the mental lexicon have to be given from the beginning, at least under the computational approaches under consideration.

Yet all three studies seem to, in aggregate, capture important facts about lexical development in the child. They stress the role cumulative knowledge plays in accelerating the rate of lexical acquisition and Fazly et al.'s as well as Siskind's are able to explain fast mapping. Their differences seem to boil down to implementation details: all three gather information as data is processed, becoming better learners at each stage but committing a varying amount of that information to memory. Siskind's seems to have more rules, however at the same time it models more complex meaning representations, homonymy, more realistic referential uncertainty and noise. The probabilistic models focus on noun (or noun-like) learning and as a result, do not have to consider competing utterance meanings or completely noisy pairings.

We have reviewed models which capitalize on the overall cooccurrence of words and referents. But what if there was another way, one that attempts to learn words from immediately available pairings, locally selecting referents and relying less on memory? The next section explores a challenge to the assumptions made so far, putting forward a more

Table 3 – Fazly et al.'s (2010) model summarized. Its main departure from the previous two is being online while also probabilistic, as well as being the first to use a corpus of child-directed speech. However, its meaning representation is the most naive among the models so far, modeling atomic meanings for each word (verbs, nouns, grammar words etc.) in the utterance.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)
Type	deterministic, online	probabilistic, batch	probabilistic, online
Models	ref. uncertainty, noise, homonymy, lex. categories	ref. uncertainty, homonymy, synonymy	ref. uncertainty, noise, synonymy
Evaluation	behavior	behavior	behavior
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech
Semantics	compositional (Jackendovian)	atomic	atomic

parsimonious idea that nonetheless seems to stand strong.

2.1.3 The local vs. global divide

As cross-situational learning has been painted so far, the developing lexicon is a set of mapping hypotheses being gradually refined. For each word, as many meanings as necessary are stored in memory and considered in parallel. However, that is about to change with Trueswell et al. (2013), who propose dramatically reducing the role of memory, turning a sort of fast mapping into the fundamental mechanism of lexical learning. In this view, the learner would propose a random mapping from the words and referents available in the scene and then verify that mapping subsequently. In case verification failed, the learner would then propose a new random mapping and this process would continue. This mechanism – which the authors call Propose but Verify (PbV) – might look too simple to work at first glance. However, it is in fact supported by psycholinguistic experiments carried out by the authors. This section discusses PbV in brief, then reviews an expansion as well as some criticism towards the proposal.

Propose-but-Verify

Trueswell et al.'s initial motivation comes from Medina et al. (2011), who found a curious learning pattern when applying the Human Simulation Paradigm, or HSP (Gillette et al., 1999). The HSP is an experimental protocol used to more realistically simulate the lexical learning task. The participant watches a muted video clip of a caretaker interacting with their child and is asked to guess what word might have been said when a beep is played

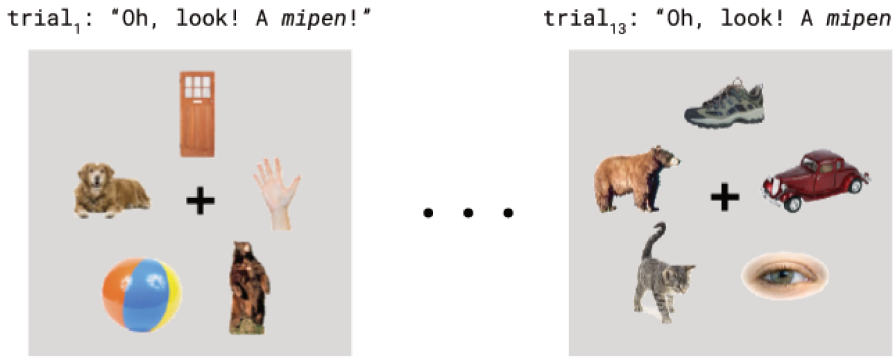
at some point in the video. Adults as well as children exhibit a better, albeit still dismal, performance for guessing nouns when compared to verbs: they are able to guess the correct word 50% of the times or more for only 7% of all nouns. That is, guessing what has been said is a hard task, even for learners already equipped with lexical knowledge. Medina et al. take the HSP and modify it. In place of a beep, a pseudoword such as "mipen" is heard. This allows participants to accumulate evidence as they watch more video clips.

Their unexpected result is that, upon correctly guessing what was said, participants tend to keep their interpretations in the next trials; however, whenever they guessed the wrong meaning, they tend to guess at random when re-encountering the same pseudoword. One would expect that participants' performance would get better as trials went by, if one assumed they kept a list of meaning hypotheses for every pseudoword. That is not what was found: participants do not seem to have memory of previous cooccurrences. When they guessed incorrectly in a previous trial, their correct guesses in the next trial were at around 11%. This result is comparable to the 9% correct guesses found in participants who only heard a beep, who were thus not able to make cross-situational inferences.

However, according to Trueswell et al., evidence from Medina et al.'s (2011) experiment could not allow the different cross-situational mechanisms proposed in the literature to be satisfactorily compared. The authors criticize the fact that experiments like the HSP place the participant as an observer rather than learner; the fact that controlling size and salience of the stimuli under consideration is not possible; and also argue that experiments in favor of a strong cross-situational mechanism, such as Yu and Smith (2007), do not observe per-trial participant behavior, but rather a final aggregate. Considering only the final average performance may mask initially random, PbV-like behavior which in the end looks like a gradual improvement in performance. Therefore, a series of experiments is proposed to investigate the matter more appropriately.

The experimental design consisted of presenting 12 pseudowords and 12 corresponding objects in five blocks, in a total of 60 trials. Thus, each word had a gap of 11 trials in between presentations (see Figure 6). Participants were instructed to try to determine the meaning of each pseudoword from the objects available in the screen. Referential uncertainty, that is, the number of objects available, was kept constant at 5. The correct object was always made present, such that noise rate was null. The results showed that the learning curve improved after each block, that is, that the participants were gradually getting better at guessing. Crucially though, only after having guessed the correct object would participants guess above chance in the next trial. In other words, upon realizing their mistake, participants did not show any evidence of using other hypotheses stored in memory. This suggests that only one mapping hypothesis is considered at each time. Two variations of the experiment were also carried out in order to test the possibility that there were too many objects in each trial for participants to remember all possible mapping hypotheses. These subsequent experiments decreased the degree of referential uncertainty, however the results still pointed in the same direction.

Figure 6 – Five referents are presented as the subject hears a sentence containing the pseudoword, in this case *mipen*, meaning "bear." Each word is presented five times, with 11 trials in between each presentation. For all words, the order of presentation remains the same throughout the experiment. However, the cooccurring referents change, except for the correct one. Figure adapted from Trueswell et al. (2013, p. 131).



PbV was then implemented as a computational model to verify the author's proposal that it could capture the experimental results described above. The algorithm:

1. Randomly selects a referent available in the scene for the word under consideration,
2. Upon re-encountering that word, remembers the mapping formed previously with a probability α,
3. If the mapping is confirmed, that is, if the referent is in the current scene, α is incremented and the referent is selected; otherwise, a new referent is randomly selected.

α is the only free parameter in the model and represents the probability of remembering a previous mapping hypothesis. It was set according to the behavior observed in the experiments just described. The model did reproduce the learning curve as well as behavioral patterns exhibited by humans.

This almost trial-and-error learning mechanism, paired with the findings from the HSP, seem to suggest that the lexical learning task should be a slow process. Then how do children develop their lexicon quickly and without trouble? Trueswell et al. argue that the source of difficulty seen in the experiments above is, in fact, the lack of the contextual information the child normally has access to, such as joint attention, the meaning of other words in the utterance, syntax etc. Thus, the HSP would be a fair simulation of the initial stages of word learning, when children do not yet have unrestrained access to these channels of information. During these early phases, children do indeed have a slow rate of word acquisition. In this case, PbV could be the underlying learning mechanism.

However, as we will see next, there still are severe limitations which could mean this simple model's implausibility. For example, how could it explain homonym learning, or resistance to noise-induced errors? If the child always overhauled her knowledge of a word given each and every piece of negative evidence, the task could be rendered impossible.

Pursuit

To address PbV's limitations, Stevens et al. (2017) put forward an updated version called Pursuit. They also introduce terminology to set these proposals, dubbed *local*, apart from *global* models like the ones reviewed in the previous sections. Pursuit, like PbV, also randomly selects a referent. However, more in line with global models, instead of throwing away the current hypothesis in face of negative evidence, the model punishes it and adds a new random referent into the list of possibilities. This small change allows Pursuit to resist noise and conceivably learn homonyms.

Pursuit, along with PbV and two global models, the original probabilistic model from Fazly, Alishahi, and Stevenson (2010) and a modified version, were evaluated in two simulations of lexical acquisition on different corpora. The first simulation used the Rollins Corpus from CHILDES (MACWHINNEY, 2014) as input. Data consisted of child-directed speech from caretakers playing with their infants with toys from a box, one at a time. Each utterance was manually paired with concrete nouns judged to be available to the learner. In total, 680 utterances were annotated, 496 for the training set and 184 for the evaluation set. The models were evaluated against separate data sets in order to control for over-fitting, that is, the fact that statistical models can become so well-fitting to a particular data set that they fail to make predictions. Results were given as precision (the proportion of words correctly learned), recall (coverage of how much of the lexicon was learned) and a synthesizing F-score. Among all models, Pursuit showed the best precision and the best F-score, while PbV had the best recall. Both local models had better performance than the global models, even though the latter gather more data.

A second simulation used the HSP corpus from Cartmill et al. (2013), having simpler data than the previous one. 560 video clips of 50s each were transcribed. In each clip, the caretaker says 1 out of 41 concrete nouns. While the average referential uncertainty degree from the last corpus was 3.1 referents/utterance, Cartmill et al.'s had an average of 7.4 referents/utterance. Results showed that Fazly et al.'s global model is better than Pursuit (F-score of 0.41 against 0.39). According to the authors, a plausible explanation might be the higher noise rate found in this experiment's corpus: 40% against 10% in the previous corpus. However, Pursuit was shown to be more robust against the interaction of higher noise and higher referential uncertainty than the global model when tested in other experiments. The authors then suggest that in the Rollins Corpus, greater cooccurrence of *meanings* (that is, CAT and DOG cooccur more frequently) is such that the global model has to distribute mapping weights more sparsely, whereas Pursuit only tests one hypothesis at a time, being blind to this competition for weights. According to the authors, the greater meaning cooccurrence is a natural property of scenes in the real world. In other words, the task using Cartmill et al.'s corpus is less ambiguous, thus favoring the global model.

Having shown that Pursuit is competitive against global models when learning a lexicon from a corpus of utterances and cooccurring referents, the authors then turn to another

Table 4 – Pursuit summarized and compared to other models. Two main differences can be seen: it is a local model, assuming much less memory and a totally different learning mechanism than its global counterparts; and it also introduces the idea of using psycholinguistic tasks in order to evaluate the model.

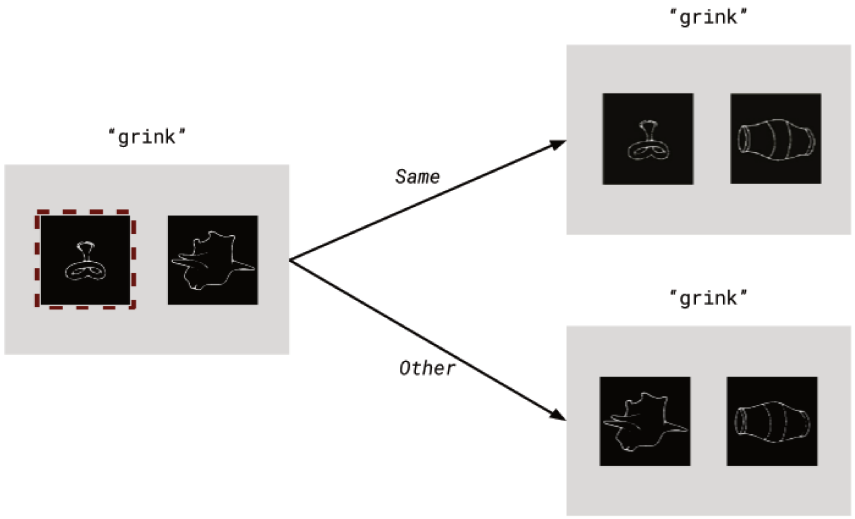
	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit
Type	deterministic online global	probabilistic batch global	probabilistic online global	probabilistic online local
Models	ref. uncertainty noise homonymy lex. categories	ref. uncertainty homonymy synonymy	ref. uncertainty noise synonymy	ref. uncertainty noise
Evaluation	behavior	behavior	behavior	behavior experiment
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli
Semantics	compositional (Jackendovian)	atomic	atomic	atomic

set of experiments designed to test its plausibility. This was done by running simulations of a series of psycholinguistic tasks. The reasoning behind this is that human performance sets a benchmark that models can be compared to.

Three different tasks were simulated: Yu and Smith (2007), the first to offer experimental evidence for cross-situational learning; Trueswell et al. (2013), just discussed; and a similar study from Koehne, Trueswell, and Gleitman (2013). I will return to these three studies in depth in section 2.2, where I present results from an implementation of Siskind's (1996) model run against some of the same experiments. For now, suffice to say that in all three tasks, both PbV and Pursuit showed behavior closer to the benchmark set by human participants, whereas the global models predicted performance above human average. Pursuit was also the only model to capture the results found by Koehne et al.'s study, the last of the three. The authors conclude, then, that Pursuit is better placed as a mechanistic explanation of human word-learning behavior, while also showing good performance learning a lexicon from a corpus. More fundamentally, they argue that "a local model that keeps track of few options is better equipped to capitalize on the rare but highly informative learning instances, which are diluted under models that keep track of all options" (p. 31), a point echoed in Yang's (2019) ample argument in favor of a local approach to lexical acquisition. I reference this later in Chapter 4.

Before we move on, briefly consider Table 4. Pursuit, but not PbV, has been added to it, since the first is an updated version of the latter. Notice further that new *global* or *local* tags have been added to all model types to identify where they fall in this respect.

Figure 7 – This experiment allows for two types of inferences. In the Same condition, the referent selected by the subject (indicated in red) is repeated in the next screen. In the Switch condition however, the referent *not* selected is presented, allowing the subject to draw on previously formed hypotheses. Figure adapted from Yurovsky and Frank (2015, p. 55).



Challenges to the local account

Despite the supporting evidence just cited, the idea that word learning is mainly based on a local selection mechanism has also seen some challenge. Yurovsky and Frank (2015) present two experiments to test whether adult participants solving a word learning task may commit more than one mapping hypothesis to memory. The authors argue that the conflicting findings from previous experiments may suggest that learners switch between two mechanisms, one local and one global, depending on the complexity of the situation at hand. Two ways in which complexity is manifested are referential uncertainty and the interval before a word is shown again; thus, their experiment manipulates these two variables to test the hypothesis that an "integrative account" might better explain lexical acquisition. The evidence seems to confirm that that is the case, introducing a problem to models like PbV and Pursuit.

In the simplest mix of conditions (Experiment 1), the participant hears a pseudoword while two distinct images are presented in the screen. They then have to select the referent they think matches the word, before moving on to the next screen, which falls under one of two types of conditions (see Figure 7). If it is of the *Same* condition, then the referent the participant selected is shown again, mixed with a completely new referent. If however the next screen falls under the *Switch* condition, the referent *not* selected will be shown, mixed with a completely new referent. Thus, in the Same condition, a local learner is able to confirm the mapping just established. However, the Switch condition allows another kind of inference – but only if the learner kept a memory of the other possible mapping.

In the experiment, participants are first shown four practice trials with familiar words and referents. Then, they are instructed to keep choosing correct referents, but this time

with novel pseudowords and unknown referents. Eight novel words are shown twice each. Four of the follow-up trials fall under the Same condition, whereas the other four fall under the Switch condition. This way, each participant's performance can be compared across conditions and changes in their learning strategies can be observed. A total of sixteen mixes of conditions were tested. Referential uncertainty, that is, number of referents in the screen, varied among 2, 3, 4 or 8. The interval between two presentations of the same word also varied, being back-to-back or intervened by 1, 2, 4 or 8 other words.

Participant's rate of correct answers across all variations in referential uncertainty remained better than chance in the Switch condition, which means they have stored information about previously considered mappings other than the mapping explicitly chosen. However, the number of referents in the screen affected their performance: the higher the uncertainty, the worse their chance of remembering alternative mappings. However, the best rate of correct answers appeared in the Same condition, suggesting that although participants can encode alternative mappings, it was their explicit choice which held the strongest association. Furthermore, in the Same condition, the interval between first and second presentation of that word had the biggest impact on performance, evidencing a memory effect.

The authors then present a computational model (see Table 5) based on Frank, Goodman, and Tenenbaum (2009), discussed ahead[7]. Three versions are implemented: a Statistic Accumulation model, which considers all referents in the scene as possible; a Single Referent model, which selects a referent as being the most plausible; and an Integrated model, which assigns a stronger probability score to a single referent and the remaining probability to the rest of the referents available. A memory effect is also added so that the lexical entries are progressively forgotten. The model thus chooses according to memory (a global strategy) or, when it has no memory of a word, randomly chooses a referent (a local strategy). Yurovsky and Frank report that the Statistic Accumulation model does not predict the differences observed in the Same and Switch conditions, whereas the Single Referent model does not predict the above-chance behavior seen in the Switch condition. However, the Integrated model predicts both behaviors, being worse than human participants only when referential uncertainty is 3 or 4 in the Switch condition. Building on these results, the authors suggest that "learning is fundamentally distributional, but the fidelity of learners' distributional estimates depends critically on their limited attention and memory" (p. 60). Because of this, complex situations with high referential uncertainty make the learner behave more locally, whereas simpler ones allow the learner to infer more word-referent mappings.

Yurovsky and Frank's findings suggest humans can encode several mappings at once and are a thorn in the side of local models of word learning. Stevens et al. (2017) take the challenge head on, acknowledging such results are left unexplained by Pursuit and

[7]It should be noted that Frank, Goodman, and Tenenbaum (2009) is a batch model of word learning, similar to Yu (2008). This modeling decision has received some criticism, as discussed in section 2.1.2.

Table 5 – Summary of Yurovsky and Frank (2015). This model implements a hybrid global-local strategy that better captures the results found in experimental tasks. In addition, the model implements a memory effect so that a lexical entry is progressively harder to remember.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit	Yurovsky and Frank (2015)
Type	deterministic, online, global	probabilistic, batch, global	probabilistic, online, global	probabilistic, online, local	probabilistic, batch, local, global
Models	ref. uncertainty, noise, homonymy, lex. categories	ref. uncertainty, homonymy, synonymy	ref. uncertainty, noise, synonymy	ref. uncertainty, noise	ref. uncertainty, memory
Evaluation	behavior	behavior	behavior	behavior, experiment	experiment
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli	simulated experimental stimuli
Semantics	compositional (Jackendovian)	atomic	atomic	atomic	atomic

should be further explored. However, the authors also argue that Yurovsky and Frank's experiment, beyond reducing the interval between same-word exposition, more importantly unrealistically simplifies the task by showing an object in only one context. In real life settings, a word learner is likely to see an object in many different contexts where it is not being referenced; yet that does not happen in the experiment, which creates an "optimal" situation for "recalling the prior context" (p. 34). Pursuit is defended as being the "idealized" core learning mechanism and it is possible that under the correct conditions, memory mechanisms kick in, aiding learning. There is also evidence that 2–3 year-olds do not seem to remember mappings other than the selected one (WOODARD; GLEITMAN; TRUESWELL, 2016) in a child-friendly version of Yurovsky and Frank's experiment.

