Bootstrapping the syntactic bootstrapper: Probabilistic labelling of prosodic phrases

Ariel Gutman[*] Isabelle Dautriche[†] Benoît Crabbé[‡]
Anne Christophe[†]

Abstract

The *syntactic bootstrapping* hypothesis proposes that syntactic structure provides children with cues for learning the meaning of novel words. In this paper, we address the question of how children might start acquiring some aspects of syntax before they possess a sizeable lexicon. The study presents two models of early syntax acquisition that rest on three major assumptions grounded in the infant literature: First, infants have access to phrasal prosody; second, they pay attention to words situated at the edges of prosodic boundaries; third, they know the meaning of a handful of words. The models take as input a corpus of French child-directed speech tagged with prosodic boundaries, and assign syntactic labels to prosodic phrases. The excellent performance of these models shows the feasibility of the syntactic bootstrapping hypothesis, since elements of syntactic structure can be constructed by relying on prosody, function words and a minimal semantic knowledge.

[*] *Zukunftskolleg* & Department of Linguistics, University of Konstanz, Germany.
Address: Ariel Gutman, Zukunftskolleg, Box 216, 78457 Konstanz, Germany. Email: ariel.gutman@uni-konstanz.de

[†] *Laboratoire de Sciences Cognitives et Psycholinguistique*, *École Normale Supérieure*/PSL Research University/CNRS/EHESS, Paris, France.

[‡] ALPAGE, INRIA/Université Paris-Diderot, France.

1 Introduction

Children acquiring a language have to learn its phonology, its lexicon, and its syntax. For a long time researchers, focusing on children's productions, thought that children start by learning the phonology of their language, then work on their lexicon, and only once they have a sufficient store of words do they start acquiring the syntax of their language (in correspondence to their productions; up to one year: babbling; 1 to 2 years: isolated words; at 2 years: first sentences). However, a wealth of experimental results has shown that children start acquiring the syntax of their native language much earlier. For instance, at around one year of age they recognize certain function words (determiners) and appear to use them to categorize novel words (Shi & Melançon 2010). Indeed, it has been previously suggested that children may use the syntactic structure of sentences to facilitate their acquisition of word meanings (the *syntactic bootstrapping* hypothesis; Gleitman 1990). In this paper, we address the question of how children might start acquiring some aspects of syntax before they possess a sizeable lexicon.

How might children infer the syntactic structure of sentences? Since prosody correlates with syntactic structure, and young children are sensitive to prosody, phrasal prosody has been suggested to help bootstrap the acquisition of syntax (Morgan 1986; Morgan & Demuth 1996). However, even though phrasal prosody provides some information regarding syntactic constituent boundaries, it does not provide information regarding the nature of these constituents (e.g. noun phrase, verb phrase). In this paper, we address the question of whether such information can be retrieved from the input. Computational modelling is an essential step in answering this question, as it can test the usefulness of hypothesized sources of information for the learning process. Specifically, we propose a model that attempts to categorize prosodic phrases by relying on distributional information and a minimal semantic knowledge.

Several models have shown that distributional information is useful for categorization (St. Clair et al. 2010; Chemla et al. 2009; Mintz 2003; Mintz et al. 2002; Redington et al. 1998; Schütze 1995). For instance, in the *frequent frame* model proposed by Mintz (2003), the model groups together all words X appearing in a context, or frame, of the type [A X B], where A and B are two words frequently occurring together. This model builds highly accurate syntactic categories based on only a few highly frequent contexts (e.g., [the X is] selecting nouns, or [you X the] selecting verbs). Importantly, young infants have been shown to use distributional information for categorizing words in a number of experiments using artificial languages (e.g. Marchetto & Bonatti 2013; Gomez & Gerken 1999) and natural languages (e.g. Van Heugten & Johnson 2010; Höhle et al. 2006). A common feature of all these approaches is that the most useful contexts for categorization turn out

to contain function words, such as determiners, auxiliaries, pronouns, etc.

The models we present here rest on three major assumptions: 1. infants have access to phrasal prosody; 2. infants pay attention to 'edge-words', words situated at the edges of prosodic units; and 3. infants know the meaning of a handful of words, the *semantic seed*.

