diacrítica
REVISTA DO CENTRO DE ESTUDOS HUMANÍSTICOS
Vol. 33, n.º 2, 2019, p. 229-251. DOI: doi.org/10.21814/diacritica.415

LEARNING PARTS-OF-SPEECH THROUGH DISTRIBUTIONAL ANALYSIS. FURTHER RESULTS FROM BRAZILIAN PORTUGUESE

APRENDIZAGEM DE CATEGORIAS DE PALAVRAS POR ANÁLISE DISTRIBUCIONAL
RESULTADOS ADICIONAIS PARA PORTUGUÊS BRASILEIRO

Pablo Faria
pablofaria@iel.unicamp.br

A child learning a language has to figure out what the syntactic, or part-of-speech, categories in her language are and assign words to one or more of them. The question we aim to answer here is how much of this learning can be accomplished through the distributional analysis of utterances. To this end, a reimplementation of Redington, Chater and Finch (1998) computational model was conducted and applied to Brazilian Portuguese input data, obtained from publicly available corpora of both child-directed and adult-to-adult speech. Results from all experiments are presented and discussed. These experiments investigate many variables and aspects involved in this learning task: types of distributional contexts, the number of target and context words, the value of distributional information for different categories, corpus size, etc. A comparison between child-directed speech and adult-to-adult speech is also carried out. In general, our results support Redington *et al.*'s (1998), although we find some possibly important, and maybe contradictory, differences. We also evaluate the cosine metric, comparing it with performance obtained with the Spearman rank correlation metric used in Redington *et al.*'s (1998) study. The latter seems to produce better performance. In this paper we focus on a quantitative analysis of our results.

Keywords: Language acquisition. Part-of-speech learning. Distributional analysis. Cognitive modelling.

Uma criança adquirindo a língua deve descobrir quais são as categorias sintáticas em sua língua e atribuir palavras a uma ou mais delas. A questão que nos propomos a responder aqui é o quanto dessa aprendizagem pode ser realizada através da análise distribucional de enunciados. Para este fim, uma re-implementação do modelo computacional de Redington, Chater e Finch (1998) foi conduzida e aplicada a dados do Português Brasileiro, obtidos de corpora disponíveis publicamente, tanto com fala dirigida à criança, quanto com fala entre adultos. Os resultados de todos os experimentos são apresentados e discutidos. Estes experimentos investigam mais variáveis e aspectos envolvidos nesta tarefa de aprendizagem: tipos de contextos distribucionais, o número de palavras-alvo e de contexto assumidas, o valor da informação distribucional para as diferentes categorias, tamanho do corpus etc. Uma comparação entre a fala dirigida à criança e a fala entre adultos também é feita. Em geral, nossos resultados dão suporte aos de Redington *et al.* (1998), embora tenhamos encontrado algumas diferenças possivelmente importantes e até contraditórias. Também avaliamos a medida *cosseno*, comparando a performance obtida com ela à performance obtida com a correlação de Spearman usada no estudo de Redington *et al.* (1998).

Esta última parece produzir melhor performance. Neste artigo, focamos numa análise quantitativa dos nossos resultados.

Palavras-chave: Aquisição da linguagem. Aprendizagem de categorias. Análise distribucional. Modelagem Cognitiva.

•

1. Introduction

A specific question is investigated here: how much information is a child able to extract from the input she hears through the distributional analysis of utterances with no recourse to subword morphology, semantics, and other possible sources of information? To this end, we present a reimplementation of the distributional learner described in Redington, Chater and Finch (1998).[1] In their work, Redington and colleagues studied many aspects involved in learning part-of-speech categories from English input data, with the goal of casting light over the language acquisition process. Aiming at investigating the same subject and on producing cross-linguistic results, we decided to start with Redington *et al.*'s (1998)[2] model, applying it to Brazilian Portuguese (BP) data. At the moment, a full replication of all its nine experiments are in place, allowing us to reach a first general overview of this matter, which we present in the following sections.

Before moving forward, however, it is important to justify the choice to reimplement Redington *et al.*'s (1998) method. First, a reimplementation is an effective way of achieving a deeper understanding of a model, for in this process we have to go back and forth from the paper to the implementation in order to find cues about details not so easily identifiable in the paper. In doing this, we also assess how clear and complete is the model presented in the paper. Given the increasing number of modeling studies being conducted nowadays, it becomes more and more important to be able to assess their replicability, a goal for which this paper makes a humble contribution. Finally, although related works have appeared before and many did since then, Redington *et al.*'s (1998) study is – to our knowledge (Frank 2011; Kaplan, Oudeyer & Bergen 2008; Seidenberg 1997; Wintner 2010; Yang 2012) and *in this specific subject* – the first and most comprehensive computational study on the distributional properties of child directed speech and how it relates to language acquisition. In this regard, this model is a computational cognitive model. In addition, this study is connected to the general problem of finding associations between words through distributional analysis (Lenci 2018; Turney & Pantel 2010).

Computational cognitive modeling, as an area of research, aims to develop models that incorporate what we understand about learning, language, and human cognition. As psychologically plausible simulations, then, models may cast light onto early aspects of language acquisition, which are otherwise empirically difficult – if not impossible – to

[1] The source code of our model (v1.0) is publicly available at <https://gitlab.com/pablofaria/dlearner>.
[2] Given that this is the only work of Redington and colleagues considered here, from now on it will be referred to simply as "Redington *et al.*".

investigate. Consequently, models are effective tools to inform learning theories, thus helping improve their reach, and make their claims more precise and consistent. By applying the method to BP, this study highlights some distributional properties of BP, which we can use to discuss its commonalities and differences regarding English. Such cross-linguistic understanding is a central goal of language acquisition theories and is important also for developing NLP techniques.

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 start by summarizing results from experiments 1, 5, and 6, presented in Faria and Ohashi (2018) and Faria (2019). Next, we introduce results from remaining experiments and conduct a quantitative discussion, focusing on a comparison with Redington *et al.*'s (1998) results. In section 5, final remarks are made about the present study, pointing to qualitative aspects and plausibility matters to be fully discussed in a future work.

2. Language acquisition and distributional part-of-speech learning

As a natural part of a typical human child development, learning a language – whether oral or gestural – 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 on general cognitive capabilities (Pullum 1996; Tomasello 1995; and others), while other theorists restrict the role of the input, arguing for a specialized biological endowment as necessary for language to be acquired (Berwick, Pietroski, Yankama & Chomsky 2011; Yang 2002; and others). At the core of such debate we see the need for precise and exhaustive investigations on the informativeness of the input the child receives. Unfortunately, comprehensive computational and corpora studies with the goal of modelling language acquisition are crosslinguistically restricted and scarce, even though there are many studies about distributional properties of words in the computational linguistics literature (see, for instance, Clark 2003; Lenci 2018; Turney & Pantel 2010).

Acting on this gap, our study investigates the informativeness of distributional information to the task of syntactically categorizing words of BP, 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 (Bernal, Lidz, Millote & Christophe 2007; Brown 1957; Landau & Gleitman 1985 to cite some). Consequently, although distributional information is broadly known to be insufficient for correctly categorizing all 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 (Lenci 2018; Turney & Pantel 2010). Although we expect to find overlaps between this study and DSMs in general, one must nonetheless take into account distinctions between these related tasks. For instance, for part-of-speech learning it is fundamental that functional words are categorized properly, while in DSMs they are in general left aside. Certainly, syntactically categorizing words involves, in part, detecting semantic associations between them, a task for which distributional information is already proved to be useful. However, relations between words are not merely semantic and, in order to detect syntactic relations, we need to find out how distributional information helps us to cluster words that behave syntactically the same together. The present study is our first general approximation of this problem, given that, among other limitations, it does not process subword morphological information. Of course, distributional information is surely not sufficient for fully solving the learning problem at stake. Nonetheless, it is important to understand how, how much, and in which conditions, is distributional information useful here.

