The Role of Utterance Boundaries and Word Frequencies for Part-of-speech Learning in Brazilian Portuguese Through Distributional Analysis

Pablo Faria
University of Campinas
Linguistics Department
Campinas, SP, 13083-859
pablofaria@iel.unicamp.br

Abstract

In this study, we address the problem of part-of-speech (or syntactic category) learning during language acquisition through distributional analysis of utterances. A model based on Redington et al.'s (1998) distributional learner is used to investigate the informativeness of distributional information in Brazilian Portuguese (BP). The data provided to the learner comes from two publicly available corpora of child directed speech. We present preliminary results from two experiments. The first one investigates the effects of different assumptions about utterance boundaries when presenting the input data to the learner. The second experiment compares the learner's performance when counting contextual words' frequencies versus just acknowledging their co-occurrence with a given target word. In general, our results indicate that explicit boundaries are more informative, frequencies are important, and that distributional information is useful to the child as a source of categorial information. These results are in accordance with Redington et al.'s findings for English.

1 Introduction

Complementary to more standard methods of investigation in the field of language acquisition (such as manual corpora analysis and experimental studies), computational approaches aim to provide models that incorporate what is currently known about acquisition, language, and human cognition. In this way, they can be taken as psychologically plausible simulations that may throw light onto early aspects of language acquisition which are otherwise empirically difficult to observe. In the study described below, we developed a computational model to address the problem of learning the syntactic categories of words during language acquisition through the distributional analysis of utterances. In the present approach, this problem may be seen as a more specific instance of the general problem of finding associations between words through distributional analysis (Turney and Pantel, 2010; Lenci, 2018).

Although it is primarily meant to inform language acquisition theories, we expect that the present work may be of relevance for the general task of categorizing and grouping words through the use of distributional information. Particularly, as we apply the method to Brazilian Portuguese (BP) input data, it may help comprehending cross-linguistic differences between languages, which is a central goal of language acquisition theories and also an important one for the development of NLP techniques. Given that BP has a relatively fixed word order, we expect distributional information to have an important role in signaling the syntactic category of words.

Our model is a (local) reimplementation of the distributional learner described in Redington et al. (1998).[1] We present preliminary results from two experiments, originally, experiments 5 and 6 of the nine experiments carried out in Redington et al.'s study. We decided to reimplement their algorithm as both a way of achieving a deeper understanding of their method and also to assess its replicability, given the description found in their paper. Although being relatively old, Redington et al.'s study was chosen for being – to our knowledge – the first and most comprehensive computational study on the distributional properties of child directed speech. It investigates many aspects of the problem, such as the effects of distinct context windows, corpus sizes, number of target and context words, etc. In this sense, the present work contribution is very specific: aside from attesting the replicability of Redington et al.'s study, it also shows that distributional information is useful for

[1] The source code of the present model will be available at https://gitlab.com/pablofaria/dlearner.

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 152–159
Minneapolis, USA, June 7, 2019. ©2019 Association for Computational Linguistics

a child learning BP, a picture that will become fully clear as we publish results of the remaining experiments.

The paper is organized as follows. We first situate the present study regarding the field of language acquisition (section 2). Next, the corpus used and its preparation are described, along with a presentation of the distributional learner implemented (section 3). In section 4, we describe the two experiments and conduct a discussion on their quantitative and some qualitative results, focusing on a comparison with Redington et al. (1998). Final remarks come in section 5.

2 Language Acquisition and the Role of the Input

As a natural part of a typical human child development, learning a language - whether oral or gestual - emerges as a spontaneous, effortless, rapid, and ultimately successful process. In the field of language acquisition studies, theorists diverge on the actual explanations for this phenomenon, some arguing for mainly inductive processes based on qualities of the linguistic experience the child is exposed to and general cognitive capabilities (Tomasello, 1995; Pullum, 1996, and others), while other theorists minimize the role of the input, arguing that a specialized biological basis is necessary for language to be acquired (Yang, 2002; Berwick et al., 2011, and others). As one can see, at the core of such debate is the need for precise and exhaustive investigations of the informativeness of the input the child receives. Surprisingly, comprehensive computational and corpora studies are still restricted and scarce. For instance, although there are many studies about distributional properties of words in the literature (Clark, 2003; Turney and Pantel, 2010; Lenci, 2018, for instance), the study presented here is the first to our knowledge to investigate the distributional properties of a language other than English, *in the context of computational modelings of language acquisition*.