It should be clear by now that with the introduction of local models, memory has become a central issue. Recall that Siskind's (1996) model includes a garbage collector mechanism designed, among other things, to make learning under conditions of homonymy and noise possible. Furthermore, Siskind's model is different from the models discussed so far in that it keeps a table of *possible* meanings taken from the first encounter with a word. Relatively few meaning fragments are thus being considered at each time, unlike the other global models, which update their probability distribution to include newly-observed word-referent mapping hypotheses. These differences suggest that an investigation of Siskind's model learning under simulated experimental settings like those in Stevens et al.'s study is warranted. The next section discusses how Siskind's model was reimplemented and adapted to investigate its behavior in these simulations.

2.2 Could Siskind (1996) still be relevant?

So far, we have reviewed a number of computational approaches to cross-situational strategies of word learning. In effect, the ground has been set for a comparison of Siskind's (1996) work with the results found in the literature of lexical acquisition modeling. More specifically, I touched on the fact that two features of his model set it apart from the other proposals. Namely, these features are: the way words being acquired have potentially less dispersed mapping hypotheses compared to the other global models, and the garbage collector, which acts as a forgetting mechanism. Given these contrasting characteristics, the question then is: how would Siskind's model fare in the experiments simulated by Stevens et al. (2017)? In this section, I start by explaining the four core rules proposed by Siskind, their inner workings and how they are able to acquire word meanings. Next, the original forgetting mechanism is explained in some detail. Then, I discuss my reimplementation of the model and what adaptions had to be made. Finally, I end by reporting the results obtained by running this model against two of the three experiments simulated by Stevens et al.: Yu and Smith (2007) and Trueswell et al. (2013).

2.2.1 Siskind's heuristics

Siskind's original heuristics are composed of six rules that update lexical entries as utterance-meaning pairs are processed. As a quick recap, in its most basic form, each lexical entry for a word is composed of two tables, one for storing (observed) possible meanings for a given word, $P(w)$, and another for storing (inferred) necessary meanings, $N(w)$. The first four rules are intended to reduce referential uncertainty, remove incorrect meaning fragments from $P(w)$ and add them to $N(w)$. Their operation is enough to acquire all the fragments which add up to be the meaning of a word. The last two rules determine how these meaning fragments assemble into a full conceptual expression. For example, rules 1–4 will determine the meaning fragments of *walk* to be the set {GO, TO}; once this is determined, rules 5–6 find its structure, GO(x, TO(y)).

Since the experiments simulated here investigate word learning isolated from the sentencial context and since the simulations in Stevens et al. (2017) implement atomic (that is, non-compositional) meaning representations, only the first four rules are enough to implement an adaptation of Siskind's model. That means the problem has been reduced to finding mappings of the kind *word* → CONCEPT.

In order to understand how these four heuristic rules can learn word-concept mappings, let us consider some simplified examples and then formally specify each rule following Siskind (1996). At each processing step, the model gets two words (an "utterance") paired with four referents (a "scene"). Since the correct referents are always available, noise rate is zero; further, referential uncertainty is of two extra meanings. The task is to learn the target lexicon $[w_1 \rightarrow r_1, w_2 \rightarrow r_2, \ldots, w_n \rightarrow r_n]$. Assume for now that the lexicon generating these two-word utterances has no homonyms and further that the model's lexicon is not null, but contains the entry $w_3 \rightarrow r_3$. The following learning instance is processed:

(1)

Utterance	w_1	w_2
Scene	$\{r_1, r_2, r_3, r_4\}$	

The model has no knowledge of these two words, so they are assumed to possibly refer to anything available in the scene. The current lexicon then is:

	$N(w)$	$P(w)$
w_1	$\{\}$	$\{r_1, r_2, r_3, r_4\}$
w_2	$\{\}$	$\{r_1, r_2, r_3, r_4\}$
w_3	$\{r_3\}$	$\{r_3\}$

So far, no more information can be extracted from this pair, so it is discarded. Next, the model gets:

(2)

Utterance	w_1	w_2
Scene	$\{r_1, r_2, r_7, r_{11}\}$	

At this time, Rule 1 can be applied in order to reduce uncertainty. It is formally defined as follows (p. 57):

Rule 1 Ignore those utterance meanings that contain a conceptual symbol that is not a member of $P(w)$ for some word symbol w in the utterance. Also ignore those that are missing a conceptual symbol that is a member of $N(w)$ for some word symbol w in the utterance.

Since both r_7 and r_{11} are missing from the possible meanings of w_1 and w_2, these cannot possibly be utterance meanings and are thus left unattended by the learner. The scene has been reduced to mean $\{r_1, r_2\}$. Now, Rule 2, intended to remove possible meaning fragments, can be applied. Rule 2 states that:

Rule 2 For each word symbol w in the utterance, remove from $P(w)$ any conceptual symbols that do not appear in some remaining utterance meaning.

Recall that so far the possible meanings of w_1 and w_2 are identical: the set $\{r_1, r_2, r_3, r_4\}$. By applying Rule 2, the inference can be made that r_3 and r_4 cannot possibly be meanings of these two words. This yields the lexicon:

	$N(w)$	$P(w)$
w_1	$\{\}$	$\{r_1, r_2\}$
w_2	$\{\}$	$\{r_1, r_2\}$
w_3	$\{r_3\}$	$\{r_3\}$

Next, we have the pair:

(3)

Utterance	w_1	w_3
Scene	$\{r_1, r_3, \not{r}_{13}, \not{r}_{17}\}$	

Rule 1 reduces referential uncertainty by removing r_{13}, r_{17}; Rule 2 then removes r_2 from $P(w_1)$. Now, Rule 3 allows the model to make a further inference: that r_1 is a necessary meaning of w_1. It states that:

Rule 3 For each word symbol w in the utterance, add to $N(w)$ any conceptual symbols that appear in every remaining utterance meaning but that are missing from $P(w')$ for every other word symbol w' in the utterance.

Since r_1 is missing from the possible meanings of w_3, the model is able to infer that it should be added to $N(w_1)$: w_1 uniquely contributes that meaning, so it *must* be a necessary part of that word's meaning. Notice that since the possible and necessary meanings are identical, the model has converged on the meaning of that word. We now have the following lexicon:

	$N(w)$	$P(w)$
w_1	$\{r_1\}$	$\{r_1\}$
w_2	$\{\}$	$\{r_1, r_2\}$
w_3	$\{r_3\}$	$\{r_3\}$

Next, the model processes the pair:

(4)

Utterance	w_1	w_2
Scene	$\{r_1, r_2, \not{r}_3, \not{r}_4\}$	

Again, Rule 1 reduces referential uncertainty; Rule 3 then adds r_2 to $N(w_2)$, since it can only be contributed by that word. Notice, however, that w_2 has not converged yet and that Rule 2 cannot be applied to remove the meaning r_1 from $P(w_2)$. Rule 4 lends a hand:

Rule 4 For each word symbol w in the utterance, remove from $P(w)$ any conceptual symbols that appear only once in every remaining utterance meaning if they are in $N(w')$ for some other word symbol w' in the utterance.

Since r_1 appears only once in the scene and is necessarily contributed by w_1, the model can infer that it cannot possibly be part of w_2's meaning. The final lexicon is:

	$N(w)$	$P(w)$
w_1	$\{r_1\}$	$\{r_1\}$
w_2	$\{r_2\}$	$\{r_1\}$
w_3	$\{r_3\}$	$\{r_3\}$

All words have converged on their meanings. These four rules are enough to acquire words in two of the three experiments simulated by Stevens et al. (2017) and discussed below. That is because Koehne et al.'s experiment (2013) seems to have some degree of homonymy. Homonymous words, as explained elsewhere, are not acquirable under these strict heuristics. In some specific cases, Siskind's model is able to determine whether it might have encountered a situation in which proposing a new word sense is called for. I discuss how this proved (at least for now) hard to model in section 2.2.5. For the first two experiments, the underlying acquisition mechanism works as described in this example.

2.2.2 Selecting referents and forgetting meanings

One particular kink needs to be ironed out before presenting the simulation results. The real-life versions of the experiments simulated here tested participant performance by providing some kind of referent selection task. Participants' answers were compared to the target lexicon to get a measure of correct answers. In the computational studies carried out by Stevens et al. (2017) (see discussion in section 2.1.3), referent selection for global models

was probabilistic: since each word meaning is a probability distribution of mappings, the mapping with the highest score would be the most likely to be chosen, followed by the second highest scored mapping and so on. PbV, on the other hand, always has a single referent under consideration, which would be that model's only choice. Pursuit has a single preferred referent hypothesis as well, but can represent other meanings under consideration. Siskind's model, on the other hand, has no provisions for selecting a referent. Since some experiments, like Trueswell et al. (2013), ask participants to explicitly choose a referent at each trial, Siskind's proposal had to be extended to model referent selection.

The current selection mechanism is quite unsophisticated and exploits the fact that each lexical entry has two tables for storing meanings. In the experimental iterations so far, the algorithm has been: select the meaning available in $N(w)$ whenever possible; if it is empty, then randomly select one meaning in $P(w)$. This is meant to model a degree of sureness: if the learner has determined some necessary meanings, then always choose from that pool. Since meaning representation is atomic in these simulations, $N(w)$ will always contain exactly one meaning. If, however, the learner has not determined the necessary meaning, then this could be seen as representing confusion about that word's meaning: thus, choose randomly from the possible meanings still under consideration.

Other mechanisms could be devised. One variation could exploit, for example, the fact that words are only acquired once the two tables have converged. I could use the moment when a meaning has entered $N(w)$ but $P(w)$ still has some other members to assign a higher weight to the necessary meaning (say, make it twice as likely), but still model some possible confusion. This is meant to capture memory, which is a factor built into the operation of models like PbV and Pursuit. However, my current adaptation of Siskind's model does not explore these alternative selection mechanisms.

Memory – more specifically forgetfulness – is a factor which must influence lexical acquisition. In the original simulations of PbV, (TRUESWELL et al., 2013), a recalling parameter was set according to the chance a human participant had of remembering a correct mapping selected previously. This parameter was increased if the second encounter lead to another correct selection. In Pursuit (STEVENS et al., 2017), memory is modeled by awarding or punishing mappings, such that associations will become stronger or weaker. In the global models implemented by Stevens et al., memory is not modeled directly as far as I can see, however since a word's meaning is a probability distribution, the learner has some probability of choosing the wrong mapping. If a global model is to be compared to results obtained by forgetful models, it has to somehow account for such memory effects.

In Siskind's original work, a "pruning mechanism" (p. 65) periodically discards spurious word senses. Some background is necessary to understand this: to deal with inconsistent utterances, the model is able to propose as many new word senses as necessary. Inconsistency can arise in some situations, notably when a noisy utterance-meanings pair is presented or when an unknown word sense is encountered by the model. Say "the boy ducked behind the chair" is given, paired with the correct meaning representation. However, because the

current lexicon only includes the noun *duck*, Rule 1 will discard the pair as noise. For that reason, the model is able to propose a $duck_2$. Nevertheless, since noisy situations are indistinguishable from homonymy, the system will inevitably propose new spurious word senses in isolated presentations; Siskind's insight is then to distinguish them using the history of presentations: noise is expected not to recur, while consistent alternative senses are. After processing 500 utterances, the model discards word senses not frozen. Senses are frozen if they meet two conditions: they have converged on their meaning *and* have been used to explain an utterance meaning more than twice[8]. In this way, spurious senses are periodically removed from the lexicon thanks to pruning.

In the experimental simulations below, there is no room for waiting for 500 utterances before purging potentially spurious words, since the longest experiment is 60 trials long. On somewhat the opposite direction, each "utterance" has few words and referential uncertainty is low, so that the task is simpler than Siskind's original simulations. Memory is thus modeled quite simply: forget unfrozen words after some number k of trials. This is then a free parameter and its value is not trivial to determine. For that reason, several values of k were tried in order to test its effect. This is not directly comparable to how local models forget mappings, however it is the most straightforward way of implementing memory while following Siskind strictly.

What follows is the characterization of a reimplementation of Siskind's (1996) lexical acquisition model[9] in two simulated psycholinguistic experiments, similar to what Stevens et al. (2017) did. They are reviewed and explained, their computational reproduction is discussed and simulation results are reported. A third experiment was originally set to be simulated, however for technical reasons discussed in section 2.2.5, this proved hard under the current adaptation of Siskind's model.

2.2.3 Simulation of Yu and Smith (2007)

Although the idea of cross-situational learning had been extensively proposed in the word learning literature previously, actual human capabilities had not been put to test, at least for word learning tasks. Yu and Smith (2007) scored the first to carry out experiments to investigate to what extent humans are able to learn words in ambiguous contexts. The task was divided in two parts: first, participants observed cooccurring words and their referents. Then, they were asked to select the correct referent of a word given a choice of four, including the correct one. Participants were above chance in completing this test and, as a consequence, this work is often cited as empirical evidence in favor of cross-situational learning in humans. For this reason, it was chosen to be one of the simulations run by

[8]These values are rather arbitrary and were set manually. In particular, the low confidence factor of 2 reveals that a sense has to have been employed twice in explaining utterance meanings before it can be frozen. This suggests spurious senses created by noise and homonymy are quite rare.

[9]The code for the model and experiments is available at IEL-NLP at Gitlab.

Figure 8 – 54 pseudowords were presented, divided in three diferent conditions represented below. Each word was seen 6 times. All participants saw all conditions, but words were different in each condition.

Stevens et al. (2017) in their comparative study of global and local models of lexical acquisition.

In the first series of experiments, participants were tasked with learning 54 pseudowords in different ambiguity conditions. 18 items were presented per condition and participants encountered each word 6 times. Ambiguity conditions were 2 × 2 (two words and two referents), 3 × 3 and 4 × 4 (see Figure 8). Each trial consisted of words shown on a computer screen and pseudowords generated by a text-to-speech system being played. At the end of each of the experimental conditions, participants were assessed to determine how many words were learned correctly. In this test phase, participants saw each of the 18 words paired with the correct referent, plus three distractors. Results were then compared across ambiguity conditions and it was found that higher ambiguity conditions correlated with proportionally less correct mappings being learned.

The researchers then ask themselves whether spurious correlations might affect participant performance. Because the 18 word-referents pairs are randomized and presented in the trials, some cooccurrences might lead to incorrect meaning hypotheses. If these spurious mappings are then made available in the test phase, participants might choose such wrong referents. To test this possible effect, the researchers ran a second series of experiments with within-trial ambiguity kept at 4, since the 4 × 4 condition had the highest occurrence of spurious correlations. The number of words and repetitions, however, varied. Three conditions were tested: 9-words/8-repetitions, 9-words/12-repetitions and 18-words/6-repetitions – which reproduced the last condition of experiment 1. One might expect learners to do better in the 9-words/12-repetitions condition, since there are less words to learn and more opportunities to do so. However, as the authors point out, there the number of spurious cooccurrences is highest. Thus, to a statistical mechanism, this should be the poorest of all conditions for learning. Contrastingly, the results showed that participant performance does not seem to be considerably affected by these extra referents

Table 6 – Proportion of correct mappings learned by models in Stevens et al. (2017) and by the memory-restricted adaption of Siskind (highlighted in bold) in different conditions of ambiguity. Memory was restricted by setting $k = 8$, that is, forgetting non-converged words after 8 trials. This adaptation of Siskind more closely resembles the behavior of local models (PbV, Pursuit) than of global models.

	2×2	3×3	4×4
PbV	0.76	0.63	0.54
Pursuit	0.96	0.84	0.71
Fazly et al. (2010)	0.99	0.98	0.98
Modified Fazly	0.99	0.97	0.96
Memory-restricted Siskind	**0.98**	**0.72**	**0.60**
Originally reported	0.89	0.76	0.53

competing for mappings, suggesting that within-trial ambiguity is a more impactful factor than spurious correlations.

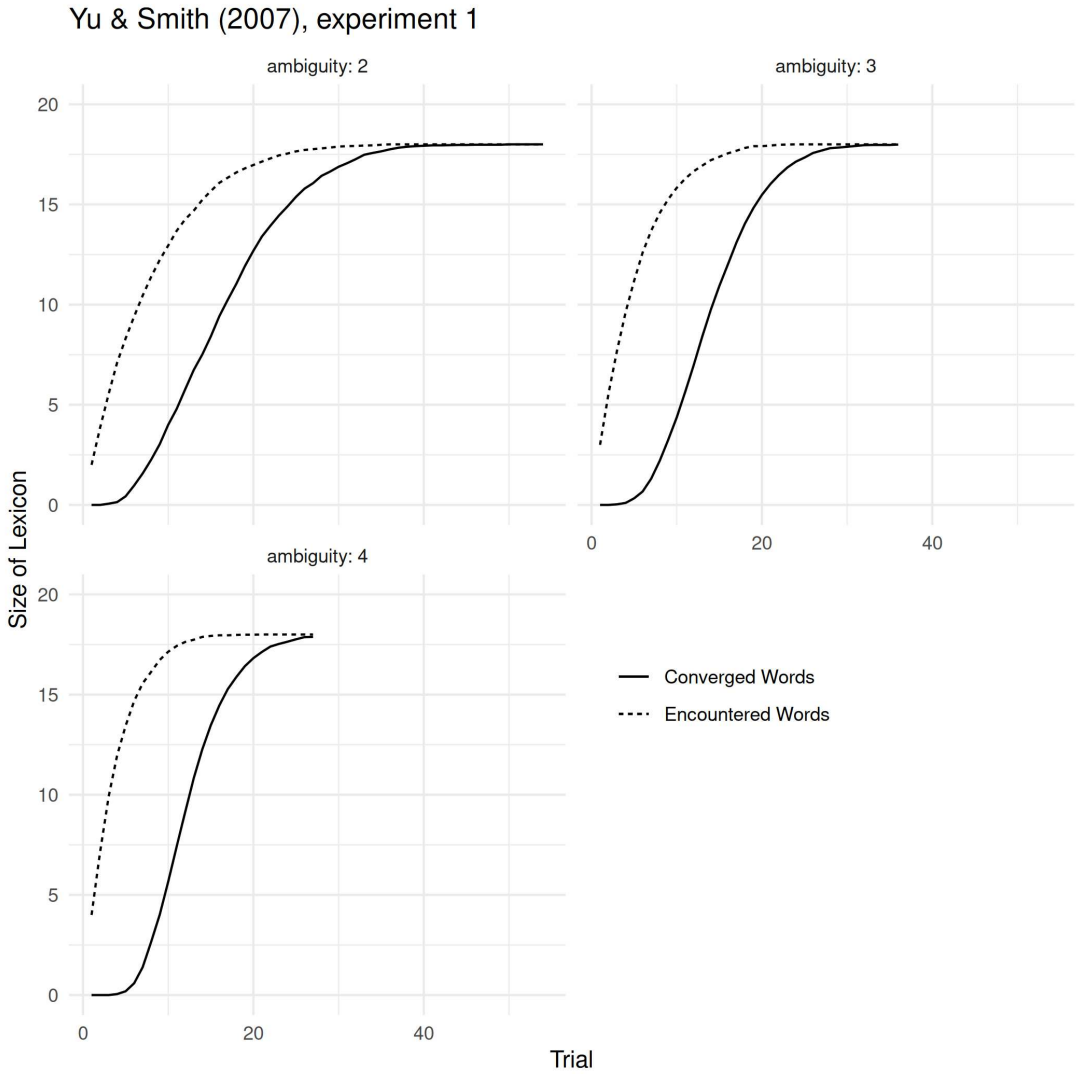
Computationally, each trial in this experiment can be represented as a list of words paired with a list of referents. For example, a 2×2 trial is represented by the lists $[w_1, w_2], [r_1, r_2]$. A small program was implemented to generate these trials automatically following the specifications in the paper. This input was then fed into the adapted Siskind (1996) model. Each condition was run 100 times and results were averaged, producing the data reported below, which are the findings for simulations of experiment 1. Let us start by looking at the "perfect memory" version of the model.