Regarding the first assumption, infants display sensitivity to prosody from birth on (e.g. Mehler et al. 1988). For example, four-month-old children are sensitive to major prosodic breaks, displaying a preference for passages containing artificial pauses inserted at clause boundaries over passages containing artificial pauses within clauses (Jusczyk et al. 1995). Sensitivity to smaller prosodic units is attested at 9 months of age (Gerken et al. 1994). Slightly older infants use prosodic boundaries to constrain lexical access. That is, 13-month-old infants trained to recognize the word *paper* correctly reject sentences where both syllables of *paper* are present, but span across a prosodic boundary as in [the man with the highest *pay*][*per*forms the most] (Gout et al. 2004; see also Johnson 2008). Finally, older children have been shown to use phrasal prosody to constrain their online syntactic processing of sentences (de Carvalho et al. 2013; Millotte et al. 2008). This early sensitivity to prosodic information has been mirrored by computational models succeeding in extracting information regarding syntactic boundaries from the speech signal (Pate & Goldwater 2011). In order to integrate this prosodic information directly, our models operate on a corpus of child-directed speech automatically tagged with prosodic boundaries. To our knowledge, no model to date has incorporated prosodic information in a model of category induction (but see Frank et al. (2013) for the incorporation of sentence type).

The second assumption states that words situated at the edges of prosodic phrases play a special role. We are specifically interested in these words for two distinct reasons. First, words at edges tend to have a special status: depending on the language, syntactic phrases typically either start with function words (or morphemes) and end with content words, or start with content words and end with function words. Focusing on words at the edges of prosodic phrases is therefore an easy way to enhance the weight of functional elements, which is desirable because function words are the elements that drive the classification in distributional models of syntactic categorization (Chemla et al. 2009; Mintz 2003; Redington et al. 1998). Second, the infant literature shows that infants are especially sensitive to edge-words. For instance, words situated at the end of utterances are easier to segment than words situated in sentence-medial position (Seidl & Johnson 2006; Shukla et al. 2007; Johnson et al. 2014). Our model therefore relies on the edge-words of prosodic phrases to compute the most likely category of each prosodic phrase.

The third assumption states that children learning the grammatical categories of their language, presumably before their second birthday (e.g. Brusini et al. 2009;

Bernal et al. 2010; Oshima-Takane et al. 2011), are equipped with a small lexicon to help them with this task. This assumption is highly plausible, as recent evidence has shown that infants as young as 6 to 9 months know the meaning of some nouns in their language (Bergelson & Swingley 2012; Tincoff & Jusczyk 2012; Parise & Csibra 2012). It seems, moreover, that they start learning the meanings of verbs at the age of 10 months (Bergelson & Swingley 2013). Children could group words together according to their semantic category as soon as they start to know the meaning of basic words. For example, they could start grouping together *toy*, *car* and *teddy-bear* because they all refer to concrete objects, and *drink*, *eat* and *play* because they all refer to actions. Because nouns are likely to refer to objects and verbs to actions, these basic semantic categories may constitute a seed for the prototypical "noun" and "verb" grammatical categories. In order to estimate the benefit of a small lexicon in our models of prosodic phrases categorization, we use this additional semantic knowledge, *the semantic seed*, in our second model.

The basic idea of the model is thus that prosodic boundaries signal syntactic boundaries (following Morgan 1986; Morgan & Demuth 1996), while function words (appearing at the edges of prosodic phrases) serve to *label* the prosodic phrases. For instance, in the example below, a sentence such as *He's eating an apple* may be split into two prosodic phrases: [He's eating] [an apple]. The first words of each of these prosodic phrases happen to be function words: *he* and *an*. These words may allow the models to attribute the first prosodic phrase to a class containing other verbal nuclei (VN, a phrase containing a verb and adjacent words such as auxiliaries and clitic pronouns), and the second one to a class containing other noun phrases (NP).