3. Methodology

Besides developing the computational learner itself, the present study depends on data for its implementation. Two BP corpora were assembled for our study: a corpus of child directed speech (CDS), partially obtained from the CHILDES Database (MacWhinney 2000) and partially obtained from CEDAE/UNICAMP[3], and an adult-to-adult speech corpus, obtained from "Projeto Norma Linguística Urbana Culta – RJ".[4] Having to deal with three distinct schemes of transcription and annotation, the preprocessing of this material included removal of metadata, children's utterances, and all kinds of identifiable annotation and comments. We also needed to normalize orthography (*e.g.*, "nene/baby"[5] to "nenê"), specially for the CEDAE/UNICAMP corpus. This was carried out in a semi-automatic way (through a manually specified conversion table) in order to cover the most recurrent cases. No lemmatization was carried out, which allows for a more direct comparison with the study for English.[6]

[3] "Centro de Documentação Cultural 'Alexandre Eulalio'". Collection "Projeto de Aquisição da Linguagem Oral", accessible at <http://www3.iel.unicamp.br/cedae/> Last accessed on 05-Nov-2019.
[4] Sections "Diálogos entre informante e documentador (DID)" and "Diálogo entre dois informantes (D2)". Available at: <http://www.nurcrj.letras.ufrj.br/> Last accessed on 05-Nov-2019.
[5] The content after "/" is the English meaning of the Portuguese word.
[6] A comment is necessary about using lemmas instead of inflected forms in this kind of study. As pointed out in the text, the original study does not use lemmas. It may be claimed that this choice needs supporting evidence, but we would argue, instead, that using lemmas would mean to assume the ability, by a child, to analyze words into roots and affixes, which is surely true for later stages but not for the initial ones. Given that the model is momentarily instantaneous with regard to the use of input data, using inflected forms is a kind of minimum assumption here, one that would make the learning task harder. Thus, if the method

In addition to speech data, it is also necessary a "benchmark classification" against which the performance of the learner is evaluated. The tagged version of the Tycho Brahe Corpus (TBC) (Galves, Andrade & Faria 2017), consisting of part-of-speech annotated text from various authors and centuries, was used. 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/little boy"). Ambiguous and other idiosyncratic forms were left unclassified. In general, we basically followed the procedures found in Redington *et al.*'s (1998) work.

It is worth mentioning a distinction between English and Portuguese which posed a methodological and conceptual problem not faced – or at least not acknowledged – by 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 (i.e., a lemma-like approach), in all cases where there was no change in the word category. However, some inflected forms exhibit specialized meanings, such as "calcinha" (literally "small pants") which means (woman) underwear. In these cases, even if belonging to the same categories, inflected variants may have significantly distinct distributions. For this reason, inflected forms were kept in the corpus and the model must reflect the child's ability to learn both the regular behavior of inflected forms and also exceptions (when distributively distinct). Furthermore, as mentioned earlier, this model of the lexical acquisition process abstracts away from morphological decomposition of words, as a child in her first steps into language.

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 comprised approximately 1.4 million tokens, including punctuation. In Redington *et al.*'s (1998) study, they used a corpus of 2.5 million tokens. As we shall see later (section 5.3), this difference in the amount of input data does not prevent the model of being directly comparable to Redington *et al.*'s (1998) nor of providing interesting insights about the problem under investigation.

3.1. The distributional learner

Our model is basically an effort to implement Redington *et al.*'s (1998) learner, by following their presentation. Therefore, only the core details of the model, necessary for the understanding of the model, are presented here. The learner must go through three stages in accomplishing the learning task: (i) measuring the distributional contexts for each target word; (ii) comparing distributional contexts for all possible pairs of words; then (iii) grouping words together based on distributional similarity. The first stage produces a *contingency table* (or 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 target word. Thus, if only the preceding word is used

proves, as it does, to be still effective for inflected forms, one would expect a significant improvement in its performance if lemmatized forms are explored, as well as semantics and other sources of information.

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. Cells in this matrix store the measurements obtained in the first stage.

Once the table is built, the second stage generates similarity measures for all possible pairs of target words, which implies comparing contextual vectors. Although cosine similarity is currently a standard for comparing word vectors (Lenci 2018; Turney & Pantel 2010), for replication purposes we use the *Spearman* rank correlation coefficient, ρ, which Redington *et al.* (1998) argue as the most successful measure in their study. Finally, in the third and last stage, target words are grouped together using a standard hierarchical cluster analysis, known as average link clustering. This is a recursive procedure that takes the momentarily two closest elements – whether words or clusters previously formed or a mixture of a cluster and a word – and form a new cluster. The clustering ends when a single cluster containing all others is obtained. The hierarchy produced can be represented as a dendrogram (see Figure 1 below). Finally, the learner's classification is extracted from the hierarchy by finding the optimum cut level (also shown in Figure 1) for which the clusters obtained are closer to the "benchmark classification" provided by the tagged corpus.

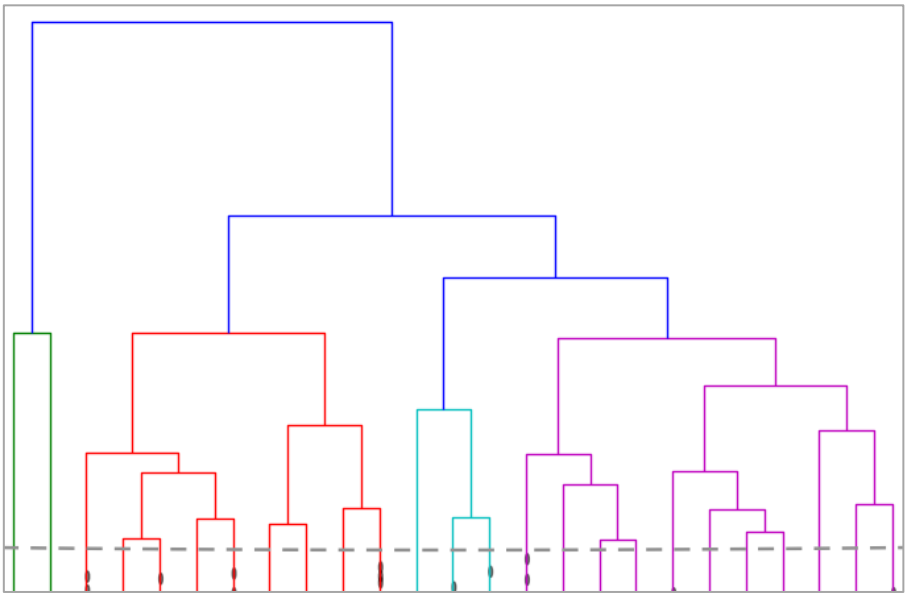

Figure 1. An example of a dendrogram with a cut level line (dashed gray) showing where clusters would be extracted.

3.2. Measuring performance

A difference worth mentioning about Redington *et al.*'s (1998) model and ours regards their use of a measure called *informativeness*. This measure is proposed as a way of balancing precision ("accuracy" in their paper) and recall ("completeness" in their paper). For some still unknown reason, we were unable to obtain a satisfactory implementation

of it.[7] Although there seems to be good reasons to believe our performance measures are relatively equivalent to theirs, given the general picture of results obtained, it is still important that we are able to implement this feature and have an even more strict replication and comparison.

Alternatively, we assess the learner's performance using the *F*-score measure, over the traditional measures of precision and recall, all applied across categories. To better understand these measures, let us see the learner's task of categorizing words as a task of determining, for any possible word pair derived from the list of target words, whether those two words are of the same category or not. Thus, when the learner guesses a pair (by joining two words in the same cluster), it is saying they have the same category. Otherwise, it must keep those words apart.

Given this view and the benchmark classification, *precision* measures how much of the learner's guessed word pairs (GP) are indeed correct (CP): CP/GP. Complementarily, *recall* measures how much of the desired pairs (DP) did the learner guess right (CP): CP/DP. The *F*-score measure is used to integrate these two according to the following general formula:

$$\mathrm{F}_{\beta} = (1 + \beta^2) * \frac{\text{precision} * \text{recall}}{(\beta^2 * \text{precision}) + \text{recall}}.$$

In our model, we use a $\beta = 0.3$ coefficient to favor precision over recall. This option seemed in our simulations to compensate for the unbalanced nature of grammatical categories, in the sense that some are *open-ended* (i.e., content words) and might, in principle, cover an unlimited number of mostly infrequent elements; on the other hand, functional categories such as "article" or "preposition" are *closed* classes, that is, they have a fixed (and often small) number of frequent elements. This is easily seen even for a small sample, as shown in Table 1 below. This unbalance tends to favor recall over precision, something we try to avoid by manually balancing the β coefficient.[8]

3.3. Benchmark and baseline classifications

The TBC has its own tagging system. Consequently, in order to use the same categories assumed in the original study, tags are converted from the TBC system to Redington *et al.*'s (1998) system. Basically, tags were stripped off of their subtags (*e.g.*, from "N-P" to "N") and then substituted by Redington *et al.*'s (1998) (*e.g.*, from "N" to "noun"), according to the schema presented in Table 1. As a benchmark classification, the words are divided in ten classes.