Acting on this gap, our study investigates the informativeness of distributional information to the task of syntactically categorizing words, also termed part-of-speech learning. As Harris (1954) points out, the "distribution" of an element can be described as "the sum of all its environments", where by "environment" Harris means an array of co-occurring elements and their positions in respect to a given (target) word. There are plenty of evidence showing that not only a distributional structure exists in language data, but also that speakers are sensitive to it (Brown, 1957; Landau and Gleitman, 1985; Bernal et al., 2007, to cite some). Although distributional information is broadly known to be insufficient for correctly categorizing words, it is important to investigate how much information it can contribute to the success of this task and that is precisely what the experiments shown below help understand.

Finally, we would like to emphasize that the problem dealt with here is similar but not the same as the problem of finding (semantic) associations between words, as seen in the long tradition of distributed semantic models (DSMs) developed in the last 30 years (Turney and Pantel, 2010; Lenci, 2018). For instance, here it is fundamental that the model categorizes function words correctly, while in DSMs they are in general left aside. Certainly, syntactically categorizing words involves, in part, detecting semantic associations between them. However, in order to detect the abstract syntactic nature of words we need to move beyond purely semantic association to find out what level of similarity allows us to cluster words together that behave syntactically the same. This is not a simple task and, of course, distributional information is surely not sufficient for fully solving the problem, in particular, because syntactic categories may differ substantially in their distributional properties and in their number of elements. For this reason, we expect to find many overlappings between our study and DSMs in general, without nonetheless taking into account important distinctions between these related tasks.

3 Methodology

For simulations, it was necessary to prepare a corpus of child directed speech (CDS) in Brazilian Portuguese, partially obtained from the CHILDES Database (MacWhinney, 2000) and partially obtained from the "Projeto de Aquisição da Linguagem Oral"[2]. The preprocessing of this material included the removal of metadata, children's utterances, and all kinds of annotation and commentaries made by those who built these corpora. There was also the need for a normal-

[2] Available online (in Portuguese) for visualization at `https://bit.ly/2sx0KBi`. Last accessed on January 17th, 2019.

ization of the orthography of transcriptions (e.g., "nene/baby" to "nenê"), specially for the second corpus mentioned above. It was carried out in a semi-automatic way in order to cover the most recurrent cases. No lemmatization was carried out.

Besides speech data, it is also necessary a "benchmark classification" against which the performance of the learner is evaluated. For this, we use the tagged version of the Tycho Brahe Corpus of Historical Portuguese (TBC)[3], consisting of part-of-speech annotated text from various authors and centuries. For some uncovered target words in the experiments, we manually assigned their most common tag for all non-ambiguous cases, such as proper nouns and diminutive forms of nouns (e.g., "menininho" which means "little boy"). Ambiguous and other idiosyncratic forms were left unclassified. In general, we basically followed the procedures found in Redington et al. (1998).

It is worth mentioning a distinction between English and Portuguese which posed a methodological and conceptual problem not faced in Redington et al. (1998). In Portuguese, nouns can be inflected in many ways, such as diminutive, augmentative, for grammatical gender, and so on. We first thought that all inflected forms could be replaced by a default form, in all cases where there is no change in the class of the word. However, there are inflected forms that exhibits specialized meanings, such as "calcinha" (literally "small pants") which means (woman) underwear. Thus, inflected forms were kept in the corpus for the model must reflect the ability of the child to learn both the regular behavior of inflected forms and also the exceptions (when distributively distinct). Furthermore, we aim to model the lexical acquisition process from its first steps, when morphological decomposition of words is not yet available.