Input sentence	He's eating an apple.
Prosodic structure	[He's eating] [an apple]
Syntactic skeleton	[*He*'s eating]$_{\text{VN}}$ [*an* apple]$_{\text{NP}}$

In fact, the present model follows the *syntactic skeleton* proposal, according to which children may combine their knowledge of function words and prosodic boundaries to build an approximate shallow syntactic structure (Christophe et al. 2008). We present two 'modelling experiments' testing whether access to phrasal prosody, edge-words, and a semantic seed is sufficient to label prosodic phrases. The first model relies only on the first two assumptions: it has access to prosodic boundaries, and gives a special status to edge-words. The second model further adds the semantic seed assumption.

In addition to these three major assumptions, both models also incorporate an additional, less crucial, constraint. In natural languages, function words tend to appear either at the beginning (on the left) or at the end (on the right) of syntactic phrases, and several experiments suggest that infants can deduce this by the age of 8

months (Gervain & Werker 2013; Bernard & Gervain 2012; Hochmann et al. 2010; Gervain et al. 2008). In French, function words tend to appear phrase-initially and content words phrase-finally. Accordingly, both our models incorporate a left-right asymmetry (although they could be rendered symmetric, see discussion).

The two models are presented in detail below.

2 Experiment 1

In this experiment, the model uses a clustering algorithm which explicitly relies on the intuition that in a head-initial language like French, the first word of a prosodic phrase is often a function word which is informative of the category of the prosodic phrase. This intuition is illustrated in Table 1. Consequently, in this experiment, classes are built by grouping together prosodic phrases that start with the same word (using frequent phrase-initial words). For instance, the prosodic phrase ***le*** *petit oiseau désolé* '**the** sad little bird' would be assigned to a class labelled *le* 'the (masc.)'.

[***Le*** *petit* <u>*oiseau*</u> *désolé*]$_{\text{NP}}$ [***est*** *prêt à* <u>*pleurer*</u>]$_{\text{VN}}$ '**The** desolate little <u>bird</u> / **is** nearly <u>crying</u>.'
[***Elle*** <u>*prend*</u>]$_{\text{VN}}$ [***le*** *petit* <u>*cheval*</u>]$_{\text{NP}}$ '**She** <u>takes</u> / **the** small <u>horse</u>.'
[***Tu*** <u>*veux*</u>]$_{\text{VN}}$ [***que*** *je* <u>*reste*</u> *là*?]$_{\text{VN}}$ '**You** <u>want</u> / **that** I <u>stay</u> here?'

Table 1: Examples from the prosodically augmented corpus. The text is divided into prosodic phrases, which are labelled for evaluation according to their <u>underlined</u> lexical heads: VN = Verbal Nucleus; NP = Noun Phrase. For comparison, the function words which may help our classification model are given in **bold-face**. These markings are reproduced on the corresponding word in the English translation.

2.1 Material

2.1.1 Input corpus

We used the Lyon corpus collected by Demuth & Tremblay (2008) (available at `http://childes.psy.cmu.edu/data/Romance/French/Lyon.zip`), containing conversations with four children aged between 1 and 4 years, forming part of

the *Childes* database (MacWhinney 2000). From the corpus, we extracted the orthographically transcribed raw text (ignoring all meta-data), without the speech of the child itself, as we are interested only in child-directed speech. This resulted in approximately 180,000 utterances, consisting of approximately 700,000 words.

2.1.2 Prosodic tagging of the corpus

The model takes as input a corpus of orthographically transcribed speech (i.e., divided into word-like units),[1] to which prosodic information (i.e., the segmentation of the speech onto prosodic phrases) was added. For the sake of simplicity, prosodic boundaries were automatically derived from the corpus relying on current linguistic theory, as explained below.