[7] Our implementation of this measure for some reason produced useless (i.e., non-discriminating) values for finding the best cut level for dendrograms. As soon as we understand why, we expect to be able to have it working and compare it with our *F* scores.

[8] The fact that we obtain this balance with a low recall is of course unsatisfactory. For this reason, this is a momentary resolution in need of further development. As pointed out by one of the reviewers, we will probably need to integrate other sources of information in order to obtain higher measures for both precision and recall. We hope to give some answer to this question with future investigations.

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

Category	BENCHMARK corpus tags	n
Noun	N, NPR	375
Adjective	ADJ, OUTRO	82
Numeral	NUM	14
Verb	VB, HV, ET, TR, SR	331
Article	D	45
Pronoun	CL, SE, DEM, PRO, PRO$, SENAO, QUE, WADV, WPRO, WPRO$, WD, WQ	53
Adverb	ADV, Q, NEG, FP	62
Preposition	P	11
Conjunction	CONJ, CONJS, C	11
Interjection	INTJ	16

In order to demonstrate the relevance of the distributional information, it is also important to show that the learner can perform above chance performance. Therefore, 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 that cut level is the mean performance obtained.

4. Quantitative results and discussion

In this section we focus on a quantitative analysis and on comparing our results with Redington *et al.*'s (1998). Following their study, nine experiments were designed – for reasons we discuss later in section 5. Each one focus on a particular aspect or variable of the learning task. The experiments are:

1. Different contexts
2. Varying the number of target and context words
3. For which classes is distributional information of value?
4. Corpus size
5. Utterance boundaries
6. Frequency versus occurrence
7. Removing function words
8. Does information about one category help the acquisition of the others?
9. Is learning easier with child-directed input?

For experiments 1, 5, and 6, results were presented and discussed in Faria & Ohashi (2018) and Faria (2019). Thus, we start by summarizing those results, before presenting results from the remaining experiments – 2, 3, 4, 7, 8, and 9 – along with a discussion about our findings, how they relate to the original study, and which questions are left opened. We also evaluated an additional condition in experiment 6, which casts doubt on the original study conclusions regarding this experiment.

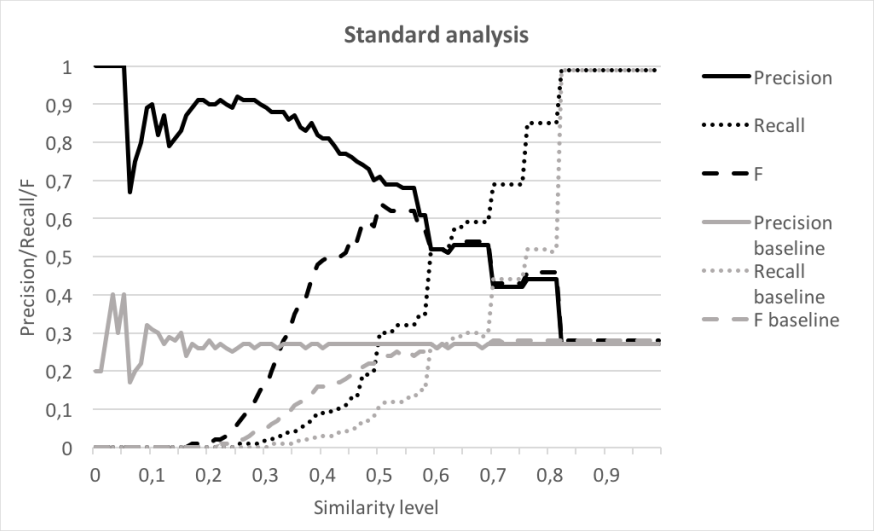

Figure 2. Performance of the learner for the standard analysis. With a cut level of 0.5, 25 clusters are obtained, with F=0.64 (prec. = 0.71, recall = 0.30).

Figure 2 above shows the learner's performance for the "standard analysis", used as a reference for the evaluation of other experimental conditions. This analysis uses the 1000 most frequent words as target words for categorization, along with the 150 most frequent words as (relevant) contextual words. The context window comprises the two immediately preceding and the two immediately succeeding words (i.e., w_{i-2} w_{i-1} w_i w_{i+1} w_{i+2}). 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. The entire CDS corpus is used.

4.1. Summary of experiments 1, 5, and 6

Experiment 1 (Faria & Ohashi 2018) was designed to evaluate the informativeness of various types of context relative to a given target word. It evaluates how the distance between a target word and a contextual item affects informativeness and also helps identifying which context maximizes learning. It starts by assessing the four contextual positions that follow target words. Then, the four preceding positions are evaluated. Finally, combinations of positions as "contextual windows" are analyzed. As Table 2 shows in its last line, our experiment indicates that the most informative context, in general, comprises two immediately preceding and one immediately following words, regarding a target word. It obtains the highest *F*-score (together with context [-1,1]), with the best balance between precision and recall and, importantly, a number of clusters closer to the benchmark. Note also that only the more local environment is informative: distant contextual items do not help the learner go beyond baseline performance ($F \sim 0{,}3$). These results are very similar to Redington *et al.*'s (1998) for English: highly local contexts are informative, contrary to less local ones. The preceding context is also more informative than the context following the target word. A small difference is that for English, the best context obtained included two preceding and two succeeding words.

Table 2. Learner's performance for various types of context.

Context	*F*	Precision	Recall	Cut	Clusters
[1]	0,47	0,49	0,34	0,35	24
[2]	0,32	0,32	0,28	0,54	12

[3]	0,30	0,29	0,68	0,64	3
[4]	0,31	0,30	0,60	0,39	5
[-1]	0,67	0,81	0,23	0,42	43
[-2]	0,49	0,56	0,21	0,71	26
[-3]	0,31	0,30	0,43	0,67	9
[-4]	0,32	0,30	0,79	0,49	4
[-2, -1]	0,61	0,75	0,20	0,54	42
[-1, 1]	0,68	0,79	0,27	0,48	33
[-1, 1, 2]	0,62	0,68	0,33	0,56	23
[-2, -1, 1]	0,68	0,72	0,40	0,54	14

In experiment 5, assumptions about utterance boundaries are investigated. As mentioned above, the "standard" assumption was to remove all final punctuations and treat the data as a single long utterance. Statistics are observed across boundaries. Of course, this is a simplifying assumption, because the child is sensitive to phonological properties characteristic of utterance boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss & Kennedy 1987; Seidl & Johnson 2006). Thus, two alternative assumptions are investigated in this experiment: one utterance at a time without final punctuation ("within utterance only") and one utterance at a time with final punctuation ("explicit markers"). In both cases, contextual information is limited to the boundaries. Figure 3 shows the results obtained (*cf.* Faria 2019). As expected, utterance boundaries improve learning. Furthermore, with explicit markers, the learner reaches *F=0.69* (prec.=0.72, recall=0.44, 18 clusters), demonstrating that the sensitivity of the child may be used as contextual information for learning word categories. As seen in Figure 4, a very similar pattern was verified for English (at the 0.7 level of similarity).

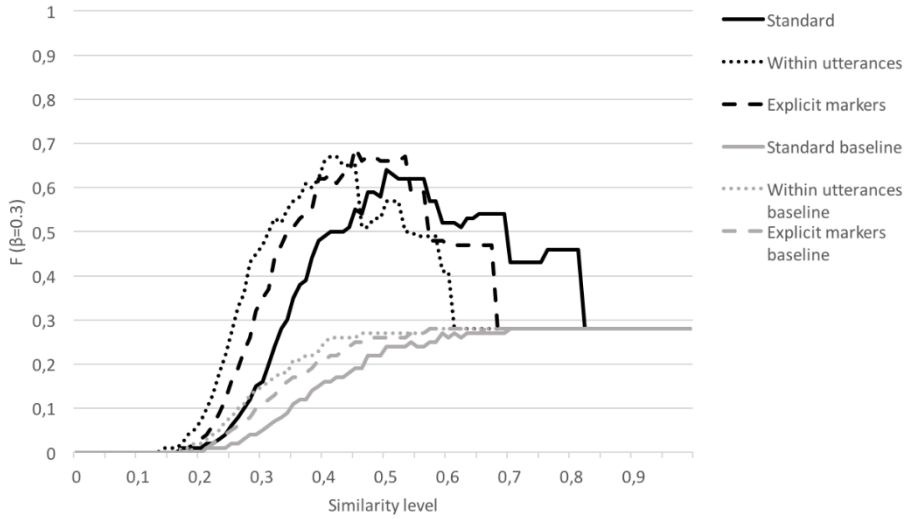

Figure 3. Learner's performances for three different assumptions regarding utterance boundaries.