Across all three conditions, the model was able to acquire all 18 word-referent pairs. The first mappings were established around trials 6–7 and all mappings had already been acquired by the time the model reached around 3/4 of the experiment. The learning curve is illustrated in Figure 9. The fact that the model behaved similarly in all three conditions suggests the degree of ambiguity does not affect its learning capabilities. This makes sense, since this first simulation does not include Siskind's original pruning mechanism which, as I argued above, can be seen as the natural forgetting associated with memory-demanding tasks.

The experiment was then rerun with the pruning mechanism activated. Recall from above that in this adaptation of Siskind's model, words are forgotten if they have not converged after k trials. Several values of k were tested and the best was chosen, as is standard for free parameters in modeling work. Values of k tested included 3, 4, 5, 6, 8 and 10. Compared to other models (STEVENS et al., 2017) and reference results with human participants (YU; SMITH, 2007), the best fitting value of k was found to be 8. The proportion of correct mappings for $k = 8$ is given in Table 6, with comparison values for the models as well as the results reported in the original experiment.

This memory-restricted version of adapted Siskind behaves more similarly to local models like PbV and Pursuit than to the global models. If forgetting is removed, essentially 100% of the words are acquired irrespective of condition, as seen above, approximating

Figure 9 – Learning curve of adapted Siskind in simulated Yu and Smith (2007), experiment 1. Notice that despite increasing difficulty, the shape of the lines look the same. This goes against the benchmark assumed here, that is, the results reported for humans: as ambiguity goes up, difficulty of learning is impacted. Lines look shorter because as the degree of ambiguity increased, there were less trials in the experiment. In the 2 × 2 condition, there are 54 trials; in 3 × 3, 36 trials; and in 4 × 4, 27 trials.



adapted Siskind to the other global models. This suggests that an underlying global learning mechanism which forgets word-referent mappings as a byproduct of memory can, at least in principle, fit the results found by Stevens et al. (2017). In fact, this is a possibility anticipated by those researchers, who recognize that "*post hoc* memory constraints could be imposed on a global learner to make the numbers match better" (p. 24). The difference here is that the memory constraint was not imposed on this adaptation of Siskind, but rather is an essential feature of his model. It is the interplay of the heuristics and the pruning mechanism that assures spurious senses will not be added to the lexicon, potentially harming other word mappings and licensing erroneous meaning interpretations, while also assuring homonyms can be acquired.

Table 7 – Proportion of correct mappings learned by the memory-restricted SAM ($k = 8$) in the different conditions of experiment 2 of Yu and Smith (2007).

	9-words/8-repetitions	9/12	18/6
Memory-restricted Siskind	0.97	1.00	0.58
Reported	0.57	0.61	0.52

Interestingly however, for experiment 2 the results of Siskind's adapted model (SAM) do not seem to be in line with what Stevens et al. found, nor with Yu and Smith's expectations. Experiment 2 was devised to study the impact of the higher number of spurious correlations while decreasing the number of words to be learned. Yu and Smith argued that a statistical learner should have more difficulty learning in the 9-words/12-repetitions condition, even though the number of words would be smaller and the number of repetitions was the biggest. In fact, the study did not find a considerable difference between conditions. On the computational side, Stevens et al. report that PbV and Pursuit stayed both near 60% correct guesses – very near what was found for human participants. SAM, however, seems to suffer when the task has fewer repetitions per word and a bigger vocabulary, as seen in Table 7. Results are too strong for smaller vocabulary sets (around 100% of mappings learned) and doubling the number of words to be learned seems to almost halve this proportion. Notice that the forgetfulness value k was set to 8, in accordance to the findings just reported for experiment 1. These results seem to contradict what was found in experiment 1, suggesting the forgetting mechanism is not enough to make SAM a model of the behavior shown by human participants. Alternative mechanisms could of course be devised, although admittedly at the risk of being *ad hoc*.

I now turn to reporting the results of the simulation of the second experiment chosen by Stevens et al. (2017), which looks at the dependence between a correct or wrong referent selection and subsequent selections. Whereas in the experiment described in this section we do not look at per-trial performance, but rather at the overall lexicon acquired by the end of the experiment, next we ask the question of what choices SAM makes as it learns a vocabulary.

2.2.4 Simulation of Trueswell et al. (2013)

Yu and Smith's (2007) conclusions were challenged by Trueswell et al. (2013), as we saw in section 2.1.3, on the basis that an alternative cross-situational learning strategy could account for their results. This strategy, known as PbV, randomly picks only one referent to be the current meaning hypothesis but never becomes too attached to it: as soon as contrary evidence is found, the learner randomly chooses another hypothesis. On the other hand, if subsequent evidence corroborates the current hypothesis, then the learner has no reason to change it. Trueswell et al. designed an experiment which showed that there was a strong correlation between having randomly picked the correct referent previously and

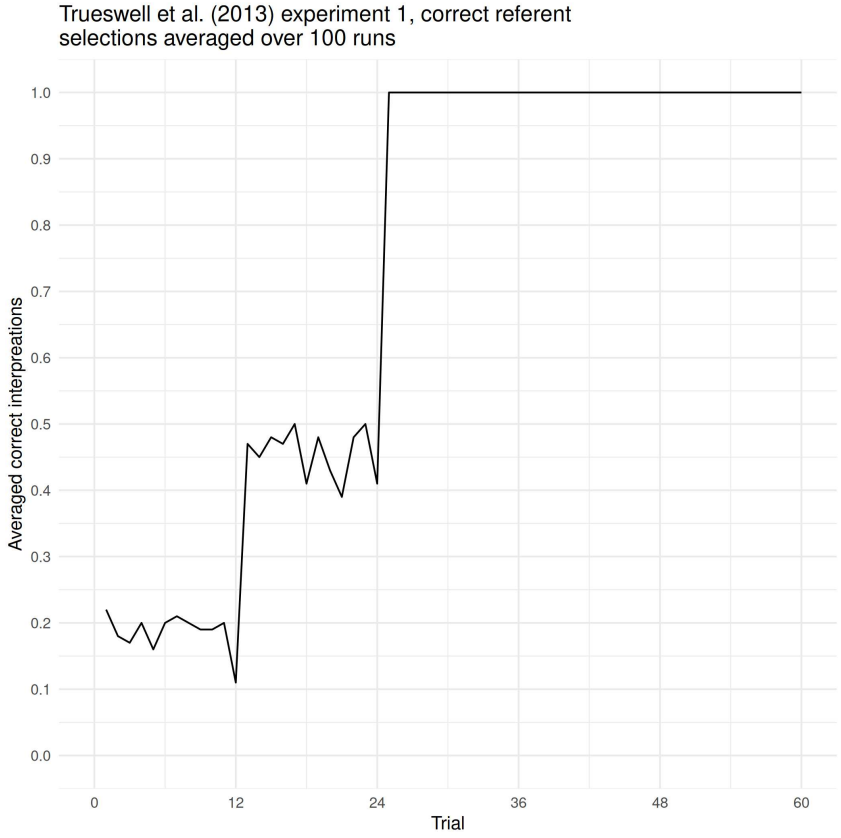
selecting the correct referent later. The authors argue this result is uniquely compatible with a PbV-like approach, since global approaches have multiple mapping hypotheses under consideration which means learners should be above chance when tasked with selecting the correct referent for a word already seen.

In the experiment, participants have to learn the referents for 12 words. Each word is shown 5 times and at each trial, the participant is asked to select the referent she thinks is the correct one. There are, in total, 60 trials, each composed of a word and 5 possible referents presented on the screen. Words always cooccur with the correct referent and never with other referents more than twice. That is, correct mappings are always present, but spurious associations are capped at 40%. At the beginning of the experiment, presentation order is randomized, constituting the presentation block. This block is then repeated in the same order four more times. The interval between the first, second etc. presentations of the same word is thus kept constant, with 11 intervening trials before it is seen again. This will be vital to understand the results below.

Implementation of the experiment strictly followed specifications in the paper. SAM was run 100 times over the simulated experiment and then answers were averaged. At each trial, the model first ran the heuristics, updating its lexical knowledge, then chose a referent according to the following logic. If a word's necessary meaning has been found, then choose from $N(w)$; if not, randomly choose from the possible meanings in $P(w)$. This is an extension of Siskind's original model, since his did not have to face a referent selection task. A history of referent selection as a function of trial can be seen in Figure 10. The behavior of the model is quite clear. The first time a word enters the model's lexicon, since there is zero information about that word, choice is completely random and stays at chance level, that is, the model has a 20% chance of choosing correctly. Then, after the 12th trial, the model has seen every word and is now going to encounter them in the same order. At this moment learning starts, since the heuristics are able to make inferences based on previous knowledge. Performance jumps to about 45% correct answers, which means the model has consistently narrowed word meaning down to about two possibilities. A third brisk spike in correct selection happens at the 25th trial. Starting from there, the model only makes correct selections.

Such results are at odds with what Trueswell et al. (2013) found in their real-life experiments. Figure 11 shows the proportion between previous correct or wrong answers and subsequent correct answers. Although the model does seem to have worse selection performance after having previously selected an incorrect referent, the results are nowhere near what was observed for human participants, who are at chance after having chosen incorrectly. Compare this to Figure 12, the results found by Stevens et al. (2017). Accuracy for SAM is as high as the highest among global models, which goes directly against the benchmark set by human behavior. In this task, SAM falls in line with the other global models.

Figure 10 – History of proportional correct selections of SAM in Trueswell et al. (2013), experiment 1. There are two well-defined spikes in performance: one after the 12th and another after the 24th trial.

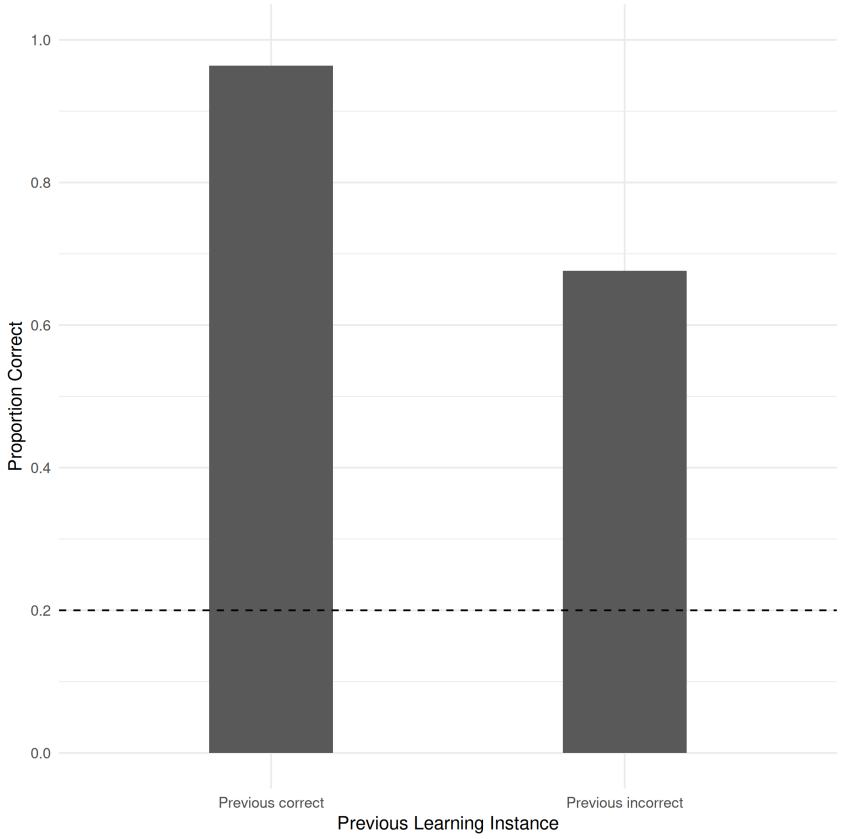


Two obvious features of SAM might be responsible for these implausible results. First, the simple selection mechanism adopted here might not model the conjectures a human makes when selecting referents. It might be the case that another mechanism, perhaps one that waits until words have converged before blindly choosing from $N(w)$, could yield a different behavior. That seems unlikely, however, given that words converge quite quickly under this experiment. A second, connected problem is that the pruning mechanism was not turned on at all. In fact, the forgetting mechanism described above – forget non-converged words after k trials – mortally wounds SAM, even when k is set to the best value found of 8. That is because it never has a chance of converging any words, since it takes 11 trials to observe a given word again. What this reveals is that this arguably artificial order of word presentation impacts SAM fundamentally. The heuristics work only under the assumption that utterances are, at the same time, not ordered nor one-word long. This is not a completely naive assumption, since it is known that the distribution of words in speech follows a power law (ZIPF, 1949)[10].

In fact, the case could be made that Trueswell et al.'s presentation order is too unrealistic to model real-life lexical acquisition, given that different distribution frequencies are a

[10]I tested the effect of randomizing trial order and although word convergence looked a little more natural and almost linear, referent selection as a function of previous selection was not impacted.

Figure 11 – Referent selection as a function of previous choice. SAM has much better performance than the human learner, who is at chance level of choosing the correct referent after having chosen the incorrect one.



necessary feature for incremental word learning. The fact few words appear much more frequently than the rest constitutes a stepping stone for learning. Recurring nouns could furnish the child with plenty of opportunities for their acquisition. Learning subsequent words gets easier as the child's lexicon grows, since known words provide clues in all directions: they can help the child map their language's syntax, help narrow down inferences about speakers' intentions, and provide more semantic context. It could be the case that while PbV captures something essential about learning words in experimental settings, it would fail to embrace the full breadth of realistic lexical acquisition. It remains a fact, still, that under these less than realistic experimental conditions, human participants solve the task in a particular way, captured by local models but so far incompatible with global approaches.

2.2.5 Why Koehne, Trueswell, and Gleitman (2013) cannot be simulated

In order to set Pursuit apart from PbV, Stevens et al. (2017) run a simulation of experiment 1 from Koehne, Trueswell, and Gleitman (2013). The main point of the experiment was to determine if participants would remember prior hypotheses. In order to test this, two referents are assigned to each word. One, called HPR (Hundred Percent Referent),

Figure 12 – (A) Performance of human participants; (B) PbV; (C) PbV without memory constraints; (D) Pursuit; (E) Modified Fazly et al.; (F) Original Fazly et al. Compare this to Figure 11. Extracted from Stevens et al. (2017), p. 26.

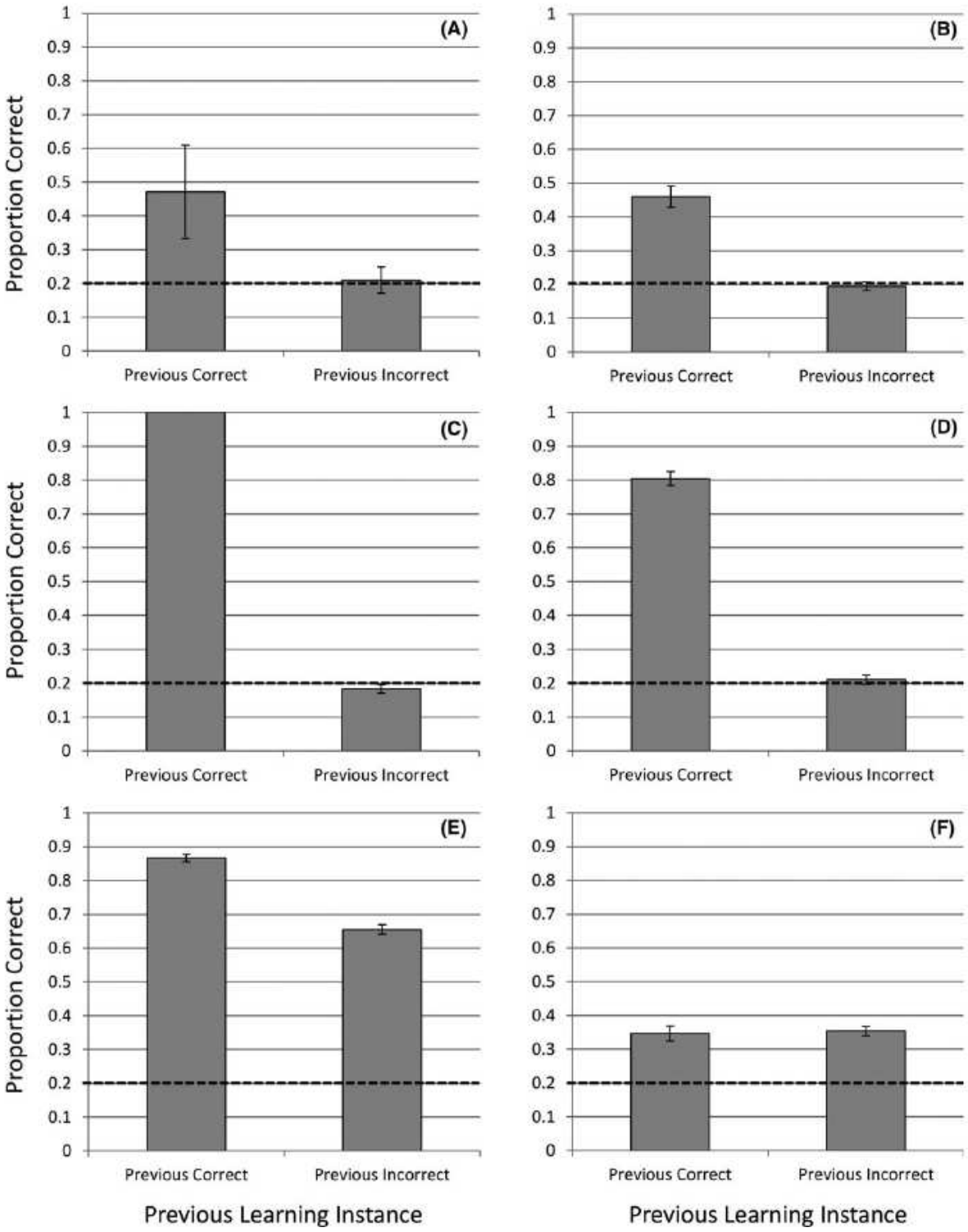
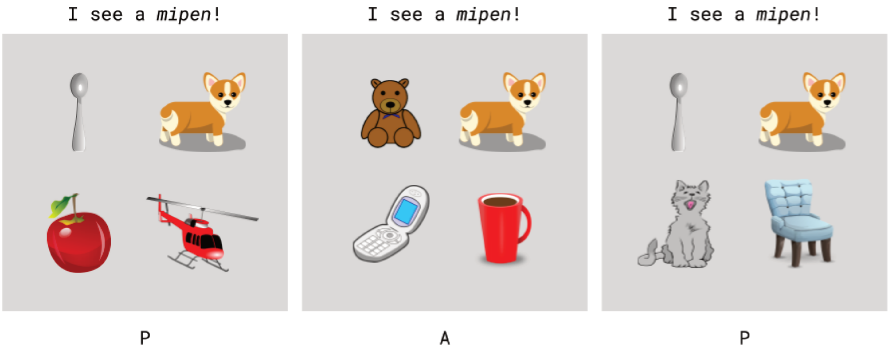


Figure 13 – In this example, "mipen" has as its Hundred Percent Referent (HPR) the corgi and as its Fifty Percent Referent (FPR), a spoon. In A(bsent) conditions, the spoon is missing. These would be the first three screens for "mipen" in the PAPAPA condition. Unlike in this example, 1–8 other words would be shown between presentations of "mipen."



occurs 100% of the times with its word whereas the FPR (Fifty Percent Referent) only cooccurs 50% of the time. All other referents cooccur only once with each word. The order in which the FPR is present (P) or absent (A) is controlled, constituting four conditions: PPPAAA, AAAPPP, APAPAP, PAPAPA (see Figure 13). At the end of experiment, the participant has to choose between 8 possible referents, but crucially, only the FPR is present. Under a strict PbV approach, only in P-ending conditions should participants choose the FPR above chance. This is because strict PbV predicts that only one referent is considered at each time. If the FPR was absent in the last presentation of the word, then it should not be chosen above chance, since it would be unavailable in the learning instance and thus erased from memory. However, the authors find that it *is* chosen even in A-ending conditions as long as it had been selected in at least one previous encounter with the word. That finding is compatible with Pursuit but not PbV.