The raw corpus was syntactically analyzed by a state-of-the-art French parser (Crabbé & Candito 2008). The text was then automatically segmented into prosodic phrases, using the notion of the *phonological phrase* defined in the theory of prosody proposed by Nespor & Vogel (2007). This theory has the merit of being relatively explicit, and is thus suitable for algorithmic implementation. In addition, it is accepted by a large part of the linguistic community.[2] Moreover, the phonological phrases are generally comparable to the syntactic phrases we are interested in, namely the NP and VN (cf. Selkirk 1984). According to this theory, "[t]he domain of ϕ [=Phonological phrase] consists of a C [=Clitic group] which contains a lexical head (X) and all Cs on its nonrecursive side [i.e., left side] up to the C that contains another head outside of the maximal projection of X." (Nespor & Vogel 2007:168). The automated process also took into account the following optional reconstruction rule, whenever the prosodic phrase was followed by a short complement (up to 3 syllables): "A nonbranching ϕ which is the first complement of X on its recursive side [i.e., right side] is joined into the ϕ that contains X." (Nespor & Vogel 2007:173). The lexical head X (i.e. a noun, verb, adjective, adverb or interjection) which appears in the definition, allows us to assign a syntactic category to each prosodic phrase (namely, the phrasal category of X, as provided by

[1] We assume that our child model has knowledge of word boundaries. This assumption is reasonable in the case of function words because of their frequency (Hochmann et al. 2010). However, the age at which children have adult-like segmentation of the full speech signal is unknown (Nazzi et al. 2006; Ngon et al. 2013). Note that in our model this assumption is not crucial since we aim to categorize prosodic phrases rather than words.

[2] See, however, the contrasting view of Lahiri & Plank (2010), who oppose the view that prosodic phrasing is strictly dependent on syntactic constituency. They claim that in Germanic languages functional elements often cliticize to the syntactic constituent preceding them, even though they syntactically belong to the following constituent, such as in the clause [drink a][pint of][milk a][day] where square bracket mark prosodic units (p. 376). If this is the case in child-directed speech, our predictive model would have to be adapted such that it takes into account frequent words both before and after the initial boundary of each prosodic phrase.

the parser), which we consider as the correct category of the phrase for evaluation purposes. This lexical head often appears at the end of the prosodic phrase, though this is not always the case due to the above-mentioned reconstruction rule. Table 1 presents some examples taken from the prosodically tagged corpus.

As a final clean-up step, we discarded all utterances that consist of a single word, which amount to approximately 22% of our corpus. While single word utterances may play a role in word learning (Lew-Williams et al. 2011), they are not interesting for our purposes: Since they appear without context, they can hardly be classified syntactically without knowing their content. Moreover, they mostly consist of categories which are not of interest to us: a third of these utterances are interjections (*oui* 'yes', *oh* etc.), and another third are proper names, according to the parser. Only 11% are VNs (mostly imperatives, e.g. *Regarde!* 'Look!').

Our procedure resulted in a corpus with 246,013 prosodic phrases. In most of the experiments we divided the corpus into 10 non-consecutive mini-corpora, each containing about 24,601 prosodic phrases, to estimate the variability in performance.

Although the results of the prosodic phrase segmentation procedure are good, they are not perfect, in part because the syntactic parser we used was not specifically designed to deal with spoken language. Nonetheless, a comparison of our algorithmic segmentation with segmentation conducted by human annotators on a sample of randomly selected sentences showed that our method gives satisfactory results for our needs: The human annotators annotated the prosodic boundaries of 30 written sentences following an example provided to them. The average agreement rate between the annotators and the algorithm was 84%, only slightly lower than the agreement rates between the annotators, 89%.

We also evaluated the quality of the syntactic labelling of the prosodic phrases by the parser: Two annotators categorized the head word of each prosodic phrase as noun, verb or another category (since these are the categories which interest us most). Their inter-annotator agreement rate was 91% and the average agreement rate with the label assigned by the parser was 79%, which we considered sufficient for our purposes.

2.2 Method

2.2.1 A probabilistic model

We use a Naive Bayes model to categorize prosodic phrases. In our case, we use this model to specifically express the class C of each prosodic phrase in our corpus

conditional to a series of m independent predictor variables V_i.

$$p(C = c|V_0 = v_0 \dots V_m = v_m) = \frac{p(C = c) \prod_{i=0}^{m} p(V_i = v_i|C = c)}{p(V_0 \dots V_m)} \quad (1)$$

For the specific case of predicting a class $\hat{c}$ given some known predictor variables, the decision rule amounts to maximizing the following formula:

$$\hat{c} = \underset{c \in C}{\text{argmax}}\ p(C = c) \prod_{i=0}^{m} p(V_i = v_i|C = c) \quad (2)$$

This equation says that the predicted class $\hat{c}$ is the one that maximizes the product of its prior probability $p(C = c)$ and the conditional probability of the different predictor variables given the class value.