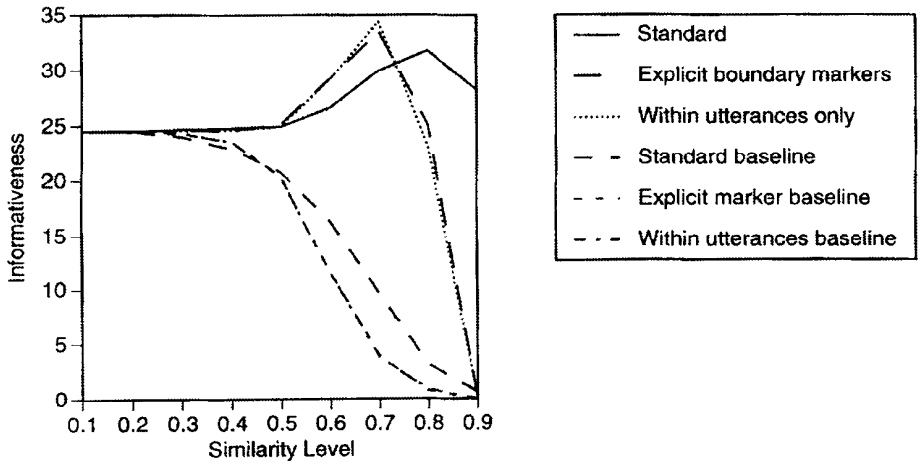

Figure 4. Results of experiment 5 in Redington *et al.* (1998:457).

In experiment 6 (Faria 2019), we evaluate whether frequencies are really necessary for the success in this task ("standard" condition) or whether the learner may succeed by only acknowledging the occurrence of a given contextual item regarding a target word, in a binary fashion ("occurrence" condition). Statistically, the difference is about having a word vector of frequencies or a binary word vector. Consequently, as argued by Redington *et al.* (1998), different vector similarity measures are necessary in each case. Thus, instead of using the Spearman rank correlation for binary vectors, they use the "cityblock" measure. However, these two conditions are now difficult to compare directly: they assume distinct vectors and distinct metrics. A third condition is then suggested by Redington *et al.* (1998) to mitigate this problem: the "cityblock" condition, where the cityblock metric is used with frequency vectors.

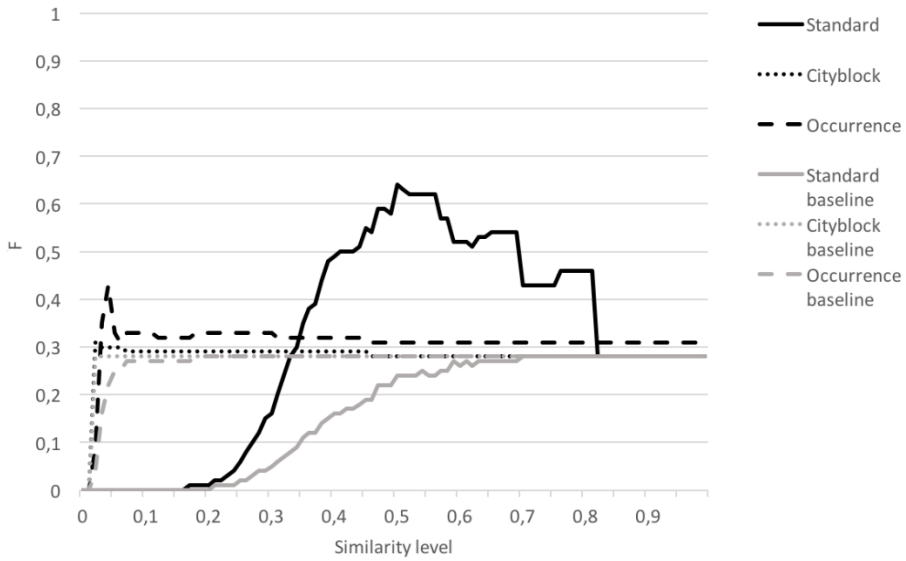

Figure 5. Learner's performance for frequency versus binary word vectors and two different metrics (Spearman vs Cityblock).

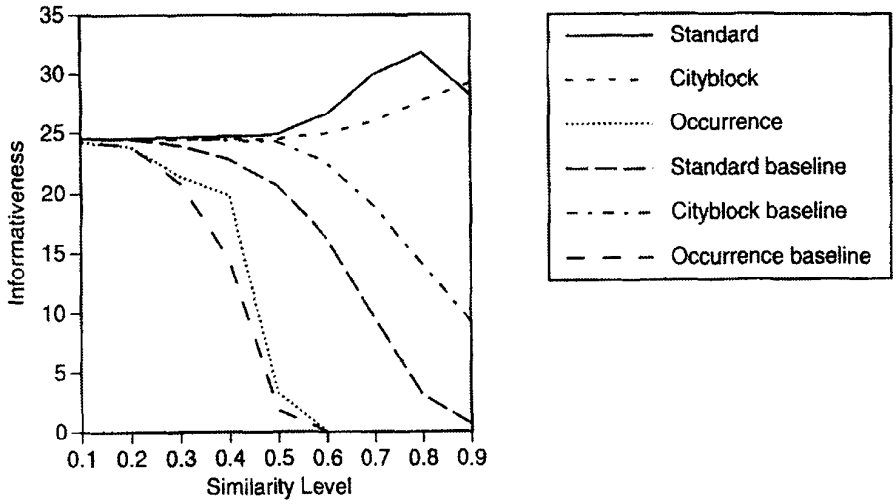

Figure 6. Results of experiment 6 in Redington *et al.* (1998, p. 459).

Given the picture in Figure 5, we must conclude that the cityblock metric is not suited to this task, for in both "occurrence" and "cityblock" conditions performance is basically baseline. This is partially in conflict with Redington *et al.*'s (1998) findings. There, the cityblock metric performed quite well with frequencies (see Figure 5), even though the authors claim it is best suited for binary vectors. The authors claim, additionally, that the Spearman correlation is not a good metric for binary vectors. In spite of their remarks, we decided to evaluate the *Spearman* rank correlation metric with binary vectors ("Occ+Sprm" condition). Results shown in Figure 7 contradict Redington *et al.*'s (1998) remarks: the Spearman correlation seems to be working just well with binary vectors and, surprisingly, obtains a higher *F* (0.69) then the standard analysis. Note also that in this condition the cluster differentiation area (i.e., the area below the curve) shifts back in the similarity axis and is less large, when compared to the standard line. This means that clusters are distinguished on shorter distances between words, which makes sense given the use of binary vectors. Most importantly, this result leaves us uncertain about what is really making a difference here: frequencies or the *Spearman* metric? This is a question we will try to answer with more investigation.

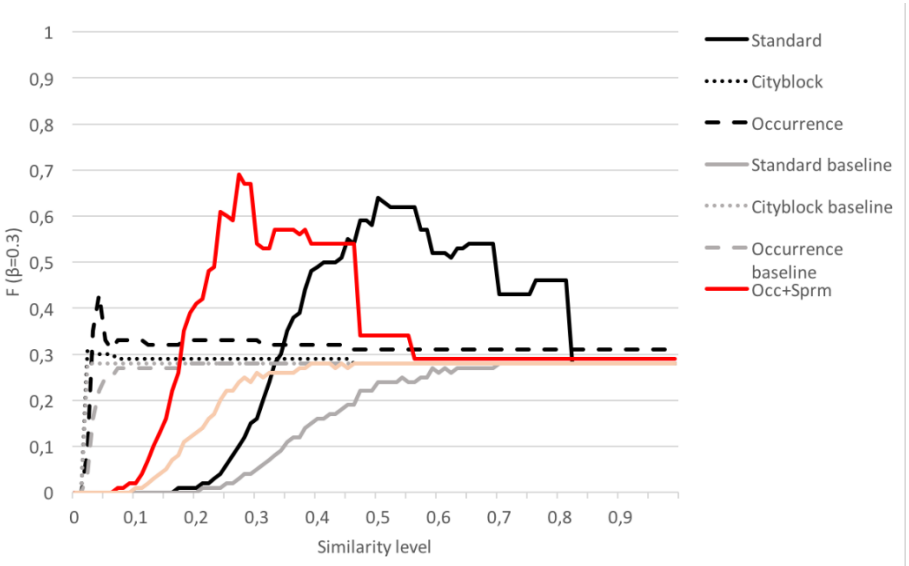

Figure 7. One additional condition in experiment 6: "Occ+Sprm", where binary vectors are compared using *Spearman* rank correlation.