Experimental design consisted of tasking participants with learning 16 novel words presented with a spoken utterance, such as "I see a mipen!" The experiment is divided into two blocks: first 8 words are presented, then the selection test is done, and this is repeated with 8 new words. Each word is shown 6 times and belongs to one of the four conditions. Further, all participants experience all conditions. Also, the presentation of a word is not back-to-back, but rather intermixed with a random number of 1–8 trials in which other words are shown. This design not only allows for comparing participant performance in different presentation orders but also their learning path, since at each trial they are instructed to click on the referent they think is correct.

At a first glance, it seems that in order to run this experiment, SAM would have to be expanded to include learning under conditions of homonymy. Let us explore this idea in some detail. However, as we shall see, not even that could solve the problem.

The fundamental challenge arises when trying to adopt Siskind's original strategy and assumptions. So far, SAM has implemented what Siskind calls the "noise-free monosemous

case" (SISKIND, 1996, p. 54). The heuristics described above are able to learn word meanings under the assumption that there is no noise or homonymy in the input data. However, the original model was expanded to deal with these two obstacles when some conditions are met, as we shall see. I focus on the effects homonyms have in the heuristics, however bear in mind that the expansions described below were also meant to help deal with noise.

Siskind points out (1996, p. 62) that in a homonym-free lexicon, the outcome of processing utterances and meanings follows certain patterns: the set of necessary meanings will always be a subset of the possible meanings; also, $P(w)$ will never be empty; and Rule 1, responsible for reducing referential uncertainty, should keep at least one utterance meaning around. Should any of these patterns be broken, an inconsistency has emerged. This may mean, among other things, that a homonymous word is present in the utterance-meanings pair. To deal with inconsistencies, Siskind expands the lexicon beyond being a direct mapping of words to meanings. Words are mapped to senses, such that a single word may have many senses, each then associated with a meaning. When an inconsistency is detected, as few new word senses as possible are proposed, all senses are updated as possible and processing moves on to the next utterance. However, notice that this implies that a new word sense will only be proposed if all currently known word senses are absent in the utterance, at least in the case of single-word utterances. Simply put, the learner only proposes a new word sense for w if the current usage of w is inconsistent with all known senses of w. A rupture is a necessary condition for spotting homonyms under Siskind's approach.

To make this clearer, consider how an inconsistency would never arise under Koehne et al's (2013) experiment. Say the model is processing an utterance containing a new word, *vonk*, which maps to the HPR DOG and to the FPR CUP. Both referents are present, since this is a P-condition utterance:

(5)

Utterance	*vonk*
Scene	{DOG, CUP, PEN, DOLL}

At this point, the model updates $P(vonk)$ to the set {DOG, CUP, PEN, DOLL}. The problem will arise right in the next utterance: only DOG will be present, this being an A-condition utterance:

(6)

Utterance	*vonk*
Scene	{DOG, ~~BOOK~~, ~~CAT~~, ~~TOY~~}

The heuristics exclude any possible "utterance" meanings not in $P(vonk)$, then remove all meanings from that set except for DOG, and finally add DOG to the necessary meanings of *vonk*. The word has converged and will not be updated anymore. Thus, the next utterance, containing the FPR CUP will be completely innocuous as a learning instance. It is a property of Siskind's model that unless an inconsistency arises, no new word senses

will be proposed. Since the HPR is always available, there are no opportunities to learn homonyms.

Recall that Koehne et al.'s main finding is that in the selection test phase, when the HPR is always absent, participants are able to choose the less frequent FPR above chance even in A-ending conditions. However, the example above showed that SAM does not explain this real-life experimental observation. Of course, in a more naturalistic setting, it is reasonable to assume that the data would eventually lead to a situation of inconsistency where *vonk* is paired only with CUP and not DOG. In this case, Siskind's model could posit a new sense of *vonk*. Still, it remains a fact that it cannot straightforwardly deal with the experimental task just laid out.

How then can the local and global models studied by Stevens et al. (2017) learn mappings in this task, if at all? PbV learns by *selecting* a referent. If the HPR is selected at any instance, it will be confirmed and thus learned. However, the FPR can only be learned in a P-ending condition like AAAPPP or APAPAP, as discussed above. Pursuit, which also locally selects a referent, has the best shot of learning the FPR in the PPPAAA condition. That is because probability of selecting *and then confirming* the FPR will be at its highest in this condition. Finally, global models do not select a referent but rather track absolute cooccurrences. The mechanism implemented by Fazly, Alishahi, and Stevenson (2010), the global model chosen in that study, spreads the probability distribution among the HPR and the FPR (and all other bogus referents which happened to cooccur). Then, when tasked with selecting the FPR, the model has around 50% chance of choosing it correctly (STEVENS et al., 2017, p. 29). It, however, falls short of replicating the effect that order has on human participants. AAAPPP is the worst condition for humans, whereas PPPAAA is the best, however Fazly et al. is equally good across conditions.

Unlike the models above, the full homonymy-aware Siskind model has no sensibility to the presence of two possible referents. The fact that it is a *deterministic* system means it never considers a co-occurring word to be a possible referent once it has been removed from $P(w)$. This is only logical, since secondary referents can occur in the absence of the most common referent, giving the model an opportunity to pose a new word sense. The problem seems to stem from the fact that Koehne et al.'s experiment captures sense *competition* rather than homonymy, as they hint at (pg. 810). Models which end up building a distribution of probabilities are able to capture human behavior in a very controlled, experimental task. Still, this task does not seem to be very representative of actual word learning. Even though a child may be confused by a highly co-occurring referent, the natural distribution of the world (or other cognitive/linguistic clues) might give her enough information to rule it out eventually.

2.3 Related issues and related models

2.3.1 Exploring the role of other cognitive and linguistic constraints

So far, we have reviewed and explored issues related to a putative cross-situational mechanism of lexical acquisition. Whatever its inner workings – whether it registers cooccurrences, building global associations over time, or exploits locally available hypotheses, checking them later – it is possible that such a mechanism is aided by information from other sources. In this section, extended mechanisms, usually built upon previous work, are reviewed. They investigate the importance of other cognitive processes such as shared attention and intention reading, as well as linguistic constraints, such as word stress and semantic categories. They constitute only a sample of the research that has gone into the interaction of a core learning mechanism and other data channels meant to help enrich this survey of the state of the art of word learning models as it currently stands.

The first model considered here is Yu and Ballard (2007, see Table 8), who present a study unifying cross-situational learning to information from social cues, such as deictic body movement and prosodic variations. This information is added to the input data, such that certain referents are marked as the focus of attention and receive higher weight at the same time as certain words, being prosodically stressed, are also more salient. Interestingly, the results showed that prosody bears less information than joint attention. Before discussing their conclusion, let us understand their proposal.

Two ways of learning words are discussed. First, following Bloom (1997), the fact that children are cognitively endowed with conceptual biases, intention reading and syntactic knowledge might be enough to explain word learning. Under this theoretical approach, children's hypotheses on word meaning would be steered by their capacity to read speakers' intentions, by observing the objects being attended and by constraints inherent to language. Conversely, cross-situational learning has also been proposed as a word-learning mechanism, as I have discussed extensively. In and of itself, this purely associative mechanism could not learn meanings for an unseen referent. That is, a child may hear "open" but very rarely will this verb occur with the actual event of opening something. However, if one assumes some intention reading on the part of the child, these meanings become available in the immediate context of hearing the verb[11]. The goal of Yu and Ballard's model is, then, to quantitatively characterize the interaction between a cross-situational mechanism and extra sources of information. More especially, eye gaze and prosody information are encoded in the input to their model.

Input data consisted of transcripts from the CHILDES corpus, manually paired with referents judged to be available in the scene. None of the target words, which referred

[11] Think back to Siskind's (1996) cross-situational model: it is capable of generating meaning hypotheses to represent the semantic content of utterances. Since these representations capture verb meaning in their core, it could be said that this model assumes a series of cognitive processes have already taken place in order for the meaning hypotheses to be generated.

Table 8 – Summary of Yu and Ballard (2007). This model adds a new layer of data to investigate the contribution given by prosodic and eye gaze information. Input data is annotated so that focused nouns (by prosody or eye gaze) receive more weight.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit	Yurovsky and Frank (2015)	Yu and Ballard (2007)
Type	deterministic, online, global	probabilistic, batch, global	probabilistic, online, global	probabilistic, online, local	probabilistic, batch, local, global	probabilistic, batch, global
Models	ref. uncertainty, noise, homonymy, lex. categories	ref. uncertainty, homonymy, synonymy	ref. uncertainty, noise, synonymy	ref. uncertainty, noise	ref. uncertainty, memory	ref. uncertainty, synonymy, cognition
Evaluation	behavior	behavior	behavior	behavior, experiment	experiment	performance
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli	simulated experimental stimuli	corpus of child-directed speech
Semantics	compositional (Jackendovian)	atomic	atomic	atomic	atomic	atomic

Table 9 – Results found by Yu and Ballard (2007). Precision and recall are maximal when all three sources of information – statistical regularities, prosody and joint attention – are used together, however the effect of considering joint attention is much stronger than the information encoded in prosody.

Model	Precision	Recall
Statistical	75%	58%
Statistical + prosody	78%	58%
Statistical + joint attention	80%	73%
Statistical + all social cues	83%	77%

to the toys played with during the recorded session, were among the 15 most frequent words; in the same direction, referential uncertainty is high: only about 2.5% of all possible word-referent mappings were relevant, since the quantity of words spoken far outnumbered the targets. The social cues being studied are included in the input data by changing the relative weights of individual referents and words. Thus, if referents are the object of joint attention, they receive more weight. Similarly, the same is done to words which are more salient in speech. Also, a `NON` referent is added to each scene so that words used non-referentially – such as function words – can be safely identified. The assumption is that children somehow know that some words may not refer to things in the extra-linguistic context. The data is batch processed by a model based on machine translation algorithms.

This choice of algorithm is not unique among lexical acquisition modeling, as I have discussed elsewhere, and has to do with the fact that both input streams – words and referents – can be thought of as two "languages" which can be mapped onto each other. In order to establish a baseline for later comparison, the authors implement a purely statistical model, without the weighted information from social cues. Although this version is able to learn some of the target words, it also seems to be affected by spurious correlations in the data. The blame is laid on the small data set used and the high degree of referential uncertainty. Next, the model is integrated with joint attention and prosodic cues. Four experiments are run on the same data to determine the role of each separate channel of information: the purely statistical model discussed above, statistics and prosody, statistics and joint attention and statistics plus all social cues. Results are shown in Table 9. Adding social cues not only strengthens the association between correct words-referent pairs, but also weakens wrong ones.

As previewed before, the performance of the model shows an asymmetrical relationship between information contributed by prosody information and joint attention. Although the last, most integrative experiment had the best results, it was only marginally better than the statistical plus joint attention model, showing that in these simulations, prosody did not play a major role. The authors argue that this might be because prosody is used in an ambiguous way when talking to the child, since it is sometimes used to call their attention and some others to actually highlight words in the utterance. Eye gaze, on the other hand, is a more consistent marker of relevant referents. An interesting suggestion is

that if prosodic weights were only considered when joint attention was engaged, the model could yield even better results. This is due to the fact that prosody will be maximally informative when speakers and children are attending to the same objects. This speculation has not been explored as far as could be found in the literature.

Frank, Goodman, and Tenenbaum (2009, see Table 10) present a Bayesian model to explore yet another aspect of cognition: the inferences about (non-)referential use of nouns a child is able to make. Their proposal processes a corpus of utterances extracted from two video files from the Rollins corpus, part of CHILDES. Each utterance was manually paired with objects judged to be available in the extra-linguistic context. Central to the model is that it considers not only these cooccurrences in order to find noun-object mappings, but also a probabilistic measure which reflects the speakers' referential intention when employing a word in an utterance. Word use may be referential, in which case the utterance-objects pair is considered as a valid learning instance, or non-referential, in which case the utterance is discarded. This models children's ability to gauge speakers' intention, allowing noisy pairings to be discarded. The authors argue adding this in yields better results than other approaches.

In fact, in a comparison to other selected models, Frank et al.'s does indeed produce a more accurate lexicon given the task of acquiring mappings from an annotated corpus. In order to carry out this comparison, three models are built implementing different probabilistic mechanisms. The translation model on which Yu and Ballard (2007) is based is also used. Their chosen measurement, the F-score, a harmonic mean of precision and recall, is highest for the authors' model. Precision is the proportion of pairings learned which are correct, whereas recall is the proportion of correct pairs compared to the gold-standard lexicon. This means the model is both more accurate as well as has a better coverage of the target lexicon. The fact that the model is able to determine if a word was employed referentially or non-referentially seems to allow it to both keep words that do not map to objects out of the final lexicon as well as discard uninformative utterance-objects pairs, minimizing noise.

The authors recognize that while their model constitutes an argument for cross-situational learning, it is not alone in doing so, since other work has shown similar results. However, they also point out that it goes further, exhibiting a series of behaviors similar to children acquiring words. For example, it shows a preference for $1 \rightarrow 1$ mappings, encoding a sort of mutual exclusivity (MARKMAN, 1990) constraint[12]; fast learning (CAREY, 1978)

[12] Stevens et al. (2017), p. 13, argue that since Frank et al.'s model penalizes $n \rightarrow 1$ mappings and has a preference for smaller lexicons, it ends up implementing the mutual exclusivity constraint implicitly rather than that being a natural emerging quality.

Table 10 – Summary of Frank, Goodman, and Tenenbaum (2009), a Bayesian model of word learning that explores an aspect of children's cognitive abilities: their inferences about spearkers' referential intentions.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit	Yurovsky and Frank (2015)	Yu and Ballard (2007)	Frank et al. (2009)
Type	deterministic online global	probabilistic batch global	probabilistic online global	probabilistic online local	probabilistic batch local global	probabilistic batch global	probabilistic batch global
Models	ref. uncertainty noise homonymy lex. categories	ref. uncertainty homonymy synonymy	ref. uncertainty noise synonymy	ref. uncertainty noise	ref. uncertainty memory	ref. uncertainty synonymy cognition	ref. uncertainty cognition
Evaluation	behavior	behavior	behavior	behavior experiment	experiment	performance	behavior experiment performance
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli	simulated experimental stimuli	corpus of child-directed speech	corpus of child-directed speech
Semantics	compositional (Jackendovian)	atomic	atomic	atomic	atomic	atomic	atomic

is observed; object individuation, that is, the observation that children expect the number of linguistic labels to be in line with the quantity of objects in the extra-linguistic context (XU, 2002), is also captured; and intention reading, the correlation children assume between naming and gesturing/looking (BALDWIN, 1993), is reproduced.

As pointed out in the discussion of probabilistic models in section 2.1.2, depending on batch processing of input data leaves an approach open to criticism. Since Frank et al.'s strategy tries to find the smallest lexicon given some constraints, probability weights and the data, it needs to consider all relations that would be otherwise accessible only to an external observer. Thus, a mechanism such as this cannot be a complete explanation of lexical acquisition, for it does not capture online development of a lexicon, nor can it be easily compared to other models via simulations of psycholinguistic tasks.

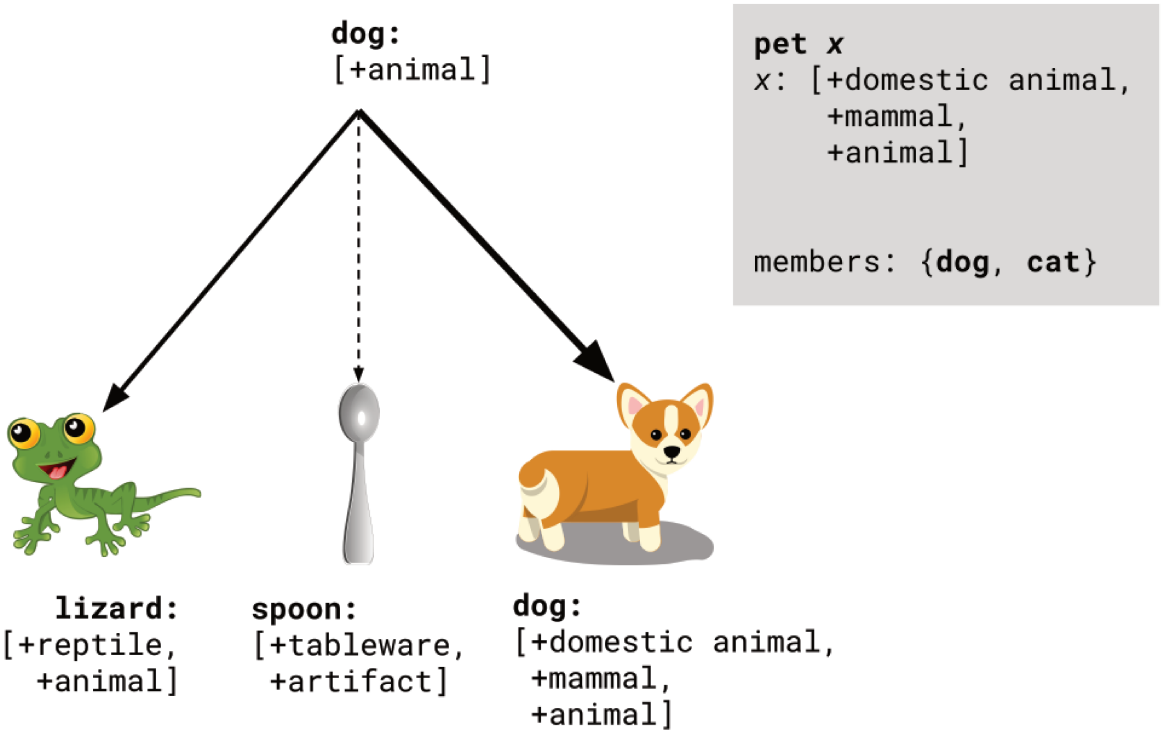
Shifting from attentional constraints to the ones provided by the sentential context, the role of semantic category membership is the focus of the model presented by Alishahi et al. (2012, see Table 11). This is an extension of Fazly, Alishahi, and Stevenson (2010), adding interpretation clues contributed by the verb. For example, verbs may impose restrictions on which semantic classes their complements can be associated with. The goal is to capture the intuition that when the child hears "daddy is wearing moccasins," the verb *wear* imposes certain semantic restrictions onto its object, such that the child can assume *moccasins* are some kind of clothing or footwear even when she does not know the meaning of that word. In the model, nouns are members of categories encoding the sentential contexts in which they have appeared as verb complements. Categories themselves have semantic content derived from the meanings of their member words. Thus, when solving the ambiguity of word-referent alignment, the model employs two sources of information: known (partial) word meanings and how much the sentential context licenses an alignment. Notice that the same word might belong to more than one category: *apple*, for instance, occurs in "cut the apple" as well as "eat the apple" and as a consequence, *apple* and *papaya* belong to the same categories (members of both "objects of eat" and "objects of cut").