Finally, punctuation is treated as in the original study: all intermediary punctuation is removed and all final punctuations (where present) are replaced by single end points. After all these procedures, our CDS corpus comprises approximately 1.4 million tokens, including punctuation. In Redington et al.'s study, they used a corpus of 2.5 million tokens.

3.1 The Distributional Learner

Our method is a local implementation of Redington et al.'s (1998) learner. Therefore, only a very brief description of the method is presented here. The learner goes through three stages in accomplishing the task: (i) measuring the distributional contexts for each target word; (ii) comparing distributional contexts for pairs of words; and (iii) grouping words based on distributional context similarity. The first stage produces a *contingency table* (a co-occurrence matrix) in which each line represents a context vector for a given target word. Each column corresponds to a context word in a particular position in respect to the word. Thus, if only the preceding word is used as context and 150 contextual words are considered, the vector will be of size 150. If two contextual positions are considered, then the vector will be of size 300, and so on.

Once the table is built, the second stage generates similarity measures for all possible pairs of target words. Although cosine similarity is currently a standard for comparing word vectors (Turney and Pantel, 2010; Lenci, 2018), for replication purposes, we use the Spearman rank correlation coefficient, ρ, which Redington et al. argue as the most successful measure in their study.[4] In the last stage, target words must be grouped together. This is carried out using a standard hierarchical cluster analysis, known as average link clustering. Once the hierarchy is produced – which can be represented as a dendrogram – the method identifies the optimum cut level which maximizes the performance of the learner in classifying words relative to the "benchmark classification" provided by the tagged corpus.

In order to demonstrate the relevance of the distributional information, that is, that the method produces results above chance classification, a "baseline classification" is calculated for each cut level analyzed. It goes as follows: for each cut level, the number of clusters obtained is kept constant but words are randomly distributed across these clusters and then performance is calculated. This is done ten times and the baseline derived for

[3] Available at http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/pos.zip. Last accessed on January 17th, 2019. The choice for TBC over other available corpora (such as Universal Dependencies) was for mere convenience (easy of access). The fact that it is historical data is not to be seen as a problem, given that we are targeting the most frequent words in our study, for which it is hardly the case that there was any historical change in their syntactic category. Nonetheless, ideally we would like to annotate the CDS data itself and use it as the gold standard for generated clusters.

[4] Of course, it leaves opened the question of whether cosine similarity would improve the model's performance, something we will address in the near future.

that cut level is the mean performance obtained.

3.2 Benchmark Classification

As a result of choosing the TBC as the tagged corpus of reference and in order to use the same categories assumed in the original study, a conversion between the two systems of classification was necessary. We have basically stripped off subtags from the TBC and established equivalence relations between the resultant tag system and Redington et al.'s classes. Table 1 summarizes the conversion schema.

3.3 Measuring Performance

The performance of the learner is evaluated through three measures, here applied across categories.[5] The first two are the traditional *precision* and *recall* measures. A third integrated measure is necessary in order to balance these two. In Redington et al. (1998), a measure called *informativeness* is proposed along with its justification. Although following the description given by the authors, we were still unable to obtain a satisfactory measure[6], reason why we decided to use the traditional F-measure, combined with a $\beta = 0.3$ coefficient to favor precision over recall. This (still tentative) option seemed in our simulations to compensate for the unbalanced nature of grammatical categories, in the sense that some are *open-ended*, that is, might in principle cover an unlimited number of elements, while others, such as "article" or "preposition", are "closed classes" with a fixed (and often small) number of elements. This fact tends to favor the recall measure over precision, because less clusters covering the largest categories will compensate for lower precision, something we would like to avoid.