4.2. Experiment 2: varying the number of target and context words

In this experiment, Redington *et al.* (1998) want to answer the following question: *What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?* The authors do not give much details about their findings, except for a general picture described by them as "an inverted U-shape": for low numbers of target words, performance is quite poor; then it increases until 1000 words and then gradually decreases towards 2000 words. They explain this behavior suggesting that, for low numbers of target items, words tend to be mostly functional and the learner does not perform well for these (as we see in experiment 3). If the target word set is small but diversified (as the 31 words in Kiss 1973 *apud* Redington *et al.* 1998), then the learner is able to perform well. On the other hand, for higher numbers of words the problem is their relatively low frequencies. Let us see now our findings in Figure 8.

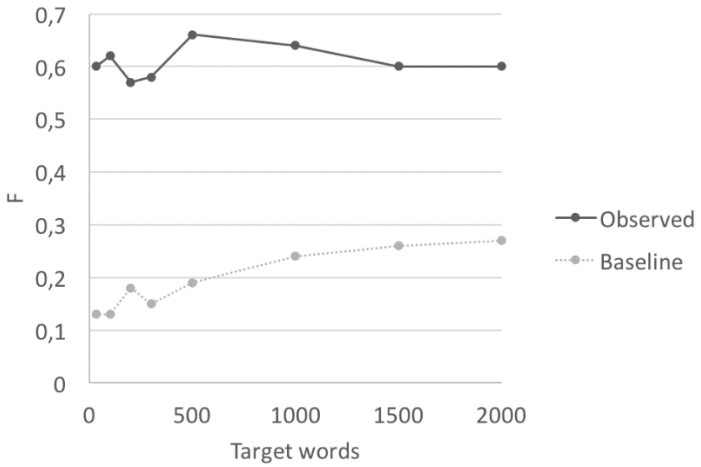

Figure 8. Varying the number of target words.

There seems to be a slight U-shape tendency in our results, given that performance starts around 0.6, has its peak at 500 words (0.66) and then steadily decreases to 0.6 as it moves towards 2000 words. At the same time, we see the random baseline steadily increasing as the number of target words grows: very infrequent items are basically verbs and nouns, thus a fifty percent chance of guessing right by chance. Our peak occurs earlier then for English (500 here, 1000 there) and we suggest that this may be related to the richer morphology of Brazilian Portuguese, which possibly advances the start of the decreasing in relative frequencies of words. We also evaluated a small set with the 31 most frequent words, distributed across 8 categories: pronouns (7), adverbs (6), articles (6), verbs (5), prepositions (3), nouns (2), one conjunction, and one interjection. This diversity of classes probably explains the good performance of the method, even for a very low number of items.

4.2.1. Context words

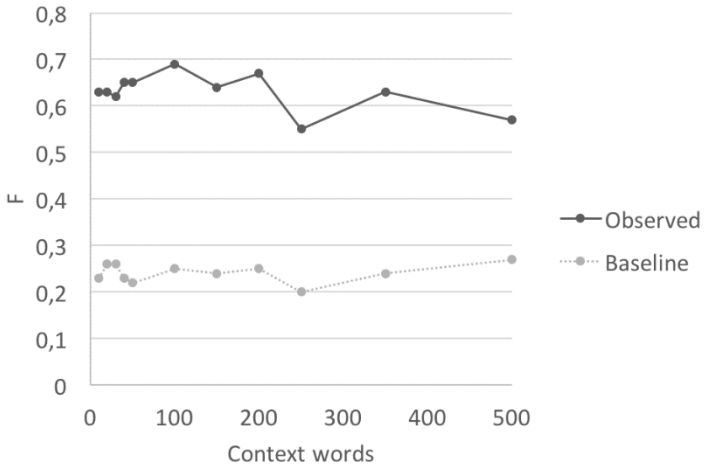

Figure 9. Varying the number of context words.

When varying the number of context words, Redington *et al.* (1998) report again a similar inverted U-shape pattern. Performance was relatively poor for low numbers, with a large gain in performance as the number increased to 50. Beyond this point, increases tended to trade better precision for reduced recall, and beyond 150 context words both precision and recall degraded. For 500 context words the difference between the method and the random baseline was very small: with 1,000 target words precision and recall were 0.40 and 0.44, with random baselines of 0.21 and 0.30. Again, here (see Figure 9), our results are not exactly similar, though they seem to indicate a similar tendency. We see a peak around 100 context words, although performance is not that poor even for only 10 context words.[9] Above 100 context words we also see a clear gradual decrease in performance while the random baseline shows a gradual increase above 250 context words. Differently than the original study, method is always performing well above random baseline.

The answer for the question stated in the beginning of this section given such results is that the method seems effective even for small numbers of target and context words, even more for BP than English. Consequently, it is possible to suggest that distributional information is of help even for the very first steps into word and word category acquisition, when children have acquired just a few items (with or without semantics, given that, as we show here, semantics is not needed for an effective use of distributional cues).

4.3. Experiment 3: performance by category

In this experiment, as in the original, we use the standard analysis setting with performance (and random baselines) calculated for each syntactic category. Redington *et al.* (1998) report the best results for nouns. Verbs are also impressive while performance on adjectives is moderately good but adverb performance is relatively poor. Overall, content words are classified better than functional words. This general picture is consistent with developmental data that shows that open classes are in general acquired and produced first. Our results are very similar but with an interesting difference: articles

[9] Performance for 10 context words holds up for adult-to-adult speech data too. This is at odds with the results for English and we do not have an explanation for that yet. We need a more detailed investigation of our corpus in order to study the distribution of these 10 words and their co-occurrence with target words. We are also working on applying our implementation to English data to see whether Redington *et al.*'s (1998) results are reproduced or not.

are classified moderately better than adjectives, adverbs and the other functional categories, as we see in Table 3.

Table 3. Performance of the learner by category for the standard analysis. Labels: n (number of words), P (precision), and R (recall).

		Observed		Baseline	
Category	n	P	R	P	R
noun	374	0,75	0,43	0,23	0,11
adverb	62	0,09	0,21	0,03	0,11
pronoun	53	0,05	0,09	0,03	0,11
adjective	81	0,09	0,24	0,04	0,11
preposition	11	0,07	0,40	0,01	0,13
verb	334	0,64	0,15	0,20	0,11
article	45	0,23	0,26	0,02	0,11
numeral	14	0,03	0,30	0,00	0,08
conjunction	11	0,01	0,15	0,01	0,14
interjection	15	0,04	0,25	0,01	0,10
Overall	1000	0,71	0,30	0,27	0,11

As an explanation, article is a much bigger category in BP than in English (45 here, 3 there) and they appear a lot as the only elements of noun phrases in BP. Consequently, these elements are easier for the learner to observe and apprehend. Regarding the general picture, as Redington *et al.* (1998) point out, children seem to acquire the major open classes, noun and verb, first. Although semantic information could also predict this ordering of acquisition, distributional learning seems also to be compatible with language development. The poor performances on functional words could be explained by their dependence on content words as their context: these are much less frequent, what makes the context of functional words relatively indeterminate.

4.4. Experiment 4: varying corpus size

Redington *et al.* (1998) report their results for analyses with 100,000 words, 500,000 words, 1 million words, and 2 million words of input. The purpose of this experiment is to determine how much data is minimally necessary for the method to be effective, given that the child will probably hear much more than 2.5 million words a year. As the authors report, the advantage of the method for English data is very slight for the 100,000 words simulation, but with 500,000 of input the advantage is more marked. With 1,000,000 words the method takes off and performs much over the baseline. Finally, if more input is given, it seems likely that small further increases in performance could be expected.