Unlike Fazly et al.'s model, however, semantic representation is not atomic here, but a collection of features extracted from that word's entry in the WordNet. Referents in the scene are represented as a set of features and the alignment between word and referent is calculated via set similarity. The logic behind this is: the higher the number of semantic features a word and a referent share, the more likely their alignment is; but this can be counterbalanced (or doubled down) by the semantic features composing the category to which the word is associated (see Figure 14). At each new utterance, the semantic features f which make up the meaning of each word are updated and this change propagates up to the relevant contextual categories.

Table 11 – Summary of Alishahi et al. (2012). Sentence-level (verb restrictions) are included in the model and tested in a simulation of an experimental task. Differently from the other models so far, the semantics of each word is a matrix of features extracted from the WordNet.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit	Yurovsky and Frank (2015)	Yu and Ballard (2007)	Frank et al. (2009)	Alishahi et al. (2012)
Type	deterministic online global	probabilistic batch global	probabilistic online global	probabilistic online local	probabilistic batch local global	probabilistic batch global	probabilistic batch global	probabilistic online global
Models	ref. uncertainty noise homonymy lex. categories	ref. uncertainty homonymy synonymy	ref. uncertainty noise synonymy	ref. uncertainty noise	ref. uncertainty memory	ref. uncertainty synonymy cognition	ref. uncertainty cognition	ref. uncertainty verb restrictions
Evaluation	behavior	behavior	behavior	behavior experiment	experiment	performance	behavior experiment performance	experiment
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli	simulated experimental stimuli	corpus of child-directed speech	corpus of child-directed speech	simulated experimental stimuli
Semantics	compositional (Jackendovian)	atomic	atomic	atomic	atomic	atomic	atomic	features

Figure 14 – In this example, the model is trying to decide what the best alignment of the partially-known word *dog* is given the three referents available in the context. Set similarity would yield two equally likely alignments: LIZARD and DOG, since they all share the feature [+animal]. However, the system also knows *dog* has appeared as complement of the verb *pet* as in "pet the dog." Since it has accumulated features about complements of *pet* (via *cat*), it is able to make the correct alignment to DOG.



Simulations of Koehne and Crocker's (2010 and 2011) psycholinguistic experiments were carried out in order to evaluate model plausibility. The original goal of these experimental tasks was to analyze the role of sentential-level information. This was accomplished in three steps:

1. Participants are first familiarized with a series of restricting and non-restricting verbs with clear correlates in their native languages; for example, *bermamema* meaning "eat" and *tambamema*, "take."
2. Next, participants see static scenes paired with SVO sentences built from the verbs just introduced and from two new nouns in the subject and object positions. The nouns in the subject position always mean either "man" or "woman," however a larger meaning variety is found in the object position.
3. Finally, participants go through a forced referent selection phase. They hear a target noun while a series of potential referents are presented. The task is to click on the referent considered to be the correct one and report their degree of confidence in the selection.

In the real-life tests, human participants are able to project the semantic restrictions imposed by their native-language verbs onto the newly learned ones. To simulate this, Alishahi et al.'s model is first trained to acquire these restrictions, represented by noun

membership to certain semantic categories as discussed above. Next, the noun learning phase is run, with the model receiving a scene consisting of a series of objects, including the correct referents as well as two distractors. After these learning trials, comes the referent selection task. Here, the probability distribution of choosing a referent r given a word w is taken to be the possible answers given by the model.

Results from two distinct experiments showed that sentence-level information carried considerable weight in word learning. In the first experiment (experiment 2 of Koehne and Crocker, 2010), the relation between referential uncertainty and information given off by restricting verbs was investigated. Importantly, these two sources of information – word-referent cooccurrences and verbal restrictions – were complementary, meaning that both were necessary for successful learning. This was achieved by controlling verb type and referential uncertainty in three different conditions. In the No-RU (Referential Uncertainty) condition, nouns were always preceded by a restricting verb and only one referent matched the verb restriction. In the Low-RU condition, two referents were compatible. In the last condition, High-RU, verbs were non-restrictive and so all four referents were plausible. In the first two conditions, complementary sentence-level information was available, while in the last condition, only a cross-situational strategy could find the correct pairings. Further, the Low-RU condition forced the learner to use both strategies if learning was to be successful. Participants were tasked with learning 12 novel nouns after being familiarized with four new verbs, two restricting and two non-restricting. Human participants did better in conditions with restricting verbs than with non-restricting verbs, showing that sentence-level information was used in learning words. Alishahi et al.'s model falls in line with these expectations and similar to the results with humans, it fares better in the No-RU condition than in the Low-RU condition, meaning that cross-situational learning does not fully compensate for the ambiguity left by having two compatible referents.

An extra study was carried out to investigate how incomplete semantic categories could, if at all, guide word learning. In the initial experiment, the model builds category information by processing 5,000 training input items. The resulting categories serve to model the interpretations licensed by restricting verbs. Then, these categories undergo a manual cleaning process to ensure no spurious meanings are associated with them. Verb restrictions are thus highly informative, capturing adult intuitions. However, the authors ask themselves what the effect of categories generated with less information would be. The model was trained with ten times less items and the categories formed were not cleaned. Learning performance was significantly degraded, which might capture what the authors call an *effect of age*, that is, the fact that in earlier stages of acquisition knowledge of which complements are licensed by which verbs is still nascent and cannot be very informative.

In the second experiment (experiment 2 of Koehne and Crocker, 2011), the effect of sentence-level information vs. cross-situational learning when they provide redundant information was studied. For that end, each of the target nouns had two potential meanings: a High Frequency Referent (HFR) which was paired with the noun 83% of all times and a

Low Frequency Referent (LFR), which cooccurred with it in only 50% of the cases. All other objects cooccurred only 17% of the time. Again, participants first learned new verbs and then were tasked with learning the meaning of 16 novel nouns. The main finding is that in forced selection tests in the absence of the HFR, participants chose both the LFR as well as a distractor object belonging to the same category as the HFR. This indicates that verb restrictions compete with cross-situational learning, thus being another, parallel source of information when acquiring a lexicon. Again, simulations with the model replicated these results. In conditions where the verb was non-restrictive, in the absence of the high frequency referent, the LFR was favored against the other distractors. However, in conditions with restricting verbs, the LFR shares probability with another referent if it belongs to the same category as the HFR, indicating there's an interaction between both sources of information considered here.

2.3.2 Possible challenges to Siskind's (1996) approach

Faria (2015, see Table 12), needing to explain lexical acquisition in the broader context of trying to model language learning as a whole, presents an adaptation of the core ideas from Siskind's (1996) model. As a consequence of his particular task, Faria's implementation faces different challenges and thus has to extend and change the system of rules originally proposed. The input data is more complex, featuring different sentence types, such as declaratives, interrogatives, imperatives and sentence fragments. Besides being syntactically richer, the input also had a higher lexical sparsity due to the choice of using words from Brazilian Portuguese (BP). Note that in Siskind's model, words were abstract enumerated items ($w_1, w_2, \ldots$). Thus, each verb or noun occur only in one form. BP has richer morphology, which makes it rarer for a specific word form to appear. Polyssemy and a more complex conceptual symbol inventory were also important factors, as discussed below. The inclusion of more conceptual symbol types was motivated by the fact that Faria also simulated the learning of function words, such as determiners. In these more realistic conditions, model performance was worse when learning from English and Brazilian Portuguese corpora. This seems to suggest Siskind's proposal needs reviewing to be able to process more sparse, polyssemic input data. However, there are other sources of trouble which might account for the poorer performance.

One important departure from Siskind's study was that the data did not have neither referential uncertainty nor noise. This places this model in a different category than the ones reviewed thus far, since it does not try to tackle the mapping problem as described in this chapter. Still, Faria's work raises some interesting challenges to be investigated opportunely. To preview the discussion below, new problems might arise for the word learner when language is considered in greater syntactic, semantic and even lexical detail, aspects largely ignored by the models reviewed so far.

Table 12 – Summary of Faria (2015). This model concludes our table and, differently from all other proposals, does not attempt to solve referential uncertainty. Instead, it adds polyssemy and lexical categories as a result of its semantically more complex input data.

	Siskind (1996)	Yu (2008)	Fazly et al. (2010)	Pursuit	Yurovsky and Frank (2015)	Yu and Ballard (2007)	Frank et al. (2009)	Alishahi et al. (2012)	Faria (2015)
Type	deterministic	probabilistic	probabilistic	probabilistic	probabilistic	probabilistic	probabilistic	probabilistic	deterministic
	online	batch	online	local	batch	batch	batch	online	online
	global	global	global	online	local	global	global	global	global
				global	global				
Models	ref. uncertainty	ref. uncertainty	ref. uncertainty	ref. uncertainty	ref. uncertainty	ref. uncertainty	ref. uncertainty	ref. uncertainty	homonymy
	noise	homonymy	noise	noise	memory	synonymy	cognition	verb restrictions	polyssemy
	homonymy	synonymy	synonymy			cognition			lex. categories
	lex. categories								
Evaluation	behavior	behavior	behavior	behavior	experiment	performance	behavior	experiment	behavior
				experiment			experiment		
							performance		
Input	synthetic corpus of utterance-meaning pairs	lab-built corpus of utterance-referents pairs	corpus of child-directed speech	corpus of child-directed speech; simulated experimental stimuli	simulated experimental stimuli	corpus of child-directed speech	corpus of child-directed speech	simulated experimental stimuli	synthetic corpus of utterance-meaning pairs
Semantics	compositional (Jackendovian)	atomic	atomic	atomic	atomic	atomic	atomic	features	features

Besides having more utterance types, Faria's input data also brings more word classes, such as verbs and function words. Utterance types are represented by conceptual symbols like DECL, for example, which heads declarative sentences. Verbs contribute to the polyssemy alluded to above, since there are, for example, inchoative verbs; also, they come in various inflected forms. The same word symbol, say the verb *break*, can be used in two different ways: "the vase broke" or "the cat broke the vase." Notice that this is different from strict homonymy, since both uses of *break* share all conceptual symbols except for the one responsible for encoding causativity. Another source of polyssemy is not lexical but rather syntactic. Consider the BP sentence "as girafas comeram as folhas das árvores" (*the giraffes ate the leaves from the trees*). The DP "as girafas" has two instances of the plural morpheme *-s*. The semantic representation of this sentence, however, needs only encode the fact that there were more than one giraffe. Similarly, since in some dialects of Brazilian Portuguese the verb agrees with the subject in number, the morpheme *-am* is also contributing with plural. So, although there are three plural words, only one conceptual symbol is needed to convey the plural meaning (something like PLURAL(*giraffe*)). Recall that in Siskind's original proposal, it is paramount that each meaning contributed by a word be found in the semantic representation. Yet, here we can see that assumption falls apart as soon as we consider wider linguistic phenomena.

This model's lexicon is organized in much the same way as Siskind's: words may have multiple senses and a sense has three tables in which possible and necessary conceptual symbols, as well as the resulting conceptual expressions, are stored: $P(w), N(w),$ and $D(w)$. Nevertheless, the original heuristics are somewhat adapted. First, since there is zero referential uncertainty, Rule 1, which served to reduce uncertainty, is adapted to ignore inconsistent Possible Sense Assignments (PSAs). These constitute a mechanism to select word senses. Since words in an utterance may have multiple senses, the system generates the cross-product of all meaning possibilities from all senses; PSAs are these meanings. Rule 1 compares them with the sole paired utterance meaning and discards a PSA if it does not contain a symbol in the $P(w)$ for some word w or if not all $N(w)$ contribute to that utterance meaning. A second alteration reduces the number of rules from six to five. Faria's Rule 5 substitutes the original Rules 5 and 6. These were meant to build conceptual expressions once a word had converged on its conceptual symbols; however, in the current model these expressions are built as soon as possible from all meanings in $P(w)$ and purged as possible meanings are deleted.

Performance was compared in four different synthetic corpora: a small "head-final" corpus, a larger English corpus and two even larger BP corpora, BP I and BP II. Mean Length of Utterance in words (MLUW) varied from 5–6 across the corpora. However, the number of target words had greater variation, with head-final having around 56 words, the English corpus with 91, BP I with 133, and BP II with 464. The size of the corpora in utterances also varied: 2,071, 40,083, 100 thousand, and 100 thousand words, respectively. Both BP corpora were larger in order to guarantee enough word expositions to the model.

Results showed that performance was good for the first two corpora, but worse for the latter, more complex corpora. When processing the head-final corpus, the model was able to learn up to 96% of all words (95% was the convergence threshold proposed by Siskind) and showed no false positives, that is, partially correct meanings. The English corpus showed slightly worse performance, registering some false positives, although the threshold of 95% convergence was also exceeded. BP I is where the first big drop in performance can be seen. The model could only converge for 53% of words, even though there was only one verb stem per verb class. For the BP II corpus, where the number of verb stems per class was brought up to 8, performance was even worse, with no more than 39% convergence.

Since performance with the head-final and English corpora was similar to Siskind's, the author argues that the greater lexical sparsity of BP data may account for such discrepancy. Many more words in these corpora had higher frequency than in the BP corpora, probably due to the higher number of inflected forms the same stem can take. Since the learner dealt with morphologically unanalyzed forms, word entries were kept separated. Furthermore, since expositions to the same form were further apart, the garbage collector discarded non-converged meanings before they ever had a chance of converging. Alterations to this forgetting mechanism did not produce improvements in performance.

There are also two other issues that might explain the relative worse performance for the BP corpora. Code efficiency could be worse compared to Siskind's original model, rendering the processing of larger datasets impossible. More important still is that fact that the distribution of lexical items was not controlled[13]. Natural occurring speech follows a distribution known as Zipf's Law (ZIPF, 1949), which states that the frequency of a word type is inversely proportional to its rank. That means the first most common word happens twice as much as the second most frequent word, three times as often as the third most frequent word and so on. This is a property of language distribution which might be exploited by word learners and conceivably could have helped Faria's model. The most frequent words could be acquired first, providing a bedrock for learning less frequent ones. This possibility remains unexplored so far.

Nevertheless, the issues posed by adding richer morphology and polyssemy to the input are conceptual challenges to Siskind's approach and, inasmuch as that model is representative of the other global models reviewed above, to the whole area of modeling lexical acquisition. It may be that a core mechanism of word learning works independently from more linguistically-oriented cognitive processes. In this way, word learning can be studied isolated from other factors. If we focus single words and objects or the acquisition of concrete nouns, then we might think the core mechanism has been fairly modeled. However, children do not hear words in a vacuum. When taking that into consideration and trying to work from fuller utterance meanings, models must take into account interactions from different domains. I will return to this point in the wrap-up discussion in Chapter 4.

[13]This fact was noticed during personal communication with the author.

3 Lexical acquisition in practice and theory

We are now at a position where it is feasible to compare these computational models of word learning to what is known about children's lexical development and the theoretical necessities associated with the task. I begin with a short description of children's path learning words, what pitfalls and successes are expected and when. The models are then discussed under this empirical light. Next, an equally brief summation of the main ideas surrounding the theory of word learning is given. I focus on the problem of indeterminacy of reference, or referential uncertainty, and what solutions have been put forth. This leads us to a collation of these ideas and how they are actually materialized in the computational models.

3.1 Lexical development in children

At opposite ends of lexical acquisition are the newborn child, who knows a grand total of zero words, and high school children, with a vocabulary reaching some 60,000 words (BLOOM, 2000, p. 6). When beginning this process, children are quite modest learners. Different studies indicate children have acquired their first ten words by 13–15 months of age and that the fifty-word mark is reached at around 17–20 months (BARRETT, 1995). At 24 months old, children have acquired around 250 words[1], which reveals they become better learners as time goes on. These milestones, however, are taken from words produced by children and thus may be misleading.

An interesting fact observed time and time again in both diary and longitudinal studies is that word comprehension precedes production by up to four months (INGRAM, 1999, p. 140–143). Thus, a huge gap exists by the time children reach the first milestone of 10 words in production, being able to understand an average ranging from 60 to 110 words, figures varying depending on the study (BARRETT, 1995, p. 363). In fact, this distinction between the comprehension versus the production vocabulary reveals an intrinsic difficulty of trying to describe this process: there is no easy definition of what it means to "acquire a word." Take, for example, the fact that some productions cannot clearly be called "words" in the traditional meaning.

Among children's first productions sometimes are what are called idiosyncratic vocalizations (BARRETT, 1995, p. 364). Apart from having somewhat stable phonetics and a clear communicative intent, these vocalizations lack conventionality, that is, are not shared by the community of speakers and thus could not be classified as words having been acquired from the linguistic input. Nonetheless, the fact that these sounds are employed

[1]The range of variation among children for any of these figures is massive. For example, in one study children at 24 months had a vocabulary ranging from 41 to 668 words (BARRETT, 1995, p. 363).

in a consistent manner for seemingly communicative purposes may be an evidence young children are already making use of language.

Besides producing nonconventional words to express their feelings, children also produce conventional ones with idiosyncratic meanings. Early on, they may only use some words in very specific contexts. Barrett (1995) lists a few examples, such as only saying *duck* while hitting a toy duck on the edge of the bathtub or *car* only while looking cars passing by out the living room window (p. 364). Conversely, other early words do not have the same limitations. For example, *teddy* might be employed while looking, touching, pointing or playing with a teddy bear (p. 366). Although such word usage seems to approximate adult meaning, it might also be a case holophrastic usage (INGRAM, 1999, p. 231), in which a word is employed to mean a whole sentence. Thus the utterance "teddy" in the context of pointing to the toy might be equivalent in meaning to "give me the teddy." On its surface, holophrastic meanings are still far from the target meaning that words will eventually take on. However, other explanations could be proposed, such as that children are trying to express a need with limited lexical, syntactic, and articulatory resources.