4 Results and Discussion

In their original study, Redington et al. (1998) conduct nine experiments. From these, the authors established a "standard analysis", used as a reference in the analysis of other experimental conditions. Our standard analysis here follows the same settings: the 1000 most frequent words were used as target words for categorization, along with the 150 most frequent words as (relevant) contextual words. The context window included both the two immediately preceding and the two immediately succeeding words. Thus, each context vector consisted of 600 elements – four contextual positions for 150 words – each consisting of the frequency of a given context word in a specific position regarding the target. All final punctuations are removed and the data is treated as single long utterance.

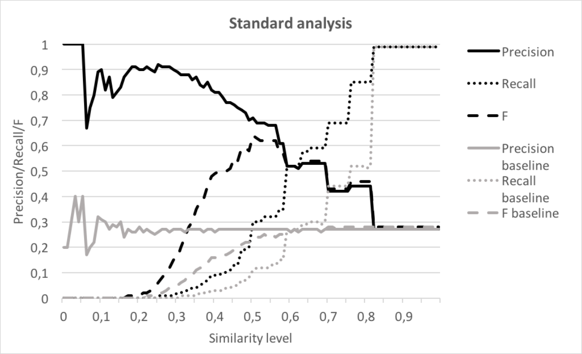

Figure 1: Performance of the learner for the standard analysis. For a similarity level of 0.5 (cut level), 25 clusters are obtained, with $F = 0.64$ (prec. = 0.71, recall = 0.30).

Figure 1 shows that the learner's performance is significantly above the baseline. As expected, categorization is much easier for the open-ended categories, specially nouns and verbs, with some clusters coming close to be "pure" (e.g., a cluster of infinitival verbs). For other categories, however, clusters tend to be mixed and more sensitive to syntactic function than to morphosyntactic properties. Thus, one of the clusters seems to capture the distribution of elements that may appear as heads[7] of noun phrases (articles, adjectives, nouns, pronouns, etc.), while another includes elements that appear in a predicative context, such as Y in "X is Y". An interesting feature observed is that many pairs of elements that vary only in gender, such as "do/da" ("of the", masculine and feminine), were very close to each other. This is an indication that distributional information can be of much help for the child to extract the grammatical gender feature

[5] One specific experiment, not reported here, assess performances for each category.

[6] Our implementation of this measure for some reason produced useless (i.e., non-discriminating) values for finding the best cut level for dendrograms. We are pretty sure it is our misunderstanding of it.

[7] For instance, in Portuguese one may say "o do Pedro" ("the of Peter"), with "o" playing the role of the head of the noun phrase. Something even more complex happens in "o vermelho do Pedro" ("the red of Peter"), where "do Pedro" can be seen as the modifier of "o", of "vermelho", or of both. The common property here is the absence of the noun itself, which impacts the distributional categorization of words. Of course, the actual syntactic analysis of such phrases will depend on the theory assumed.

Category	TBC tags	Examples	n
Noun	N, NPR	ademir, adriana, ajuda/*help*	375
Adjective	ADJ, OUTRO	alto/*tall*, amarelo/*yellow*, baixo/*low*	82
Numeral	NUM	cinco/*five*, dez/*ten*, duas/*two*	14
Verb	VB, HV, ET, TR, SR	abre/*opens*, abrir/*to open*, abriu/*opened*	331
Article	D	a/*the*, aquele/*that*, os/*the*	45
Pronoun	CL, SE, DEM, PRO, PRO$, SENAO, QUE, WADV, WPRO, WD, WPRO$, WQ	aonde/*whither*, aquilo/*that*, cadê/*where*	53
Adverb	ADV, Q, NEG, FP	agora/*now*, ainda/*still*, algum/*any*	62
Preposition	P	até/*until*, com/*with*, de/*of*	11
Conjunction	CONJ, CONJS, C	como/*how*, e/*and*, enquanto/*while*	11
Interjection	INTJ	ah, ahn, ai	16

Table 1: Categories, examples, and quantities for the 1000 most frequent words of the CDS corpus.

in the acquisition of Portuguese as well as for other similar alternations such as diminutive forms, plurals, etc. This is only a summary of some core aspects of a qualitative analysis of the clusters and categorizations obtained.