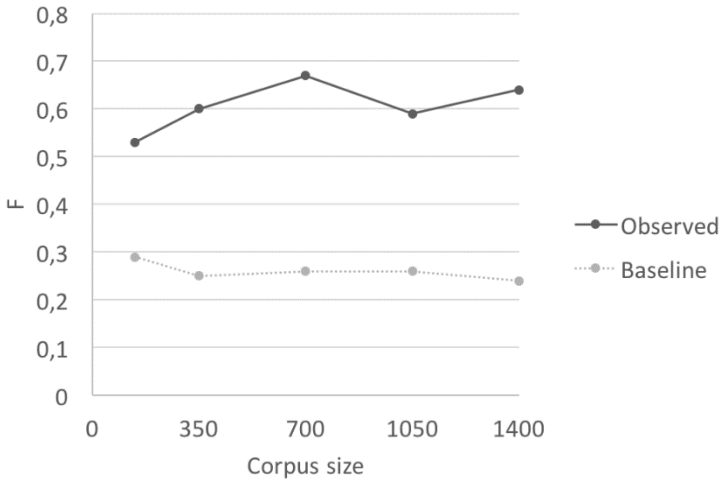

Figure 10. Varying corpus size (in thousands of tokens).

The pictures we obtain in our simulations are less clear (Figure 10). First, whatever the size of the corpus from 140,000 tokens on, the method always performs well above the baseline. There seems to be a tendency of observed performance to decrease towards the baseline for smaller input sets. On the other hand, we are not sure our simulations point to increasing performances as the input data sets grow. Performances do increase up to 700,000 tokens but the tendency is not clear after that. It is possible, after all, that this mark is an upper bound in performance. If that is true, the size of the corpus in our study, being smaller than in Redington *et al.* (1998), does not prevent the direct comparison being conducted here.

4.5. Experiment 7: removing functional words

Experiment 7 evaluates what would happen if the child just ignored functional words whatsoever. This is meant to simulate the plausible situation where a child just ignores these elements in speech, only paying attention to elements of major categories. Redington *et al.* (1998) report that removing functional words has a considerable impact on the performance. However, the analysis still provides a considerable amount of useful information when compared to the baseline (see Figure 12). Our results are very similar (Figure 11): while the standard analysis obtain an observed $F = 0.64$ and 25 clusters, after removing functional words we obtain an observed $F = 0.55$ and 28 clusters (high above the baseline of $F = 0.32$).

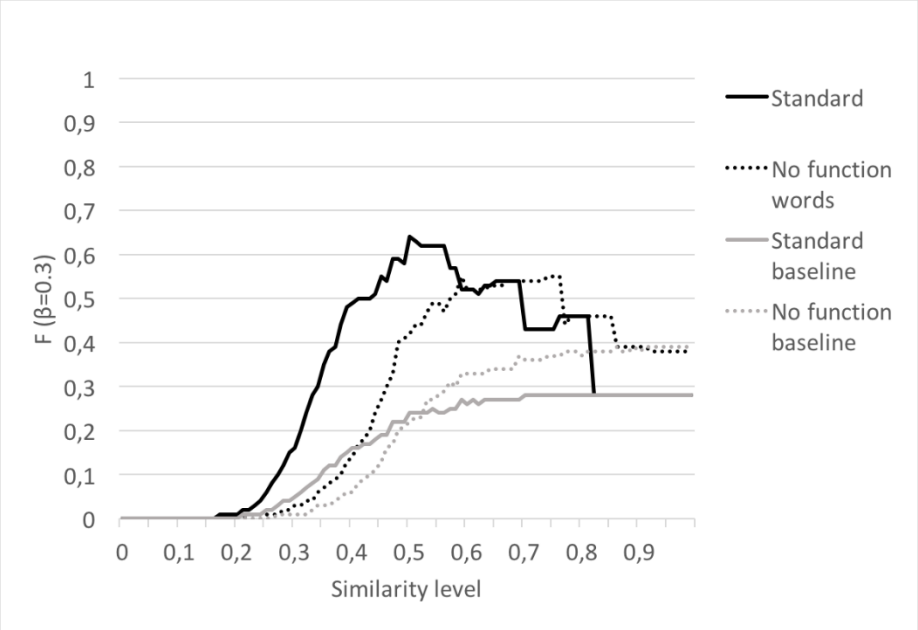

Figure 11. Performances with and without functional words.

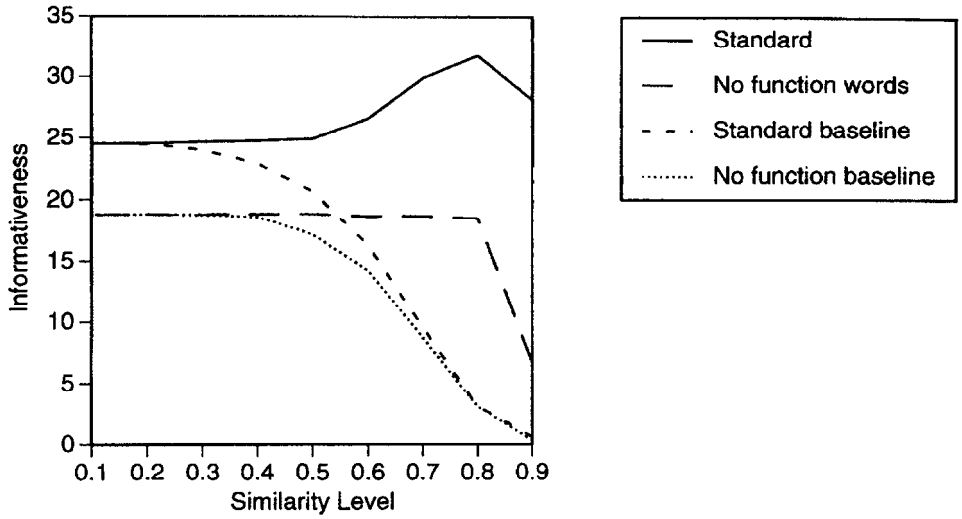

Figure 12. Results of experiment 7 in Redington *et al.* (1998:460).

4.6. Experiment 8: how one category may affect learning another

Experiment 8 was designed by Redington *et al*. (1998) as an effort to investigate how learning would be affected if the child used categorial information, instead of specific words, to classify other words. To do that, three conditions were designed: "noun hints", where all nouns in the corpus were replaced by the symbol NOUN; "verb hints", where all verbs were replaced by VERB; and "function hints", where all functional elements were replaced by FUNCTION. Figure 13 shows our results in this experiment. Compared to the standard analysis, conditions "noun hints" and "verb hints" show lower performances (F=0.50 and F=0.59, respectively). In condition "function hints", however, the learner performs relatively better, with F=0.70 (prec.=0.74, recall=0.44, 14 clusters). This result makes sense to us, given that in this condition there are only four categories to learn, exactly the open classes for which the method performs better.

Nonetheless, our results in this experiment are very different from Redington *et al.*'s (1998) (Figure 14). In their study, all conditions show a decrement in performance

and, strikingly, "function hints" is the worst. They interpret their results as suggesting that this (categorial) source of information may not be appropriate for distributional analysis. Would this difference in results be due to differences between languages? Or can it be traced back to differences in implementations? In future work we will be able answer these questions by applying our model to English data and after a deeper analysis of our results.

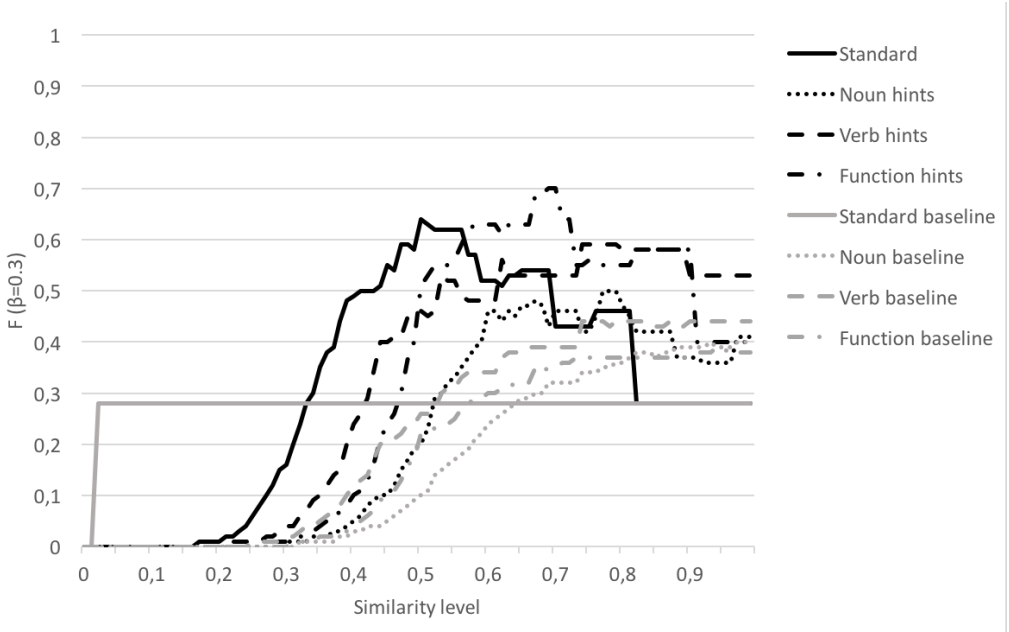

Figure 13. How one category affects the learning of others.