A question which arises from the observation that different word types are being employed by the child is that of order of acquisition. Barrett (1995, p. 367), advising caution, reports a study (FENSON et al., 1993) that suggests children's early vocabulary is, during the 50–100-word phase, dominated by nouns; verbs and adjectives then start to become more prevailing, but the proportion of verbs plateaus at around 400–500 words, while adjectives continue to rise proportionally until 500+ words. Barrett's caution notwithstanding, Waxman and Lidz (2006) are more emphatic on their assessment of order of acquisition, maintaining adjectives and verbs can only be fully learned as a function of first being able to identify nouns, which would justify the preponderance of nouns. However, it should not be assumed that young children categorize words according to such grammatical distinctions (p. 368). Ingram (1999, p. 144–147) reports on other studies which, instead of assuming these syntactic categories, introduce more semantically-oriented classes. These include specific nominals (names for unique things like "Toto" and "mommy"), general nominals (e.g. "doll," pronouns), and action words ("eat," "mommy"). On average, general nominals make up most of the production vocabulary and increase in number relative to specific nominals as new words are learned.

Nonetheless, this distribution changes when we consider individual vocabularies. A tendency is observed when we look at the proportion of words used by individual children. *Referential* children tend to use more general nominals, whereas *expressive* ones, while still mostly producing general nominals, tend to have a proportional reduction of these in favor of a larger number of personal-social words (e.g. "yes," "bye-bye," "hi") (BARRETT, 1995, p. 368). One reason for this could be differences of actual language use.

Furthermore, if we turn our attention to the comprehension vocabulary, another interesting distribution is revealed. Early in development most words understood by children are action words, rather than nominals. Ingram (1999, p. 146) argues that this

suggests a learning path beginning with more action words in comprehension, which are then superseded by general nominals by the time children start producing words. This would make sense if one looks at the input, since the words they receive are mostly general nominals and it is expected that these will make up most of the final vocabulary.

Before moving on to some telling patterns of meaning during early word learning, I will briefly touch on a topic alluded to in the beginning of this section. It is a mathematical fact that children have to somehow accelerate their rate of learning if they are to reach the target of 60,000 words mentioned above. The standard explanation has been that children go through a spurt or explosion, sometimes associated to the so-called *naming insight* (BARRETT, 1995, p. 367). However, it needs not be the case that there is a definite point at which the rate of word learning increases enormously[2] (an "inflection point" as per Ganger & Brent, 2004). First of all, there is evidence showing that at least some children do not show a spurt (BARRETT, 1995, p. 363). Second, the definition of spurt has been questioned. Bloom (2000) argues that previous studies had defined the spurt on the basis of a threshold of the type "acquiring a certain amount of words in a certain period of time." For example, Gopnik and Meltzoff's (1986) criterion is the rate of 10 or more words in a period of three weeks. The problem is, according to Bloom, that in order to reach the adult vocabulary, children will inevitably catch up with these rates and beyond. Thus, a gradual increase in rate could also explain children's improving word-learning abilities[3].

A much less controversial ability is fast mapping (CAREY; BARTLETT, 1978). Carey and Bartlett observed that, in the right situation, children can notice the presence of an unknown word and learn some of its semantic content in a single trial. In their original experiment, a set of objects (for example, trays and cups) was placed in the environment and then children were casually asked to "bring a chromium cup, not the red one, the chromium one" (p. 271). The researchers report children could successfully complete the task even though "chromium" was a novel color word, given this explicit instruction and the contrast with another presumably known color. Children were also able to remember some of the meaning associated with the novel word a week later. This effect has been experimentally reproduced for nouns and verbs (GOLINKOFF; JACQUET, et al., 1996).

So far, we have discussed vocabulary size, improvements in learning rate, and which types of words are learned in which order. It was said that children acquire words that seem to be context-bound (BARRETT, 1995), however this is not the only meaning pattern

[2]In personal communications with Pablo Faria, he pointed out it is at least surprising that a spurt should be a moot point. His expectation is that the Zipfian distribution of words would initially give children a hard time learning words. This is due to a small number of words, consisting mainly of functional items, being very frequent and the rest appearing relatively infrequently. However, as children learn these more frequent words, there could be a point where most common patterns would be known and then less frequent words would suddenly become learnable by the very fact that these would be the only varying elements. To his point, recent research on Zipfian distributions indicates they make for a better learning environment than uniform ones (HENDRICKSON; PERFORS, 2019).

[3]A study with an improved definition of spurt and showing that not all children go through it is reviewed back on page 33 of this work.

observed. Barrett notes (p. 369–371) that, besides context-bound words being somewhat extended, meanings may undergo other phenomena. For example, children may underextend or overextend the meaning of referential words. A word, say *ball*, may be exclusively used to refer to balls existing in the surroundings, or to pictures of balls in books. However, children are also found to overextend word meanings, sometimes generalizing them incorrectly, such as calling an orange a *ball*. These incorrect overextensions occur for one third of the words in the vocabulary (INGRAM, 1999), although most of the time happening only once. This suggests on-the-spot, communicative pressures might lead the child to select the closest word readily accessible in memory, even when they have already acquired the correct word as evidenced by their presence in the comprehension vocabulary.

An interesting fact about the way overextensions happen is reported by Barrett (1995). When children employ the same word for incorrect referents, these seem to share features with the original referent for the word but not among themselves (p. 371). The example he gives is of a child using "the word *clock* to refer to clocks, a circular bracelet and the sound of dripping water" (p. 372). The common factor between the bracelet and the clock are their roundness, while the dripping water has a repeating rhythmic sound that resembles a clock's tics and tacs. Yet, there is nothing in common between the two referents incorrectly named. Similarly, function is another factor that leads to overextensions.

Finally, Barrett points out that underextended words may be overextended and vice-versa. For example, *ball* may begin by being used to refer only to a unique ball. As time goes by, however, it may be incorrectly extended to "oranges, pumpkins, peas, round beads on a necklace" (p. 374) etc. The same has been observed of overextended words which have their range of reference shrunk. In particular, this appears to happen as children acquire more specific words to talk about the world. Barrett reports (same page) a case in which a child overextended *car* to trucks, however upon learning the word *truck*, started a phase of intermittent use of both terms to refer to trucks. After a transitional phase, both words became specialized. This pattern of word meaning evolution suggests a learning strategy that expands or retracts as more data is observed, until stabilizing on the correct word meaning.

In summary, children are thought to learn words following their own individual paths while falling into some distinct patterns. First and foremost, there is a gap between comprehension and production vocabulary that persists into adult life. Second, they may start speaking by producing idiosyncratic "words" but will always show idiosyncratic meanings for some proportion of words, be it due to strongly binding meaning to context, or because meaning is being under or overextended. Also, some researchers firmly argue that nouns are acquired before verbs and adjectives, while others that different trends are observed if words are categorized functionally rather than syntactically. Whichever way you cut the pie, there seems to be at least an inkling of order. Finally, whereas a spurt in the traditional sense might not be recorded for all children, it is a fact that as time goes on they become better word learners, leading to fast mapping. I now turn to the question

of which of these empirical aspects can be explained by the models discussed before, and if there are any dark corners left to be explored.

3.1.1 Empirical evaluation of the models

The validity some of the models explored in Chapter 2 is predicated on being compatible with children's developmental patterns. In this section, I argue that there is one aspect of this development which has been to some extent captured, while the other aspects are still largely left unaccounted for.

To start with the positive evaluation, almost half of the models (SISKIND, 1996; YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010; FRANK; GOODMAN; TENENBAUM, 2009) directly mention simulating either vocabulary spurt or fast mapping. All of these models argue that fast mapping is a function of accumulating word knowledge. This makes theoretical sense and is in line with the empirical observations regarding this phenomenon. Children are able to fast map if they have enough confidence on the rest of the linguistic environment (although other sources also help). If a cross-situational model has accrued enough information that a single word in an utterance is unknown, then that single word may be mapped into the correct meaning with little effort. The only requirement seems to be that either a) there is only one meaning hypothesis or b) if there are multiple meaning hypotheses (a competition for alignment) but constraints placed by the other words can neutralize this ambiguity. I would argue that only Yu's and Frank's models could fast map in situations like b).

Let us assume the utterance "Look! A cat!" is paired with the concepts CAT and DOG. Further, let us assume that the only unknown word in this sentence is *cat* and also that the mapping *dog* → DOG has already been acquired. The ambiguity is whether *cat* should map to CAT or DOG. Because Yu's (2008) model represents meanings in a network, where information about other mappings increase or decrease the likelihood of mappings under consideration, the fact that *dog* already maps into DOG will possibly make the model learn the correct mapping in this trial. Frank's (2009) model, on the other hand, processes all utterances in a batch, and the knowledge of the word *dog* will also lead to the correct mapping, given that their model includes a "soft preference for one-to-one mappings" (p. 582).

The vocabulary spurt is also explained in terms of accumulation of prior knowledge and is likely mechanistically dependent on the onset of fast mapping. Given the Zipfian distribution of language corpora, frequently-occurring words will be slowly acquired in the beginning of the process. As the system collects frequent words, contexts where new, less frequent ones appear will be more telling since unknown windows will appear in highly informative contexts. In other words, improvements in rate of learning are due to the growing vocabulary, which progressively reduces the size of the problem. This is the basic explanation proposed by Siskind (1996) and Yu (2008).

Fazly, Alishahi, and Stevenson (2010) also report a vocabulary spurt, however they are able to offer an explanation of why some children do not exhibit this pattern (GANGER; BRENT, 2004), as discussed in the previous section. Their model considers a word acquired when one of its alignment probabilities exceeds the threshold $\theta = 0.7$. With this parameter setting, the model shows a pattern of acquisition which resembles the vocabulary spurt. However, if this value is lowered to 0.5, then the line of acquisition is smoother, indicating a more steady growth in the rate of acquisition. Fazly et al. argue this captures the fact that children may be more or less conservative before producing a word, which would then be added in their vocabulary tally.

Now I turn to the other patterns observed in children's lexical acquisition. First is the fact that early acquisition may begin with idiosyncratic "words." These are not at all considered by the models, since they simulate comprehension and not production. However, given their complicated linguistic status, I will skip over this discussion. More interesting will be understanding whether these models are able to capture context-bound, under or overextended meanings.

Context-bound words are prompted by a conjunction of objects, people and actions. Recall the example of the child who said "car" only when looking at cars driving by out of a window in the living room. This seems to indicate a sort of superholophrastic use of the word, almost as it stood for the whole situation. The models are sorely lacking an explanation for this.

Underextension happens when a word is employed to only one referent, or to a limited number of referents. For example, when only Rex is "dog," and not other dogs out in the street. As we have seen, children can stretch (sometimes even too much) the meaning of initially underextended words, which means they can realize "dog" maps into the general concept. However, underextending can be a useful strategy when acquiring proper nouns, which are nothing more than nouns specific to certain individuals. However, none of the models discussed here are apt to capture this behavior, since their meaning hypotheses are already applying a kind of categorical constraint, as will be discussed in the next section.

Overextension is also not captured. One could argue models which start by distributing probabilities to meaning hypotheses and then adjust their association strength are overextending the meaning of a word, however this would be inaccurate. The stream of meanings which cooccur with each utterance is, again, filtered by this categorical constraint such that concepts which could be initially lumped together are not. For example, if the child hears "the moon looks like a ball," this utterance will be paired with the concepts {MOON, BALL}. The lack of concept formation and development, which apparently walk hand in hand with lexical growth, is a major limitation of the models[4].

Finally, since most models only deal with noun learning, they are not able to say anything regarding order of acquisition. Even Siskind's (1996) model, which does differentiate

[4]However, see Plunkett et al. (1992) and Roy and Pentland (2002) for computational examples of how the development of the lexicon and of categories may interact.

between verb- and noun-like meanings, does not report if an effect of order was identified. It should be said, however, that some studies purport to be models of *early* lexical acquisition (see FRANK; GOODMAN; TENENBAUM, 2009; TRUESWELL et al., 2013; STEVENS et al., 2017), that is, simulations of how children can bootstrap learning of nouns which then opens the possibility of learning more different types of meanings, such as argument-taking words. This ignores the fact that children are thought to have a comprehension vocabulary with a number of verbs (INGRAM, 1999) and that words employed holophrastically are in essence indistinguishable from verbs.

3.2 Theoretical problems of lexical acquisition

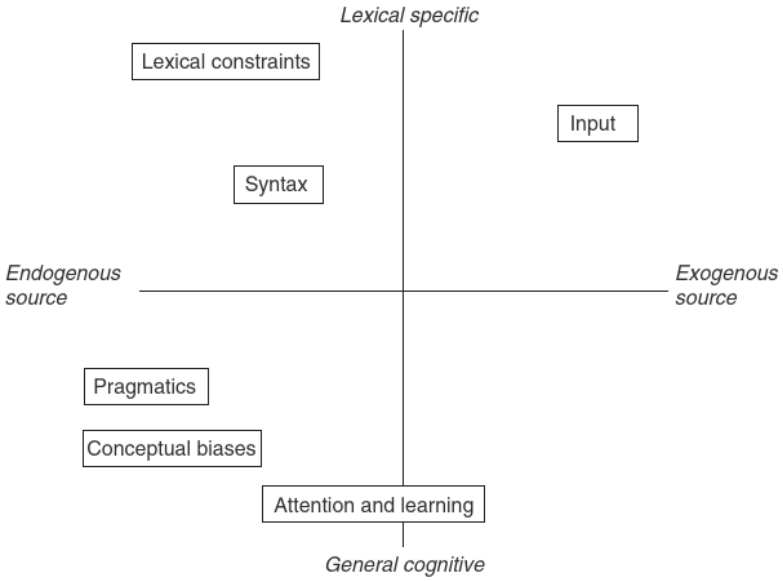
Children learn words so matter-of-factly that, as Bloom (2000, p. 4) puts it, it takes philosophers such as Quine (1960) to even question the natural assumption that lexical acquisition is an easy task. Landau and Gleitman (1985) spend half a chapter arguing that the empiricist John Locke's position – that children are presented names for things around them, thus learning these words – is in fact riddled with issues. Quine's indeterminacy of reference, and in particular its linguistic rendering called *the mapping problem*, is the central issue in many accounts of word learning.

Word-learning is a problem of induction par excellence. The form words (or signs) take have nothing to do with their semantic content, and thus any mapping form-meaning is arbitrary. We have as many words for cat as there are, have been or will be languages in the world. Furthermore, children have a vast number of perceptual as well as conceptual impressions to draw from. If we are immersed in the world and someone utters words in an unknown language, my guess is as good as yours.

Let us briefly review some complications arising from the mapping problem. Children have access to world-situation pairings, however end up acquiring broader word-meaning pairs (LANDAU; GLEITMAN, 1985). For example, although children have a limited experience of dogs in the world, they somehow have to infer that the word *dog* applies not only to the canine set so far observed, but also extends to any other dog that might be encountered. The connection of the word is not to the world but to some abstract category encompassing all of its members.

There are too many possible and logical encodings of experience available (p. 4). The same animal – say, the family's cat – can be described linguistically in higher or lower levels of the conceptual hierarchy: a living creature, an animal, a cat, a kitten. Further, there is no way to tell if a part or even a property of the cat is being referred to. One might be naming its tail, whiskers, or the fact that it is meows. In summary, there is too much evidence in favor of an array of possible linguistic construals of a scene. Another issue (p. 5) is that this can be exacerbated to the point where the child infers only wrong meanings about what was said. If she hears a sentence like "it's time for your dinner" while

Figure 15 – Six mechanisms to explain how children learn words, plotted against two axes: how specific the cognitive machinery is (lexical specific or general) and what source of information is favored (external or internal). From Diesendruck (2007), p. 258.



tending to her toys, what is there to stop the child from mapping *dinner* to an object just before her? Finally, meanings can be abstract (p. 5). Consider for instance the verb *want*, which can reference other people's intentions and thus unobservable mental states. Even seemingly simple nouns, like *pet*, have no direct extra-linguistic counterpart.

Unrestrained hypothesizing, we have seen, is not in children's best interest. It must be the case that somehow there are restraints to the kinds of word meaning hypotheses children can and do entertain. The search for these restraints or biases has guided research on how children beat referential uncertainty and efficiently solve the mapping problem. Evidently, there are many sources of information – internal or external – that children could tap into.

Diesendruck (2007) classifies a number of theoretical restraints in relation to two axes: how specific the mechanisms are, that is, if they operate on the level of lexical acquisition or of more general cognitive learning tasks; and also what source of information is favored, the two extremes being exogenous (e.g. speech) or endogenous (e.g. innate assumptions about language). The proposals are plotted in Figure 15.

With the exception of *input*-oriented accounts, most theories of lexical acquisition draw their power from endogenous sources of bias. That means most proposals are in fact looking for restraints inside children's minds, rather than in the data they can observe in the world. This is consistent with observations of the lexical development of children whose access to sensory input is curbed in some way. For instance, blindness and deafness do not affect normal word learning. Deaf children learn signed language just as well as hearing children (BLOOM, 2000, p. 7) and blind children reach the same level of linguistic ability as their sighted counterparts (LANDAU; GLEITMAN, 1985) – although admittedly, with a seemingly

later onset of linguistic comprehension and production at least partially attributable to other clinical factors. If such limitations do not impair word learning, it must be the case that either the essential information in the input data has not been tampered with by these sensory limitations, or that any missing information is compensated by internal processes or innate expectations.

One very influential proposal of hypothesis restraint has been appealing to specific *lexical constraints* (DIESENDRUCK, 2007, p. 262). These would limit the inferences children make from the very outset of the word learning process. Markman, for example, proposes constraints which are often incorporated by computational models. She presents (MARKMAN, 1990) three interlocking constraints: the Whole Object Assumption (words refer to whole objects), the Taxonomic Assumption (words are not individual-specific, but extend to the whole class of that thing), and the Mutual Exclusivity Assumption (objects only have one name each). Although these biases guide children's hypotheses, they can be overcome if data clearly point to another direction. If children are biased in these ways towards words specifically, it is expected other cognitive tasks would not be similarly constrained. Evidence has suggested linguistic inferences do differ from those related to non-linguistic ones. For example, new words are extended taxonomically as opposed to thematically. This is verified in experimental settings: when asked to pick up an object based on another one just given a novel name, children choose objects belonging to the same class. However, if asked to pick up an object which "goes with" another one, they tend to choose based on thematic similarity.

While Markman's work concerned mainly noun acquisition, other proposals have looked at *syntactic* constraints that might guide verb learning. While Pinker attempted to show semantics could fully explain the acquisition of verbs (PINKER, 1996, 1994), Gleitman and her colleagues (GLEITMAN, 1990; NAIGLES, 1990; FISHER et al., 1994) argue that semantics alone underdetermines verb meaning which, in some cases, can only be learned by considering verb argumental positions. They maintain that verbs which express a change in point of view (e.g. the pair *give* and *receive*) encode the same external situation from different linguistic perspectives. Giving implies a receiving, and to the child the difference could only be established once she realizes, for example, the semantics related to the syntactic positions of subject and indirect object in sentences such as "mother gave a gift to him" and "he received a gift from mother." The facts that verbs appear later than nouns in children's vocabularies and that they seem to be acquired only once nouns can be properly identified (WAXMAN; LIDZ, 2006) support this position, which came to be known as *syntactic bootstrapping*.