In this experiment, the set of classes $c \in C$ is defined as follows: the k most frequent words at the beginning of the prosodic phrases containing at least two words are used to define k classes, each of them initially corresponding to prosodic phrases starting with that frequent word. The parameter k is allowed to vary from 5 to 70 in this experiment. This design captures the intuition that in a head-initial language, the first words of prosodic phrases will usually be function words. Indeed, when k is small, the most frequent phrase-initial words are function words. For instance, among the 50 most frequent phrase-initial words, there are only three content words, namely *faut* '(one) must', *regarde* 'look', and *fais* 'do'.

For each data point, the predicting observations $V_i = v_i$ are word forms chosen to represent the linguistic context and content of each prosodic phrase. These variables reflect our assumption that the child is especially sensitive to both function words and content words appearing near the boundaries of prosodic phrases. In a language like French, which was the language used for conducting this experiment, first words are mostly function words while final words are mostly content words. To capture this, our learning model uses the two prosodic phrase edge-words as variables, dubbed L_0 for the first, "leftmost" word and R_0 for the final, "rightmost" word. Following preliminary experimentation with the model, we also included the second word of the phrase dubbed L_0'[3]. Intuitively, this is important since the "true" function word of a phrase can appear in the second position as well, following a conjunction, as in: *mais le bébé* 'but the baby', *que je sache* 'that I know' etc. In order to model the immediately preceding context of the phrase, we also selected the first word of the preceding phrase, L_{-1}, as a variable. Hence, the set of predictor variables is $V = \{L_{-1}, L_0, L_0', R_0\}$ (see Table 2 for an example).[4]

[3] In the special case where a prosodic phrase contains only one word, we have $L_0 = R_0$ and L_0'

	L_{-1}	L_0	L'_0		R_0		
#	***Ah***	***tu***	***me***	*donnes*	***aussi***	*une*	*cuillère*
#	**Oh**	**you**	**me**	give	**too**	a	spoon

'Oh, you're also giving me a spoon.'

Table 2: Example of variables used in an utterance divided into three prosodic phrases. The focus of the predictor is on the second phrase. Words which are used as predicting variables are given in **bold-face**, below the name of the variable. The # symbol represents the beginning of the utterance.

Clearly, the independence hypothesis of the model is too strong. The predictor variables V_i, conditionally dependent on C, are not independent. However, common experience with the Naive Bayes model has shown that this strong independence assumption entails a computationally tractable framework without impeding its predictions. This is also the case for the current study.

2.2.2 Parameter estimation

The purpose of the parameter estimation mechanism described below is to estimate the parameters of the probabilistic model (i.e. the prior probabilities $p(C = c)$ and the conditional probabilities $p(V_i = v_i|C = c)$ present in equation 2) in a case where some variables remain unobserved in the data (the class variable C in our model). Here, we use the Naive Bayes Expectation Maximisation algorithm (NB-EM) as described by Pedersen (1998). In this algorithm, each data point is initially randomly assigned to a category (initialization step). Subsequently, the model parameters are calculated according to this assignment (maximization step). Using the newly calculated model parameters, the data points are re-assigned to the various categories (expectation step). These two steps are iterated until the resulting likelihood of the data set ceases to increase.[5] Note that the number k of possible categories is chosen initially (as one of the hypotheses of the model) and does not change subsequently.

---

is void.

[4] As mentioned above, our choice of predictor variables has a built-in "leftward" bias, due to the fact that our model is designed to work with French child-directed speech. In the conclusions we discuss the plausibility of this bias, and ways we can extend our model to be more "symmetric".