4.1 Utterance Boundaries: Testing Different Assumptions

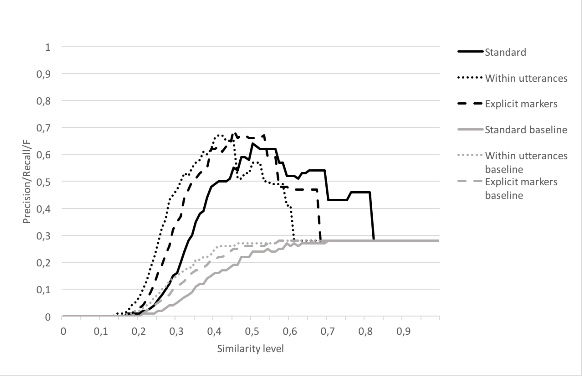

Figure 2: Comparison of performances (F-scores) when different assumptions about utterance boundaries are evaluated. Baselines for each are also shown.

In the standard analysis, all data is treated as a single long utterance, with punctuation marks removed. This is unrealistic, of course, because a child is surely sensitive to the beginning and end of utterances as well as to interruptions in speech, alternation of speakers, and so on. Thus, in order to investigate this issue, Redington et al. (1998) designed two specific conditions. First, utterances were taken one at a time, with contextual information limited to the boundaries of each utterance ("within utterance only"). This seems more realistic, although one-word utterances, being "contextless", become useless for the method. A second condition tests whether the addition of explicit boundary markers (i.e., final punctuation marks) helps the learner. In this case, punctuation marks are expressing the speaker's sensitiveness to phonological properties of speech, such as phonological phrase or utterance boundaries.

Figure 2 shows the distinct performances obtained for each condition. In general, curves are alike, although both alternative conditions have their peaks on a lower level of similarity. More specifically, in condition "within utterances only", the best $F = 0.67$ is obtained for the cut level 0.41, producing 17 clusters with precision 0.7 and recall 0.48. Recall is substantially higher (60%) than in the standard analysis, while keeping basically the same precision. Furthermore, the number of clusters decrease to 17, which is much closer to the benchmark. Considering only these general results, it seems that utterances boundaries benefit the learner. If this is indeed the case, explicit markers should help even more and that is what the condition "explicit markers" shows.

As we can see, its $F = 0.69$ is the highest obtained so far. Although recall decreases a bit, to 0.44, precision increases to 0.72, and the number of clusters is 18 (for cut level 0.45). The main difference between this and the previous condition is the use of one-word utterances, which now has an explicit boundary marker functioning as a minimum contextual information. Given these results, we can more confidently claim that utterance boundary information indeed helps the learner. This is, of course, compatible with what language acquisition theory says, specially the advocates of the important role phonology plays in the acquisition by helping the child segment the speech stream (Christophe et al., 2008, for instance).

4.2 Context Words: Attesting Occurrence Instead of Frequency

In this experiment, the goal is to observe how the learner behaves when, instead of the frequency of each context word, only the occurrence (or not) of it is recorded. Although children do extract statistics from input data (Romberg and Saffran, 2010), it may be the case that the actual learning procedure is *in between* mere occurrence and precise statistics about context words. This experiment allows us to explore this radical alternative learning strategy, see how it plays out, and hopefully learn something from it. In order to do that, after collecting statistics about context words, all context vector values greater than zero are converted to 1. And, because rank correlation is not well suited to binary vectors, following Redington et al., the "cityblock" metric is used in the "Occurrence" condition. A third condition, "Cityblock", uses frequencies *and* the "cityblock" metric, allowing for a better comparison with the standard analysis.