Another source of restraint comes from social cognition and interaction, a group of approaches labeled *pragmatics* by Diesendruck (2007). According to this account, children's main interest is determining what is in her interlocutor's mind; see the book-length discussion of theory of mind by Bloom (2000) or the effects of joint attention by Baldwin (1993, 1995). By their second year of life, children already show some sensitivity to other

people's states of mind (DIESENDRUCK, 2007, p. 265) This is instantiated, for example, by noticing and keeping track of speakers' referential intentions. Evidence supporting this mechanism shows that children suppose speakers "might know the common names of novel objects even if they have never been exposed to them, but will not know the proper names of objects under the same circumstances" (same page). Biases which appear to be lexically specific could be explained under this account. For example, mutual exclusivity could be a result of the child knowing that if the speaker wanted to be handed an object whose name the child knew, then the speaker would use that name and not a novel one. Having a theory of what the speaker knows or intends to say allows children to make inferences useful to learn words, although not limited to this task.

The sources of constraint just reviewed apparently solve the problem of unrestrained hypothesizing that plagues the idealized inductive learner. However, they do not constitute a "mechanism of learning" in the cognitive or psycholinguistic sense of the term. Such constraints could well explain how children learn a single word from a certain linguistic and extra-linguistic context. According to Akhtar and Montague (1999), this was indeed the view implicit even in experiments designed to test these constraints. However, what some of these views seem to leave out – with the expressed exceptions of Pinker's and Gleitman and colleagues' – is the fact that problems of induction can be solved by a strategy that proposes hypotheses from an initial observation, then test these hypotheses when given the chance. The only prerequisite is *memory*. If children can remember some of the prior instances of use of a word and associated meaning hypotheses, then she can exclude noncongruent ones until she is left with the right answer.

The genealogy of the term *cross-situational learning*, as far as I could determine, is: Fisher et al. (1994), then Siskind (1996) and preceding related work, and finally Akhtar and Montague (1999) and Yu and Smith (2007)[5]. When push comes to shove, word learning has to happen in a real-time manner. Siskind adopted this strategy to explain how lexical development could unfold in time, while the latter two investigated the matter experimentally in human subjects. If this lineage is correct, then we have before us an interesting case of theoretical proposal later made more concrete by the intrinsically explicit nature of computational implementations, before being picked up by psycholinguistics.

In fact, Siskind's model makes it abundantly clear that cross-situational learning is the mechanism that weaves together other (simulated) cognitive faculties (see section 2.1.1). While one faculty is responsible for segmenting speech into words, another has the duty of coming up with conceptual representations for the events going on in the world. These representations are, of course, exactly the meaning hypotheses that have been the central discussion of this section. This means any conceptual faculty would have to be restrained

[5]Nowadays, research on this hypothesis-testing learning strategy when applied to word learning can be found in troves online. Unfortunately, time limitations did not allow for a comprehensive investigation of all experimental findings; however, many of the existing cross-situational learning models have been included in this thesis.

in how many and what kind of hypotheses it generates. These restrictions could come from any number of sources – the more, the merrier – and would by reason increase the likelihood of the conceptual faculty coming up with correct meaning hypotheses. After all, when (for example) Baldwin (1993) proposes joint attention as a constraint, he is at the same time choosing a high-information source that will assist and not undermine children's chances at finding the correct meaning.

This point – that cross-situational learning works in tandem with a constrained hypothesis-generating faculty – is stressed here because even in foundational works, cross-situational learning seems to be *pitted against* these theoretical sources of constraint (AKHTAR; MONTAGUE, 1999; YU; SMITH, 2007), instead of being presented as the way hypotheses can be slowly (or quickly, given enough accumulated information) confirmed or rejected.

This matter is investigated in the next section, alongside with other considerations on the state of models of cross-situational word learning. I will attempt to answer the questions of what the current modeling work says about these theories of word learning and what aspects of theory are still being ignored and why.

3.2.1 Theoretical evaluation of the models

The natural place to start a theoretical evaluation of the models under consideration seems to be by asking the question of whether Quine's problem of referential uncertainty has been appropriately modeled. Quine posits a challenge for translation in the extreme. Word learning is analogous to translating from an unknown language if we accept that the unknown vocabulary are the words of the child's mother tongue and that the target is the mental language of conceptual representation of reality. This is the view underlying the mapping problem, as it is construed by many researchers[6], among them Landau and Gleitman (1985), Markman (1990) and Fisher et al. (1994). In this way, Quine's challenge is equivalent to the mapping problem.

In fact, computational modeling can be laureated for better defining the mapping problem. Siskind (1996) is the first to give it an implementation and in doing so to clearly define modules responsible for word segmentation and perception/conceptualization. As a consequence of this more formal description, new problems creep into the mapping problem which I have discussed in preceding theoretical descriptions. When forced to deal with how the learner stores and processes the input data, Siskind realizes the learner will struggle with the effects arising from the mapping problem in the specific case of lexical acquisition. These phenomena go beyond referential uncertainty (and the extreme case of noise): they are synonymy, homonymy, and polyssemy. It turns out that the arrows leading from words to concepts draw a complicated network of connections.

[6]However, not all researchers see word learning as a mapping problem. For an opposing view, see Tomasello (2001).

It seems fair to say that computational modeling of lexical acquisition has done justice to abstractly representing the mapping problem, which is at the core of the theories just discussed above. As a matter of fact, computational efforts have further developed the problem by revealing issues arising at the processing time.

However, how well have these models represented the two *domains* of words and concepts? I will argue that here is where the biggest issue with these models lie.

Consider again the discussion in section 1.2. The way these models construe the hypotheses available to the learner follows two main trends: either they are any object available in the extra-linguistic context, and thus only nouns, or an abstract representation of utterance meaning like Siskind's. What is this telling us about these models' conceptual representations?

Most models, whether they explicitly declare it or not, assume at least some of Markman's (1989; 1990) constraints. The most common is the Whole Object Assumption. The objects available in the extra-linguistic context are encoded (sometimes manually, sometimes automatically) in the utterance-meanings stream as indivisible conceptual representations which have so far been called "atomic" (the exception being Siskind, 1996). So far, so good.

The Mutual Exclusivity Assumption is not acknowledged so universally: Siskind (1996), Yu (2008) and Stevens et al. (2017) are the only to explicitly do so. Fazly, Alishahi, and Stevenson (2010) and Frank, Goodman, and Tenenbaum (2009), on the other hand, contend their models are able to achieve this behavior without any dedicated rules, but rather it being a byproduct of the model's design. However, Stevens et al. (2017) argue that built into these statistical models are biases that favor one-to-one mappings. Frank, Goodman, and Tenenbaum (2009), for example, "chose a prior probability distribution that favored parsimony, making lexicons exponentially less probable as they included more word-object pairings" (p. 579). If we accept Steven et al.'s compelling argument, then all these models assume both whole objects and mutual exclusivity.

Notice that these assumptions affect different parts of the models. The Whole Object Assumption is embedded in the input data, while Mutual Exclusivity is a property of the learning algorithm. In other words, the concepts simulated by (most of) these models are from the outset restricted to whole objects. Exceptions include most notably Siskind (1996), whose model learns sets of (synthetic) semantic features or conceptual symbols assembled into a hierarchical structure, and Alishahi et al. (2012) and Faria (2015), whose models learn sets of semantic features. Thus, the first assumption affects the simulated conceptual representations.

This has an important implication. For models which represent meaning atomically, under and overextension are phenomena left out of the simulation. However, this also means that some relations between terms are impossible to be captured by these models. The Whole Object Assumption has interesting corollaries. If a context is clearly focusing an object for which the learner already has a name, she has to make some decisions. The

first is that the object has more than one name (synonymy), which is normally disfavored by Mutual Exclusivity. A second hypothesis is that the name refers to a part (or maybe property) of the object (meronymy). A third possibility is that the word is either more general (hypernymy) or more specific (hyponymy) than the name already known for that object. Despite these possibilities, the fact that concepts are static, atomic entities in modeling work means these hierarchical relations between terms cannot be represented whatsoever. An important aspect of Markman's work has been completely left out of modeling.

Mutual exclusivity, on the other hand, says that referents whose names are already known are less likely to receive a novel name. Probabilistic approaches can be thought to model this, while having the advantage that mounting evidence may counterbalance this assumption, a point which Markman (1989) herself upheld. Models such as Yu and Ballard (2007), Yu (2008) and Fazly, Alishahi, and Stevenson (2010) can learn synonyms; Fazly et al. even report that their simulations showed an initial resistance followed by acquisition of synonymous words, which captures children's behavior (p. 1052)[7].

Markman's Taxonomic Assumption is yet another way in which the input data in these models is being constrained. Originally, Markman and Hutchinson (1984) observed that children solve tasks differently when given an object label. If they are asked to find another instance of an object without being given an explicit name, they search for thematically related objects. If however they are asked to "find another *fep*," then they are observed to look for an object related taxonomically. The fact that the models under consideration do not receive visual (or other perceptive) input and then have to find the categories associated with each object available means they assume a strong taxonomic rule. However, as a consequence the models cannot explain cases of under or overextension. Children cannot always hone in the correct level of abstraction for a given concept. This reveals that the Taxonomic Assumption, a phenomenon of lexical nature, is constrained by the correctness of the hypothesis a child makes. However, since these computational models always give the problem half solved, difficulties arising from the cross-interaction of under and overextension, Markman's constraints and the online nature of word learning are not simulated.

An interesting question is if models which represent meaning as features or compositionally (SISKIND, 1996; ALISHAHI et al., 2012; FARIA, 2015) can capture incorrectly extended meanings. In these models, an utterance is presented with each word carrying a series of features or conceptual symbols. Upon first encountering a word, the learner will initialize it to be the collection of all cooccurring features. However, notice that any given object will always be represented as the same set of features, which is equivalent to saying

[7] Although deterministic models would be barred from learning synonyms if they assumed hard mutual exclusivity, it seems that Siskind's (1996) model is not as far as I can see. In the case of this model, mutual exclusivity only applies for individual utterances.

the learner always makes correct inferences about the category of each object. Thus, no model can capture under or overextensions as they are currently conceived.

Moving on to other kinds of constraints, the model by Yu and Ballard (2007) investigated two sources of information: eye-gaze and prosody. These are two examples of constraints at the social-pragmatic level (BALDWIN, 1993; BLOOM, 2000, for example). The results from the model indicate that adding in these information channels does help learners become better at selecting the correct referent, with eye-gaze having a greater effect than prosody. The researchers speculate that this is due to prosody having the double function of emphasizing a word (which might be the referent being attended to), while also serving to grab children's attention. This is an instance of modeling work that produces interesting predictions to be verified observationally or tested in experimental settings. Apart from the caveats mentioned above, this seems to be a fair representation of how sources of information may be included in a model.

Siskind's (1996) model is the only to represent verb-like semantics. This means his model has complex conceptual expressions with argumental positions, analogous to many verbs. Even though his corpus of utterance-meanings pairs was generated synthetically, the underlying semantic structure mirrors the theoretical properties of meaning under Jackendovian representations (JACKENDOFF, 1983, 1990). Despite this, his simulations do not include any syntactic restrictions similar to those proposed by Gleitman (1990) and Fisher et al. (1994). In fact, Siskind suggests that his positive findings are evidence that acquisition of verbs is possible without any syntactic information (p. 84). However, this comes with the assumption that children are capable of choosing between two hypotheses in very dubious situations. For instance, imagine a child observing Sue give John a guitar. Gleitman and colleague's point is that the child could not help but conceive (at least) two possible meaning representations:

CAUSE(**Sue**, GO(**guitar**, TO(**John**), FROM(Sue))) or
GO(**guitar**, TO(**John**), FROM(Sue))

From a strictly interpretive point of view, it is impossible to select a single hypothesis. This would mean then that whenever the learner heard an utterance containing *give* or *receive*, their lexical entries would ultimately include the same number of argument positions and all four conceptual symbols (namely, CAUSE, GO, TO, FROM). The fact that Siskind's model does acquire verbs presumably in spite of this problem means that, at some point, one meaning hypothesis was either quashed or not even proposed. However, Siskind does not provide a mechanism for explaining how the perceptual/conceptual faculty would be able to do this, meaning hidden constraints are at play.

We have seen that, either on purpose or because of tentatively plausible simplifying assumptions, models of lexical acquisition incorporate different types of restrictions. These may pertain to different levels: they may curb the nature of concepts given as input data, in particular in the case of models which are restricted to noun learning; they may guide the process of building mappings, such as by assuming no synonymy or penalizing it; and sometimes these restrictions are added in with the explicit goal of assisting in the choice of correct referents. However, I have argued that all in all, these models' simplifications may impede a more sophisticated and realistic simulation of lexical acquisition.

The first problem is that meaning representations are poor. They cannot capture very important facts about children's mistakes when cutting up the world into concepts. Specifically, under and overextension could interact with the mapping process. Since prior partial or full word knowledge guides new mappings in many of these models, underextending could cause spurious meanings to be accepted or good learning instances to be discarded, and overextension could block new meanings since the learner would consider that name already learned.

Aside from this, there is no explanation for how meaning representations are generated and selected. Manual annotation might introduce the biases competent speakers carry into the data, the most obvious being that categories are always estimated at the adult target. If models are serious about modeling early noun acquisition, then they also have to provide a more robust simulation of the interaction of evolving conceptualizations and vocabulary.

Incidentally, there are issues on the vocabulary side as well. As it was alluded to above, the assumption of Mutual Exclusivity might signal cases when children will have to focus on parts of objects, or higher or lower levels of categorical specificity. These lexical connections mean words (or maybe concepts) are organized in a relational network which is surely under construction during lexical acquisition. An organized lexicon is completely missing from all models[8]. Siskind's model is an exception in that it minimally distinguishes words into senses, which then map into meanings; however, the final lexicon is still an unstructured list of senses.

What is perhaps most telling about this point is that the theories of lexical acquisition reviewed earlier do not consider learning over time, but only "on-the-spot" meaning disambiguation. This means researchers trying to model word learning have to make simplifications where theory is underspecified. Of similar concern is that basic terms, like *word* and *concept* or *meaning* are not defined. These theoretical decisions have ripple effects in modeling, where each and every aspect of the problem has to be clearly specified. Therefore, although lexical acquisition modeling still has some extra miles to go, one cannot say theory is much ahead as far as implementability is concerned.

[8]But see, for example, Li, Farkas, and MacWhinney (2004), whose neural network can explain how meaning (and phonetic) representations can self-organize in a topography of simulated neurons.

4 Closing remarks

What is the state of the art for models of lexical acquisition? In order to answer this basic question, the computational problem of word learning was broken down into its basic parts (Chapter 1); then each model was studied in detail, so that a picture of the general contributions could be painted (Chapter 2); and finally, these general contributions were compared to the empirical and theoretical status quo (Chapter 3). I now summarize my findings, commenting on the model's input data, mechanisms of acquisition and methods of evaluation. I also try my best to acknowledge this work's limitations. Lastly, I argue for a view of future research that might guide a new generation of models.

4.1 Back to the basic issues

In Chapter 1, I laid out an overview of how lexical acquisition is construed by computational models. In these studies, input data, in the form of two synchronized utterance-meanings streams, are presented to a learning mechanism, which extracts a lexicon. Each lexical entry is a mapping established between word and meaning. There is also no structure in each input stream, that is, utterances are just lists of words and meanings are just lists of concepts (except for SISKIND, 1996). Even though both input streams are synchronized, that does not mean the problem is easy, since utterances contain multiple words and meanings contain both correct (but sometimes none) as well as incorrect meanings. Furthermore, the lexicon generating the utterances in the corpus has properties such as homonymy and synonymy.

Let us first look at the mapping problem. Is it a fair representation of the challenge faced by children? In essence, the lexical mapping problem is a specific version of any problem of translation where mappings have to be established between two domains. Indeed, some models investigated here found their cross-situational engines in well-known algorithms of machine translation (YU; BALLARD, 2007; YU, 2008; FAZLY; ALISHAHI; STEVENSON, 2010). However, there is one important admonition to be made about this abstraction. Word learning does not happen between two immutable sets of elements in different domains. As the vocabulary grows, hierarchical relations among words have to be discovered and then established. Meronyms can be learned, if we follow Markman's (1989) arguments, only when a child already knows the name for a referent, and children have difficulty using superordinate terms when they already know basic-level names (BLOOM, 2000, p. 66). These are essentially lexical phenomena which may impact acquisition and thus have to be taken seriously.

Similarly, the conceptual side of learning is not static. This is revealed by the fact that children's early word meanings can be bound to context (as if a very limited, perhaps

event-based (BARRETT, 1995), concept had been created), or under/overextended. These phenomena cannot be represented by mapping a word to multiple different concepts, since when a child calls the moon a "ball" that is not a case of homonymy. Rather, as we previously discussed, children seem to select a property of the original referent (*roundness*, in this case) and use the word to refer to something sharing this property. In order to fully simulate the problem of word learning, inaccuracies in concepts, not only their overabundance, have to be factored in.

In fact, there is computational (ROY; PENTLAND, 2002) as well as empirical evidence (WAXMAN; LIDZ, 2006; FULKERSON; WAXMAN; SEYMOUR, 2006) that the process of word segmentation and concept formation are interwoven. Children are not perfect at segmenting words at 10–12 months, however become better as they notice certain concepts have certain names. Conversely, naming can hint at the existence of a concept, which might spark the finding of individuals; but also naming in the context of several objects of the same category can lead to the discovery of their unifying trait. Clearly, the mapping problem as it has been represented is naive at best and the consequences of facing the real-world organization of lexical and conceptual systems are still undetermined.

Still on the conceptual side, most models of lexical acquisition cannot claim to have a plausible solution for Quine's (1960) challenge of referential uncertainty. No model has presented a theory or mechanisms for generating conceptual representations that might capture children's meaning guesses. This is, after all, the crucial question. The fact that most models use human annotators in order to extract meanings (usually, referents) from a scene and pair them with the utterances in the corpus introduces huge bias. Adults already have a mature conceptual system. Further, these meanings represent categories which already solve the problem of extending a word to other instances of the same referent. In addition to these problems, a basically pre-theorical stance (except for SISKIND, 1996) is taken on semantics: no attempt at defining meaning is even tried. Referents are only that: objects in a very limited world.

Perhaps the models should not take the blame alone. As discussed in the end of the last chapter, the underlying theories of lexical acquisition adopted by modelers do not present a clear, well-defined notion of meaning. The same fact is true for the ever so elusive term *word*. Leaving terms undefined already has ramifications for the theory in its quarantine trial period while it is being developed and tested. However, these ramifications become glaring omissions in the unforgiving world of computational modeling. One trivial albeit fundamental issue can be observed when input comes from morphologically richer languages. Here, affixes are productive to the point of becoming an issue to model performance (FARIA, 2015).

In an effort to both abstract from contingent issues of lexical acquisition and get to the core mechanism responsible for word learning, more recent models have become guilty of oversimplifying. Consider the earliest study, Siskind's (1996) model. It featured a structured lexicon with homonymy, a theoretically-grounded conceptual representation for

meanings, clearly announced and implemented lexical and conceptual restraints. By the time the latest models came out (the likes of STEVENS et al., 2017), all of these aspects had disappeared from modeling. Despite embracing more naturalistic data from child-directed language, Siskind's successors restrict themselves to simulating concrete noun learning.