[5] The NB-EM algorithm is a standard parameter estimation algorithm that can potentially suffer from local minima. Nevertheless, our results were extremely stable, as evidenced for instance by the almost invisible error bars in Figure 1, even though each of the different sub-corpora was rather small. This suggests that the behaviour of the model itself is highly stable, and would not change with a different procedure for estimating the model's parameters.

Once the parameters are estimated, the model can be used to predict the categories of each prosodic phrase in the corpus using the decision rule given in (2), so that each prosodic phrase is assigned to one of the k classes.

2.2.3 Initial clustering according to frequent L-words

Instead of using a random initialization phase as is typical with the NB-EM algorithm, each prosodic phrase is assigned initially to a category corresponding to its first word (the L-word), if and only if this word is part of the k most frequent L-words appearing in prosodic phrases longer than one word (as one-word phrases cannot normally contain a function word). If this is not the case, the prosodic phrase is initially left unassigned (see Table 3 for examples).[6] The subsequent maximization phase is based only on those data points that had initially been categorized. Then, the NB-EM algorithm proceeds normally.

Phrase		Assigned Category
vas-y	'go! (sg.)'	Not assigned initially
tu vas apprendre	'you (sg.) will learn'	*tu* 'you (sg.)'
je vais prendre	'I will take'	*je* 'I'
le bain	'the bath'	*le* 'the (masc. sg.)'
au bébé	'to the baby'	Not assigned initially
et le crocodile	'and the crocodile'	*et* 'and'

Table 3: Examples of prosodic phrases with their initial functional category, when initializing with the 10 most frequent function words (*tu, c', et, il, on, ça, je, qu', de, le*).

2.2.4 Evaluation measures

In order to evaluate the performance of our model, and compare it to a model whose parameters are estimated with a full random-initialisation, we calculated for each resulting class (i.e., group of phrases whose predicted category is the same), its *purity* measure, which measures how well this class captures a real syntactic category (as given by our rule-based parser). Following Strehl et al. (2000), we measure

[6] The idea of creating initial classes which contain only one type of "head-word" is similar to the idea proposed by Parisien et al. (2008). However, in their algorithm, the head-word could be any word in the stream of words, while in our algorithm the "head-word" must be an L-word of a prosodic phrase.

this by comparing the size of the class to the size of the largest syntactic category represented in it. Formally, this gives the following definition:

$$\text{purity}(\text{Cl}) = \max_i \frac{|\text{Cat}_i \cap \text{Cl}|}{|\text{Cl}|} = \frac{\text{Size of largest category}}{\text{Class size}} \tag{3}$$

A class which has an absolute majority of one phrasal category (purity > ½) can be considered a reasonably good class. A good class will exhibit a purity above ⅔.[7]

The purity measures of all k classes can be averaged in order to estimate the overall success of the algorithm.

As a further baseline of comparison, we use "chance purity" which is the average purity that would result if we would distribute the prosodic phrases by chance in the k classes. This should be equal to the proportion of the largest phrasal category, which happens to be the VN category with a proportion of approximately 37%.

Precision and recall of best classes As explained above, we are particularly interested in the labelling of prosodic phrases which correspond to the VN and NP categories. In the current experiment, we have no classes which correspond uniquely to these labels, but for comparison purposes with other approaches (as well as with Experiment 2) we can *a posteriori* select the class with the highest ("best") proportion of VNs, and the class with the highest proportion of NPs, and label them as such. For those classes we can calculate the standard *recall* and *precision* measures, which are defined as follows:

$$\text{precision} = \frac{\text{Number of hits}}{\text{Class size}} \tag{4}$$

$$\text{recall} = \frac{\text{Number of hits}}{\text{Category size}} \tag{5}$$

The term "hit" in this context should be understood as a VN prosodic phrase in the best VN class, or alternatively an NP prosodic phrase in the best NP class.

Since we select the classes with the highest proportion of these prosodic phrases, our precision measure should be high. By contrast, since we look only at one class

[7] We have also used a more fine-grained measure, namely the "Pair-wise precision" measure, which measures the probability of selecting by chance a pair of phrases with the same category in a given class, or formally: $\text{PWP}(\text{Cl}) = \sum_i \left(\frac{|\text{Cat}_i \cap \text{Cl}|}{|\text{Cl}|}\right)^2$. This measure is called "precision" (Hatzivassiloglou & McKeown 1993) or "accuracy" (Chemla et al. 2009). The two measures, purity and PWP, are closely correlated. In our data, a purity measure of ⅔ approximately corresponds with a PWP measure of ½, which indicates that the probability of two randomly selected phrases belonging to the same category is ½.