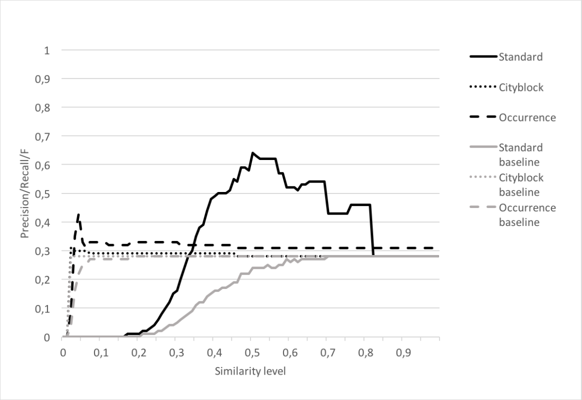

Figure 3: Comparison of performances (F-scores) for different ways of counting context words (i.e., frequency or binary context vectors). Baselines for each are also shown.

As one can see in Figure 3, the learner's performance drops significantly for both "Cityblock" and "Occurrence" conditions, when compared to the standard analysis. The "Cityblock" condition, with $F = 0.3$, precision 0.29, recall 0.91, and 9 clusters, demonstrates the inappropriateness of the "cityblock" metric when actual frequencies are taken into account, as Redington et al. (1998) point out. Its very low precision shows that it poorly categorizes words, basically creating large clusters, which explains its high recall. Instead, when binary vectors are used with the cityblock metric, performance increases, as the "Occurrence" condition shows. It obtains $F = 0.43$, with precision 0.53, recall 0.14, and 48 clusters. While still being a low performance, it demonstrates some ability to categorize (precision 0.53), although its high number of clusters prevents it from reaching a good recall. A possible interpretation is that it performs better in recognizing differences among categories than similarities between elements of the same category. Finally, these results, in general, indicate that some tracking of frequencies of contextual elements is necessary for the learner to extract the full potential of distributional information.

4.3 Related Work

In their original study, Redington et al. (1998) evaluate the effects of different assumptions about utterance boundaries. In Mintz et al. (2002), this aspect is also investigated, but they move a step further to investigate the effects of intrasentential boundaries. This is a study we plan to conduct in the near future. When we consider Redington et al.'s results on this issue (p.457-458), we find the same tendency observed in our experiment. Both conditions, within utterance and explicit markers, help improve the learner's performance, with the latter producing the best performance overall. As Redington et al. point out, "information recorded across utterance boundaries effectively act as noise."

In our second experiment, we have found that collecting actual frequencies of contextual elements improves the learner's performance. In Redington et al.'s study (p.458-459), results show similar tendencies, but with some key differences worth emphasizing. First, in the "Cityblock" condition, in which our learner performs very poorly, their learner performs quite well, although worse than for the standard analysis. This opposite behavior is puzzling to us and we cannot find reasonable explanations for it, apart from some unnoticed technical misunderstanding in our replication of their study or, in part, due to the distinct performance measures applied in each study. For the "Occurrence" condition, however, although it performs second in our study, both here and there we observe a significant decline in performance and a very small advantage of the method over the random baseline.

In general, a precise comparison of these studies is not totally straightforward. First, as already pointed out, because each uses its specific performance measure. In the future, we expect to overcome this limitation through appropriate implementations of the measures used in Redington et al. (1998) and Mintz et al. (2002). Furthermore, with these in hand, we will be able to compare measures and try to understand whether they are complementary or substitutes. Second, and more subtle, are the way values for similarity are obtained. We cannot claim our method produces equivalent similarity values, particularly in the sense that, in our study, similarity values do not generalize across experiments, as they appear to do in Redington et al.'s study. Consequently, they are able to consider a cut level of 0.8 as an "optimum" cut level for all experiments, while this is not possible in our study. We are also working on this issue.