One finding of the computational simulations (see section 2.2) presented in this thesis was that the Siskind Adapted Model (SAM) had better fitting results than any other local or global model for experiment 1 of Yu and Smith (2007). "Adapted" here does not refer to the learning mechanism originally proposed by Siskind, but rather to the fact that I had to include provisions to select referents, thus enabling SAM to be tested in simulated psycholinguistic experiments. What this result essentially means is that a comparatively ancient approach can still be competitive in the current landscape of local, allegedly more parsimonious models.

On the other hand, of course, SAM was found not to fit the data from human participants for experiment 2 of Yu and Smith (2007) or from Trueswell et al. (2013). Worse still, it could not even deal with the experimental design from Koehne, Trueswell, and Gleitman (2013). Nevertheless, I have to point out the limited nature of using experiments to validate a learning mechanism. First, all of these experiments are not fair representations of children's linguistic and extra-linguistic context. In the case of Yu and Smith (2007), a number of words (not an utterance) is presented with a number of images (the "referents"). Therefore, there is no referential uncertainty, only an uncertainty in alignment. Furthermore, there are none of the hints available in daily situations: not factors of language such as syntax or pragmatics, nor factors of interaction such as eye-gaze. The experiment could well be seen as testing people's memory.

Trueswell et al. (2013) make improvements by adding an utterance, such as "Oh, look! A *mipen*!," while also replicating true referential uncertainty by having several referents to the one word. The dubious aspect of their design is presenting n words in a block, then repeating this block *in the same order* four more times. It is precisely this unrealistic presentation order that impedes the memory-restrained version of SAM from completing the task. In real-life settings, children do not experience an orderly presentation of words. A word being used in quick succession may give them enough chances to discard incorrect meaning hypotheses and converge on the correct meaning. The same is true of SAM.

Koehne, Trueswell, and Gleitman (2013) present an elegant experimental design and find that human participants can track more than one meaning hypotheses at a time. Each word in the experiment has one referent which appears with all presentations. The twist is adding another referent which is shown only fifty percent of the time. The order in which this less frequent referent appears is manipulated to demonstrate that its recall depends not on frequency, but fundamentally on the fact that humans select a single referent and stay with it unless they notice contradicting evidence. This memory effect is very clear and not contested here. However, what does this experiment have to say about lexical acquisition *specifically*? Is it not conceivable that any other task of inter-domain

mapping would find the same results? If we accept these designs capture only very limited aspects of lexical acquisition, then using experimental simulations does little in the way of explanatory power for our models.

Having said that, the advances of PbV (TRUESWELL et al., 2013) and Pursuit (STEVENS et al., 2017) are not to be dismissed. They represent a true shift in cross-situational learning mechanisms, are in principle more parsimonious for requiring less memory, and were shown to be competitive with so-called "global" models in corpus processing tasks. A theoretical aftermath was discussed by Yang (2019), who argues that a local learner can profit from highly informative situations much more than global ones, which are too busy juggling all hypotheses at the same time. This is an interesting development, although it is still an open question how it would fare in more realistic simulations, in particular if more complex semantics were introduced.

4.2 Limitations of this work

No work is perfect, and this is no exception. There are a great many ways in which I am still dissatisfied with how far I was able to advance in the discussion. I highlight what in my opinion are this work's greatest deficiencies.

This computational survey is not exhaustive There are a great many studies on cross-situational learning and many more on lexical acquisition at large. I decided, both because of time and scope pressures, to focus on what I found were the most representative studies which modeled referential uncertainty, the central problem according to prominent thinkers (GLEITMAN; GLEITMAN, 1992; YANG, 2019). Notable omissions include but are not limited to:

Roy and Pentland (2002) discuss how sounds can constraint category formation and vice-versa;

Bailey (1997) introduces a model of verb acquisition;

Plunkett et al. (1992) and Regier (2005) and many other connectionist approaches to word learning.

Approaches drawing from machine learning and, more broadly, artificial intelligence; for example Roads and Love (2020). In the bibliography composing the body of this work, they were not considered as a result of not being cited by the papers studied here and, furthermore, as I did not expect they would explicitly tackle the issue of cognitive plausibility. However, this plausibility is, in the end, not well-defined in any uncontroversial or definitive way. That being true, such approaches can definitely bring new perspectives to the discussion[1].

[1] I have to thank professor Dan Yurovsky for pointing this out during my thesis defense. Cognitive plausibility is not a golden standard argument but rather, at this point at least, merely a modeling wish. It is absolutely within reason that inspiration drawn from machine learning or even natural language processing should be seriously considered.

Many other theories By the same token, I decided to narrow my review of theories of lexical acquisition to those few which have direct associations with the models surveyed here.

Generalizations about the models The devil is often in the details, which means any reservations expressed here about the overall present status of modeling work may be softened upon closer inspection of each model individually.

Mistakes in the model The code of SAM was written exclusively by me, which means there might be interpretative or implementational errors either helping or hindering my results.

4.3 The future

Finger pointing often requires pointing directions as compensation. This extensive discussion of the state of the art for models of lexical acquisition has led to an appreciation of two main ways of improving over cave painting towards more sophisticated perspectives.

The first is improving upon the mapping problem. This will imply on a mapping problem of its own, however this time the territory is theoretical and empirical. I propose that a more clear understanding of elemental aspects of word learning is necessary, lest our models continue to carry simplifying assumptions. A clear understanding of the theory behind the *word* or the linguistic side of the mapping problem is needed. This understanding will provide us with better chances of either implementing more realistic models or making simplifications with the full and expressed knowledge of what is left out. The same is true of the conceptual side of the mapping problem. In particular, I propose that theories and their theoretical dependencies be systematically reviewed with an eye towards computational implementability. This way, computational modelers will have a reliable resource for choosing theories or justifying abstractions. On the other hand, this work should also provide us with the chance of being more demanding of theories, which are not always so clearly defined that they can inform computational efforts. This two-way work will be the basis for my PhD research.

The second path, which is really a development of the first proposal, is looking at theories which regard word learning as a complex system. No one source of constraint seems to deplete all referential uncertainty. For this reason, more recent research programs have looked towards integrating different sources of constraint (DIESENDRUCK, 2007). These may be systems ripe for implementation, as long as we have a better representation of both the lexicon and the conceptual/semantic aspects of word learning.

References

AKHTAR, N.; MONTAGUE, L. Early lexical acquisition: the role of cross-situational learning. *First Language*, v. 19, n. 57, p. 347–358, Sept. 1999.

ALISHAHI, A. et al. Sentence-based attentional mechanisms in word learning: evidence from a computational model. *Frontiers in Psychology*, v. 3, 2012.

BAILEY, D. R. *When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs*. 1997.

BALDWIN, D. A. Early referential understanding: infants' ability to recognize referential acts for what they are. *Developmental Psychology*, v. 29, n. 5, p. 832, 1993.

_____. Understanding the link between joint attention and language. In: MOORE, C.; DUNHAM, P. J. (Eds.). *Joint attention: its origins and role in development*. New York: Psychology Press, 1995. P. 131–158.

BARRETT, M. Early lexical development. In: FLETCHER, P.; MACWHINNEY, B. (Eds.). *The Handbook of Child Language*. Oxford: Blackwell, 1995.

BENEDICT, H. Early lexical development: comprehension and production. *Journal of Child Language*, v. 6, n. 2, p. 183–200, June 1979.

BERALDO, R. Explorando um modelo computacional da aquisição lexical. *Revista do SETA*, v. 9, p. 131–145, 2019.

BERTOLO, S. A brief overview of learnability. In: _____. *Language acquisition and learnability*. Ed. by Stefano Bertolo. Cambridge: Cambridge University Press, 2001. P. 1–14.

BLOOM, P. *How children learn the meanings of words*. Cambridge: The MIT Press, 2000.

_____. Intentionality and word learning. *Trends in Cognitive Sciences*, v. 1, n. 1, Apr. 1997.

BROEDER, P.; MURRE, J. Introduction: the computational study of language acquisition. In: _____. *Models of language acquisition: inductive and deductive approaches*. Ed. by Peter Broeder and Jaap Murre. Oxford: Oxford University Press, 2002.

CAREY, S. The child as a word learner. In: HALLE, M.; BRESNAN, J.; MILLER, G. (Eds.). *Linguistic theory and psychological reality*. Cambridge: MIT Press, 1978.

CAREY, S.; BARTLETT, E. Acquiring a single new word. *Papers and Reports on Child Language Development*, v. 15, p. 17–29, 1978.

CARTMILL, E. A. et al. Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, v. 110, n. 28, p. 11278–11283, 2013.

CLARK, E. V. The principle of contrast: a constraint on language acquisition. In: MACWHINNEY, B. (Ed.). *Mechanisms of language acquisition*. Hillsdale: Lawrence Erlbaum Associates, 1987. P. 1–33.

DIESENDRUCK, G. Mechanisms of word learning. In: HOFF, E.; SHATZ, M. (Eds.). *Blackwell handbook of language development*. Malden: Blackwell, 2007. (Blackwell handbooks of developmental psychology).

DROMI, E. Early lexical development. In: BARRETT, M. D. (Ed.). *The development of language*. New York: Psychology Press, 1999. (Studies in developmental psychology). P. 99–131.

FARIA, P. A Computational Study of Cross-situational Lexical Learning of Brazilian Portuguese. In: SIXTH WORKSHOP ON COGNITIVE ASPECTS OF COMPUTATIONAL LANGUAGE LEARNING. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*. Lisbon, Portugal: Association for Computational Linguistics, 2015. P. 45–54.

FARRELL, B. A.; APTER, M. J. The computer simulation of behavior. *The Philosophical Quarterly*, v. 22, n. 86, p. 76, Jan. 1972.

FAZLY, A.; ALISHAHI, A.; STEVENSON, S. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, v. 34, n. 6, p. 1017–1063, May 2010.

FENSON, L. et al. *The MacArthur communicative development inventories: user's guide and technical manual*. San Francisco: Singular Publishing Group, 1993.

FISHER, C. et al. When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. *Lingua*, v. 92, p. 333–375, Apr. 1994.

FRANK, M. C.; GOODMAN, N. D.; TENENBAUM, J. B. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, v. 20, n. 5, p. 578–585, 2009.

FULKERSON, A. L.; WAXMAN, S. R.; SEYMOUR, J. M. Linking object names and object categories: words (but not tones) facilitate object categorization in 6-and 12-month-olds. In: 30TH Boston University Conference ON Language Development. *Supplement to the Proceedings of the 30th Boston University Conference on Language Development*. Somerville: Cascadilla Press, 2006. P. 32–38.

GAMBELL, T.; YANG, C. *Word segmentation: quick but not dirty*. [S.l.]: Yale University, 2005.

GANGER, J.; BRENT, M. R. Reexamining the vocabulary spurt. *Developmental Psychology*, v. 40, n. 4, p. 621, 2004.

GILLETTE, J. et al. Human simulations of vocabulary learning. *Cognition*, v. 73, n. 2, p. 135–176, 1999.

GLEITMAN, L. The structural sources of verb meanings. *Language Acquisition*, v. 1, n. 1, p. 3–55, Jan. 1990.

GLEITMAN, L. R.; GLEITMAN, H. A picture is worth a thousand words, but that's the problem: the role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, v. 1, n. 1, p. 31–35, 1992.

GOLINKOFF, R. M.; HIRSH-PASEK, K., et al. Children and adults use lexical principles to learn new nouns. *Developmental Psychology*, v. 28, n. 1, p. 99–108, 1992.

GOLINKOFF, R. M.; JACQUET, R. C., et al. Lexical principles may underlie the learning of verbs. *Child Development*, v. 67, n. 6, p. 3101–3119, Dec. 1996.

GOPNIK, A.; MELTZOFF, A. N. Words, plans, things, and locations: interactions between semantic and cognitive development in the one-word stage. In: KUCZAJ, S. A.; BARRETT, M. D. (Eds.). *The Development of Word Meaning: Progress in Cognitive Development Research*. New York, NY: Springer, 1986. (Springer Series in Cognitive Development). P. 199–223.

HENDRICKSON, A. T.; PERFORS, A. Cross-situational learning in a Zipfian environment. *Cognition*, v. 189, p. 11–22, Aug. 2019.

INGRAM, D. *First language acquisition: method, description, and explanation*. Cambridge: Cambridge University Press, 1999. 572 pp.

JACKENDOFF, R. *Semantic structures*. Cambridge: MIT Press, 1990.

_____. *Semantics and cognition*. Cambridge, Mass.: MIT Press, 1983. (Current studies in linguistics series, 8).

KOEHNE, J.; CROCKER, M. W. Sentence processing mechanisms influence cross-situational word learning. In: ANNUAL MEETING OF THE COGNITIVE SCIENCE SOCIETY. *Proceedings of the annual meeting of the cognitive science society 32*. Portland: Cognitive Science Society, 2010. v. 32, p. 7.

_____. The Interplay of Multiple Mechanisms in Word Learning. In: ANNUAL MEETING OF THE COGNITIVE SCIENCE SOCIETY. *Proceedings of the Annual Meeting of the Cognitive Science Society 33*. Boston: Cognitive Science Society, 2011. v. 33, p. 7.

KOEHNE, J.; TRUESWELL, J. C.; GLEITMAN, L. R. Multiple Proposal Memory in Observational Word Learning. In: ANNUAL MEETING OF THE COGNITIVE SCIENCE SOCIETY. *Proceedings of the Annual Meeting of the Cognitive Science Society 35*. Pasadena: Cognitive Science Society, 2013. v. 35, p. 7.

LANDAU, B.; GLEITMAN, L. R. *Language and experience: evidence from the blind child*. Cambridge: Harvard University Press, 1985. 250 pp. (Cognitive science series, 8).

LI, P.; FARKAS, I.; MACWHINNEY, B. Early lexical development in a self-organizing neural network. *Neural Networks*, v. 17, n. 8-9, p. 1345–1362, Oct. 2004.

MACWHINNEY, B. *The CHILDES project: tools for analyzing talk, volume ii: the database*. 3. ed. New York: Psychology Press, 9 Jan. 2014.

MARKMAN, E. M. *Categorization and naming in children: problems of induction*. Cambridge: The MIT Press, 1989. 268 pp.

_____. Constraints children place on word meanings. *Cognitive Science*, v. 14, n. 1, p. 57–77, Jan. 1990.

MARKMAN, E. M.; HUTCHINSON, J. E. Children's sensitivity to constraints on word meaning: taxonomic versus thematic relations. *Cognitive Psychology*, v. 16, n. 1, p. 1–27, Jan. 1984.

MEDINA, T. N. et al. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, v. 108, n. 22, p. 9014–9019, 31 May 2011.

NAIGLES, L. Children use syntax to learn verb meanings. *Journal of Child Language*, v. 17, n. 02, p. 357, 1990.

NELSON, K. Structure and strategy in learning to talk. v. 38, n. 1/2, p. 1–135, 1973.

PEARL, L. Using computational modeling in language acquisition research. In: BLOM, E.; UNSWORTH, S. (Eds.). *Experimental Methods in Language Acquisition Research*. Amsterdam: John Benjamins Publishing Company, 2010. v. 27. (Language Learning & Language Teaching). P. 163–184.

PINKER, S. How could a child use verb syntax to learn verb semantics? *Lingua*, v. 92, p. 34, 1994.

_____. *Language learnability and language development*. Cambridge: Harvard University Press, 1996. 435 pp.

_____. *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: The MIT Press, 1989. (Learning, development, and conceptual change).

PLUNKETT, K. et al. Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, v. 4, n. 3-4, p. 293–312, Jan. 1992.

POIBEAU, T. et al. Computational modeling as a methodology for studying human language learning. In: POIBEAU, T. et al. (Eds.). *Cognitive aspects of computational language acquisition*. Berlin: Springer, 2013. (Theory and applications of natural language processing).

PUSTEJOVSKY, J. Constraints on the acquisition of semantic knowledge. *International Journal of Intelligent Systems*, v. 3, n. 3, p. 247–268, 1988.

QUINE, W. v. N. *Word and object*. Cambridge, MA: The MIT Press, 1960.

REDINGTON, M.; CHATER, N.; FINCH, S. Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, v. 22, n. 4, p. 425–469, 1998.

REGIER, T. The Emergence of Words: Attentional Learning in Form and Meaning. *Cognitive Science*, v. 29, n. 6, p. 819–865, 12 Nov. 2005.

ROADS, B. D.; LOVE, B. C. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, v. 2, n. 1, p. 76–82, Jan. 2020.

ROY, D. K.; PENTLAND, A. P. Learning words from sights and sounds: a computational model. *Cognitive Science*, v. 26, n. 1, p. 113–146, Jan. 2002.

SALVETER, S. C. Inferring conceptual graphs. *Cognitive Science*, v. 3, n. 2, p. 141–166, 1979.

SISKIND, J. M. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, v. 61, n. 1-2, p. 39–91, 1996.

_____. Acquiring core meanings of words, represented as Jackendoff-style conceptual structures, from correlated streams of linguistic and non-linguistic input. In: THE 28TH ANNUAL MEETING THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*. Pittsburgh, Pennsylvania: Association for Computational Linguistics, 1990. P. 143–156.

SMITH, L.; YU, C. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, v. 106, n. 3, p. 1558–1568, 2008.

STEVENS, J. S. et al. The pursuit of word meanings. *Cognitive Science*, v. 41, p. 638–676, Apr. 2017.

TOMASELLO, M. Perceiving intentions and learning words in the second year of life. In: TOMASELLO, M.; BATES, E. (Eds.). *Language development: the essential readings*. Malden: Blackwell, 2001. (Essential readings in developmental psychology). P. 111–128.

TOMASELLO, M.; MERRIMAN, W. E. (Eds.). *Beyond names for things: young children's acquisition of verbs*. New York: Psychology Press, 25 Feb. 2014. 569 pp.

TRUESWELL, J. C. et al. Propose but verify: fast mapping meets cross-situational word learning. *Cognitive Psychology*, v. 66, n. 1, p. 126–156, Feb. 2013.

WAXMAN, S. R.; LIDZ, J. Early word learning. In: KUHN, D.; SIEGLER, R. S. (Eds.). *Handbook of Child Psychology*. 6. ed. Hoboken: Wiley, 2006. v. 2.

WOODARD, K.; GLEITMAN, L. R.; TRUESWELL, J. C. Two- and three-year-olds track a single meaning during word learning: evidence for propose-but-verify. *Language Learning and Development*, v. 12, n. 3, p. 252–261, 2 July 2016.

XU, F. The role of language in acquiring object kind concepts in infancy. *Cognition*, v. 85, n. 3, p. 223–250, Oct. 2002.

YANG, C. How to make the most out of very little. *Topics in Cognitive Science*, 12 Mar. 2019.

YU, C. A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, v. 4, n. 1, p. 32–62, 7 Jan. 2008.

YU, C.; BALLARD, D. H. A unified model of early word learning: integrating statistical and social cues. *Neurocomputing*, v. 70, p. 2149–2165, 2007.

YU, C.; SMITH, L. B. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, v. 18, n. 5, p. 414–420, May 2007.

YUROVSKY, D.; FRANK, M. C. An integrative account of constraints on cross-situational learning. *Cognition*, v. 145, p. 53–62, Dec. 2015.

ZIPF, G. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press, 1949.