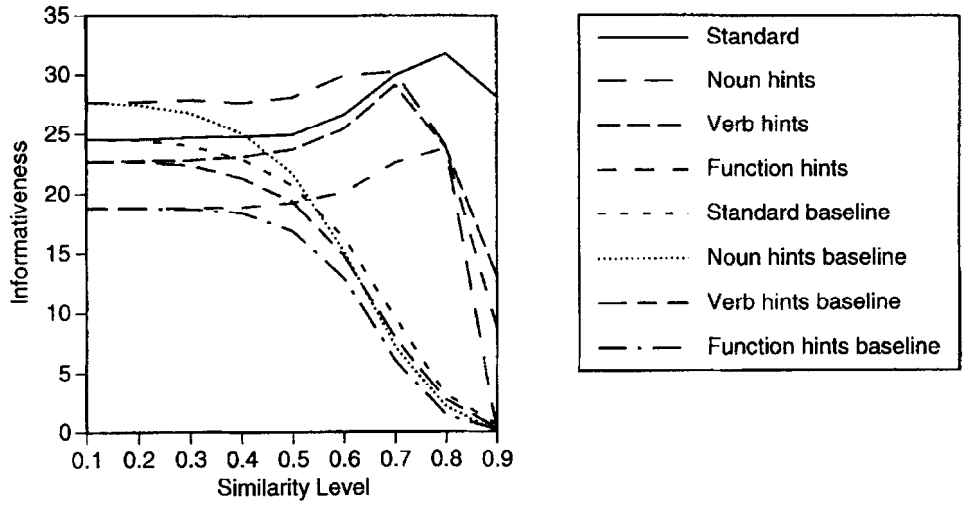

Figure 14. Experiment 8 in Redington *et al.* (1998:461).

4.7. Experiment 9: child-directed vs. adult-to-adult speech

This final experiment evaluates whether child-directed speech in BP – as in English – is indeed "marked", that is, specially modified when compared with normal speech, represented here by the corpus of adult-to-adult speech. The speech directed to the child is also called "motherese" and is characterized, among other things, by shorter utterances, simpler structures (less subordination), more restricted vocabulary, redundancy, etc. (Snow 1977). This style of speech would then be – arguably – facilitating for the child to

learn language in her earlier years. Of course, this facilitation may emerge in various forms during acquisition and we would like to know whether it is facilitating for the distributional part-of-speech learning. We can answer that by comparing the performances of the learner on CDS data and adult-to-adult data. Figures 15 and 16 show, respectively, results for BP and English. As we can see, adult-to-adult speech seems to provide a slight advantage for the distributional strategy in both languages (for English, this is measured at the 0.8 level of similarity). For BP, an *F* of 0.65 is obtained, with a precision of 0.65, recall of 0.72, and 9 clusters. This is an interestingly different pattern when compared to the other experiments, because here precision is relatively high while recall is even higher and the number of clusters obtained is closer to the benchmark number (quantitatively). This is an indication that the contribution of "motherese" to learning rests on different aspects of language, not on its distributional information, at least regarding BP and English.

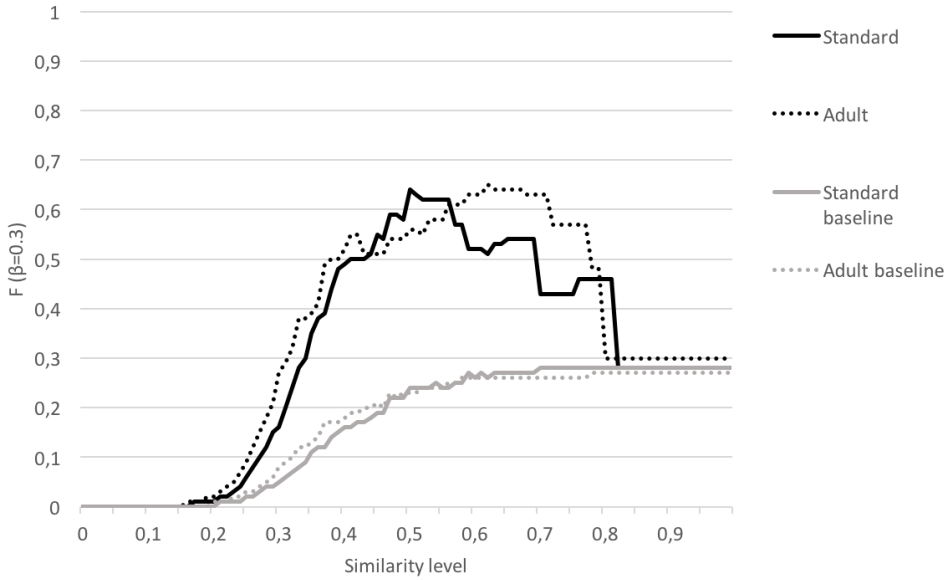

Figure 15. Comparison between child-directed speech and adult-to-adult speech.

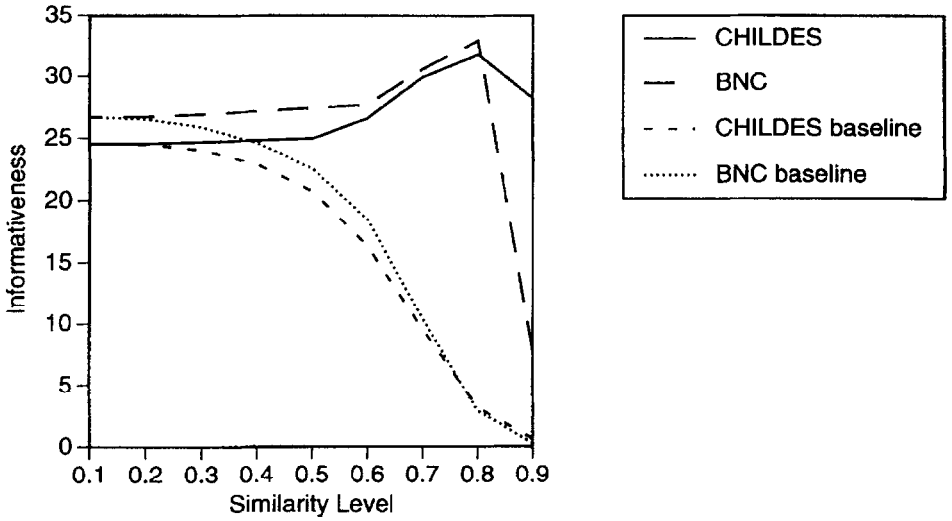

Figure 16. Experiment 9 in Redington *et al*. (1998:462).