5 Conclusions

In this paper, distributional properties of Brazilian Portuguese are investigated through the replication of the study in Redington et al. (1998). Two aspects were analyzed here: the effects on performance of different assumptions about utterance boundaries and the effects of distinct learning strategies regarding the use of statistical information about contextual items. Our results tend to support the original study, although we have pointed out some differences that deserve more investigation. In sum, results support the claims that distributional information is informative to the task of learning word categories, that explicit utterance boundaries help the learner in this task, and that frequency of contextual elements, instead of merely attesting their occurrence, is necessary in order to extract the full potential of this source of information.

Many issues remain open for future work. Some are already under investigation, such as the remaining experiments in Redington et al. (1998), the first of them (evaluation of different context windows) reported in Faria and Ohashi (to appear). A central goal of ours is to provide a more in-depth comparison between English and Brazilian Portuguese regarding the role of distributional information, specially in terms of how morphological and word ordering differences between these two languages affect category identification. Aside from completing the set of experiments, we will also expand it by evaluating the suitability and plausibility of more recent models (Baroni and Lenci, 2010; Mikolov et al., 2013; Pennington et al., 2014) to this task. In addition, other relevant factors must also be studied, as indicated in Turney and Pantel (2010) and in Lenci (2018), such as using cosine and other vector similarity measures, as well as trying mathematical techniques to deal with lower frequencies and noise, weighting, sparsity, and optimizations.[8] Given that BP has rich morphology, exploring also how such information may help the learner, as in (Clark, 2003), is also something in our sight.

Finally, it is important to note that although the present study strongly relates with DSMs and all its literature, the distributional learning of syntactic categories is approached here as part of the language acquisition process of a child learning her native language. Consequently, matters of psychological, developmental, and empirical plausibility strongly applies to the computer model which aims to increasingly approximate what we observe in real life. Moving towards a gradual presentation of input data, for instance, is a condition for psychological plausibility we aim to meet in the future and which may be in conflict with other DSMs found in the literature, primarily conceived for massive NLP tasks with manipulation of the whole set of data. Nonetheless, assessing the suitability of the various models is the kind of question we hope to be able to answer as our research moves forward.

Acknowledgments

We would like to thank the reviewers for their thoughtful feedbacks on the present study. Their comments and suggestions not only helped improving the present paper, but also indicated interesting directions for future work.

[8] All reviewers stressed the fact that these are important issues to explore, not only to better understand the phenomenon *per se* but also as a way approximating the state-of-the-art in this topic. We have good reasons to expect that as we gather the results of the full collection of experiments while expanding to new ones, we will be able to provide some interesting reflections on these.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Savita Bernal, Jeffrey Lidz, Séverine Millotte, and Anne Christophe. 2007. Syntax constrains the acquisition of verb meaning. *Language Learning and Development*, 3:325–341.

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science*, 35:1207–1242.

Roger W. Brown. 1957. Linguistic determinism and the part of speech. *Journal of Abnormal & Social Psychology*, 55(1):1–5.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51(1-2):61–75.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pablo Faria and Giulia Osaka Ohashi. to appear. A aprendizagem distribucional no português brasileiro: um estudo computacional. *Revista LinguíStica*, 14(3).

Zellig Sabbetai Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Barbara Landau and Lila R. Gleitman. 1985. *Language and experience: evidence from the blind child*. Harvard University Press, Cambridge, MA.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annu. Rev. Linguist.*, 4:151–171.

B MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk*, third edition edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Toben H. Mintz, Elissa L. Newport, and Thomas G. Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26:393–424.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

Geoffrey K. Pullum. 1996. Learnability, hyperlearning, and the poverty of the stimulus. In *Proceedings of the Twenty-Second Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on The Role of Learnability in Grammatical Theory*, pages 498–513. Berkeley, California: Berkeley Linguistics Society.

Martin Redington, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Alexa R. Romberg and Jenny R. Saffran. 2010. Statistical learning and language acquisition. *Wiley interdisciplinary reviews. Cognitive science*, 1(6):906–914.

Michael Tomasello. 1995. Language is not an instinct. *Cognitive Development*, (10):131–156.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Charles Yang. 2002. *Knowledge and learning in natural language*. Oxford University Press.