4.8. Assessing the cosine metric

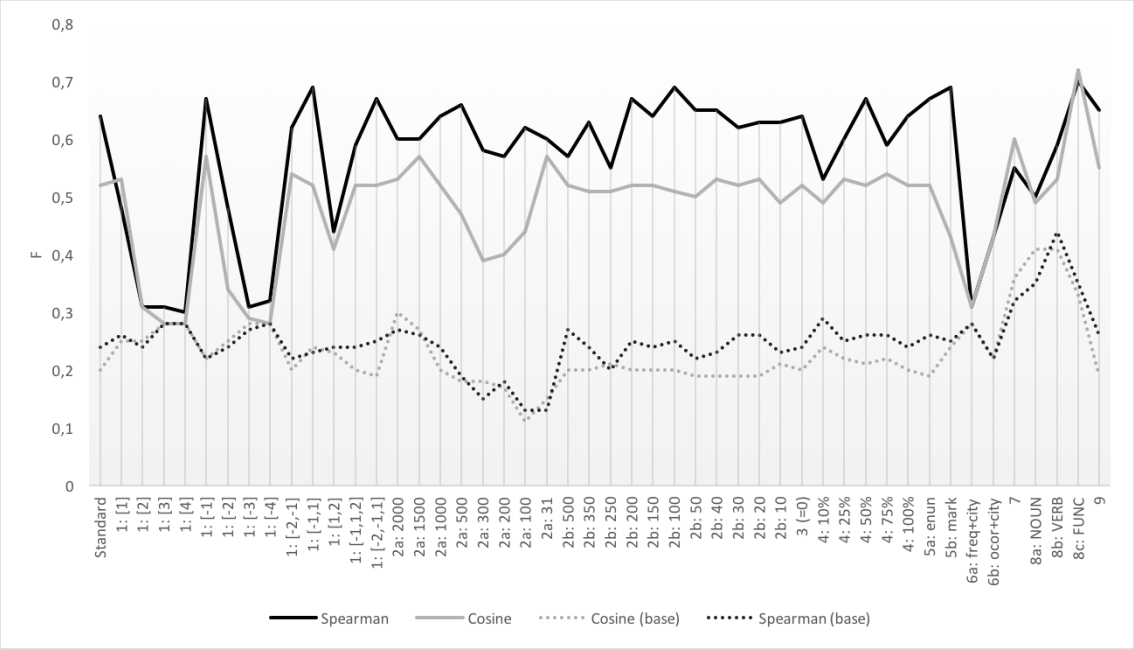

Figure 17. Comparison of the learner's performances using either Spearman rank correlation or cosine metrics. Standard deviations of differences between observed and baseline: Spearman = 0,11; Cosine = 0,09.

In section 3.1, we mentioned that we chose to work with the spearman rank correlation metric in order to more precisely replicate Redington *et al.*'s (1998) study. However, since their study, the cosine metric has been established as a standard for comparing context vectors in many kinds of distributional analysis applications. In order to see how this metric performs in our study, relatively to the Spearman correlation, we performed all experiments again using the cosine. Figure 17 shows the full quantitative comparison. As we can see, the shapes of both observed curves (as well as baselines) are very similar. The main difference is that the Spearman rank correlation seems to provide a better classification overall, as indicated by the standard deviations when we compare observed and baselines for each metric.

5. Final remarks

As indicated in the beginning of this paper, the analysis conducted here focused on quantitative comparisons and discussions. We are aware of the importance of a deeper qualitative analysis that not only provide a broader understanding of what these numbers mean, but crucially that relate our results to theoretical and empirical considerations about the acquisition of part-of-speech categories, language acquisition in general, and also about psychological plausibility. Nonetheless, quantitatively there are interesting things to say about similarities between BP and English:

- Very local contexts are much more informative (experiment 1). This is compatible with processing capabilities of young children, such as shorter working memory;
- Distributional analysis is informative even for small numbers of target and contextual words (experiment 2). This means that since its first steps, the child might be benefiting from distributional analysis of the input she hears;

- Experiment 3 shows that open-ended categories, specially nouns and verbs, are easier to learn, with some clusters coming close to be "pure" (*e.g.*, a cluster of infinitival verbs). On the other hand, functional categories are harder to learn in both languages. Whether this is a distributional fact or a problem of benchmark categories assumed or a mixture of both is still to be determined;
- The amount of data needed for the method to be effective is also much lower than what a child is expected to hear (experiment 4): with some estimates coming to 1.5 million of words directed to a child per year, we could see that 700,000 words seemed effective for BP, when learning the categories of the 1000 most frequent words;
- Explicit utterance boundaries help learning (experiment 5), indicating also that intra-sentential boundaries may also be of help. This is compatible with the attested sensitivity of children to phonological cues of phrasal and sentential boundaries;
- Sensitivity to functional elements in speech is demonstrated to be crucial for the full benefit of distributional analysis (experiment 7);
- Distributional learning is more effective for typical adult-to-adult speech than for child-directed speech (experiment 9). This means that whatever benefits the "motherese" style may bring to language acquisition, it is probably related to other properties of language that need to be learned (*e.g.*, pragmatics, word identification on the speech stream, etc.).

Nonetheless, our study also brings some possibly conflicting results, when compared to Redington *et al.*'s (1998) results for English:

- Our experiment 6 brings complicating results to the question of whether frequencies are strictly necessary, instead of a (arguably psychologically) simpler strategy of tracking binary counts of co-occurrence. While Redington *et al.*'s (1998) had reasons to conclude for the importance of frequencies, our results indicate that the Spearman correlation metric may be playing a role in the learner's performance here. More studies are necessary, in particular, we need to run our model on English data to have a more direct comparison with Redington *et al.*'s (1998) model.
- While for English, the authors report a decrease for all conditions in experiment 8, concluding that categorial information is not successfully integrated to distributional analysis, our results show a more complicated picture: for BP, the categorial contribution of functional elements, as a single class, may be of help for the learning of open classes. And we started asking the question of what would happen to learning if instead of a broad FUNCTIONAL symbol we used all functional categories themselves (article, preposition, etc.)? We will answer it in future work.

The present study is part of our effort to provide an in-depth understanding of Brazilian Portuguese regarding the role of distributional information, both by reflecting on its own properties and by comparing it with similar studies for other languages. We began by replicating Redington *et al.* (1998) study and now we are moving on to more recent ones, including evaluating the suitability and plausibility of models such as Baroni and Lenci (2010), Mikolov, Sutskever, Chen, Corrado & Dean (2013), and Pennington, Socher & Manning (2014), to this task. In addition, we will investigate other factors, such as evaluating other vector similarity measures, as well as trying mathematical techniques

to deal with lower frequencies and noise, weighting, sparsity, and optimizations. Given the richer morphology of Brazilian Portuguese, we also want to explore such information, as in Clark (2003).

It is worth noting 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. Consequently, matters of psychological, developmental, and empirical plausibility must be considered for every instance of a model we develop or use. In this regard, some of these 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.

As next step, we are working to provide a qualitative analyzes of our results, discussing the actual categorizations obtained by the method, their purity, outliers, spurious results, etc., and also qualitative properties of our *input data*. We understand that this is crucial for a deep linguistic understanding of the role of distributional learning in the process of language acquisition.

References

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

Bernal, S., Lidz, J., Millotte, S. & Christophe, A. (2007). Syntax constrains the acquisition of verb meaning. *Language Learning and Development*, *3*, 325–341.

Berwick, R. C., Pietroski, P., Yankama, B. & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science*, *35*, 1207–1242.

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

Clark, A. (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* (pp. 59–66). Stroudsburg, PA, USA: Association for Computational Linguistics.

Faria, P. (2019). The Role of Utterance Boundaries and Word Frequencies for Part-of-speech Learning in Brazilian Portuguese Through Distributional Analysis. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (NAACL'19), 152–159.

Faria, P. & Ohashi, G. O. (2018). A aprendizagem distribucional no português brasileiro: um estudo computacional. *Revista LinguíStica*, *14* (3), 128–156.

Frank, M. C. (2011). Computational models of early language acquisition. *Current Opinion in Neurobiology*, *21* (3), 381–386. https://doi.org/10.1016/j.conb.2011.02.013.

Galves, C., Andrade, A. L. de & Faria, P. (2017). *Tycho Brahe Parsed Corpus of Historical Portuguese*. < http://www.tycho.iel.unicamp.br/corpus/en/>.

Harris, Z. S. (1954). Distributional structure. *Word*, *10* (2-3), 146–162.

Kaplan, F., Oudeyer, P.-Y. & Bergen, B. (2008). Computational models in the debate over language learnabiliy. *Infant and Child Development*, *17* (1), 55–80.

Landau, B. & Gleitman, L. R. (1985). *Language and experience: evidence from the blind child*. Harvard University Press, Cambridge, MA.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B. & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26* (3), 269–286.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (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* (pp. 3111–3119). Lake Tahoe, NV, USA: Curran Associates Inc.

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. *EMNLP, 14*, 1532–1543.

Pullum, G. K. (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* (pp. 498–513). Berkeley, California: Berkeley Linguistics Society.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22* (4), 425–469.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275* (14), 1599–1603.

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*, 1–22.

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

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

Wintner, S. (2010). Computational Models of Language Acquisition. In A. Gelbukh (Ed.), *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing* (CICLing'10) (pp. 86–99). Berlin, Heidelberg: Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-12116-6_8

Yang, C. (2002). *Knowledge and learning in natural language*. Oxford: Oxford University Press.

Yang, C. (2012). Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*, 205–213. http://dx.doi.org/10.1002/wcs.1154

[received on June 1, 2019 and accepted for publication on November 6, 2019]