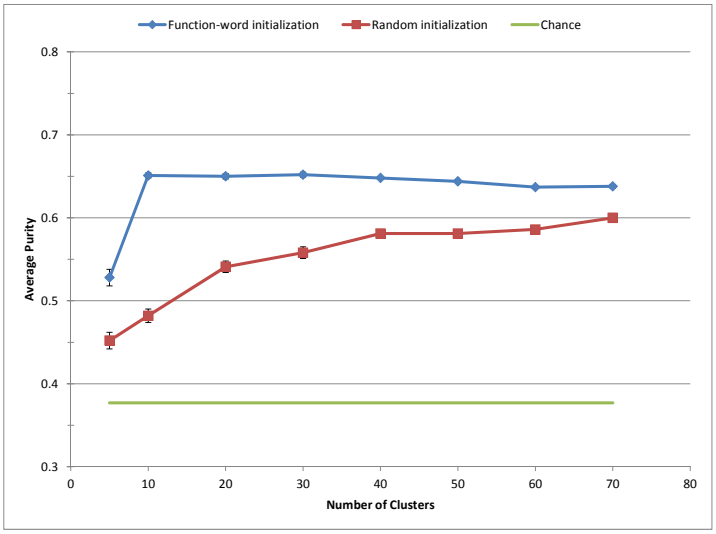

Figure 1: Average purity of the resulting classes as a function of the number of classes. The error bars (albeit being barely visible due to the relatively low variability) indicate standard errors of the mean calculated over the 10 sub-corpora.

for each category (out of our k classes), the recall measure will be very low, because each category (NP and VN) is spread out over many classes.

As a baseline, we can compare these measures to a chance distribution of the prosodic phrases into k clusters, which yields precision levels equal to the relative NP or VN proportions in the corpus, and recall levels which equal $1/k$.

2.3 Results

The average purity measure over the 10 sub-corpora as a function of the number of classes is given in Figure 1.

In general, we expect the average purity to grow with the number of classes (Strehl et al. 2000). This expectation is indeed borne out for the random-initialisation model. By contrast, this is not the case for the function word initialization model. Purity reaches a fixed level (about 0.65) with 10–30 classes, and does not increase with the addition of more classes. While the random-initialisation model shows a continuous increase in purity as a function of the number of classes, it remains

substantially lower than the purity of the function-word initialization model for any number of classes. Both models show a clear advantage over the chance baseline.

Importantly, the performance of our model decreases substantially when there are fewer than 10 classes. Using only 5 classes is insufficient – this is intuitively understandable given that none of the 5 most frequent L-words is a determiner (*tu, c', et, il, on/ça*, with the last L-word varying between the different sub-corpora). Indeed, the most common determiners (*de, le, un* depending on the sub-corpus) are ranked in positions 9–11 amongst the L-words.

Exploring the precision values of the best VN and NP classes (as defined in section 2.2.4) leads to similar conclusions. These values, together with the corresponding recall values, are presented in Figure 2. For both the VN category and the NP category, we see that the function-word initialisation model substantially outperforms the random-initialisation model in constructing precise VN or NP classes. As expected, recall is generally low and decreases with the number of classes. Both the random-initialisation model and function word initialisation model outperform chance-level baseline in all measures. For recall only, the random-initialisation model does slightly better than the function word initialisation model.

We can conclude that the L-word initialisation is highly beneficial even when considering a relatively low number of function words. Strikingly, purity and precision levels are consistently high across the whole spectrum. This holds despite the fact that the 10 most frequent function words only initially classify approximately 33% of the corpus, while the 70 most frequent L-words initially classify 70% of the corpus. This increase hardly has any effect on the average purity of the model (as illustrated in Figure 1) or on the best-classes precision levels.

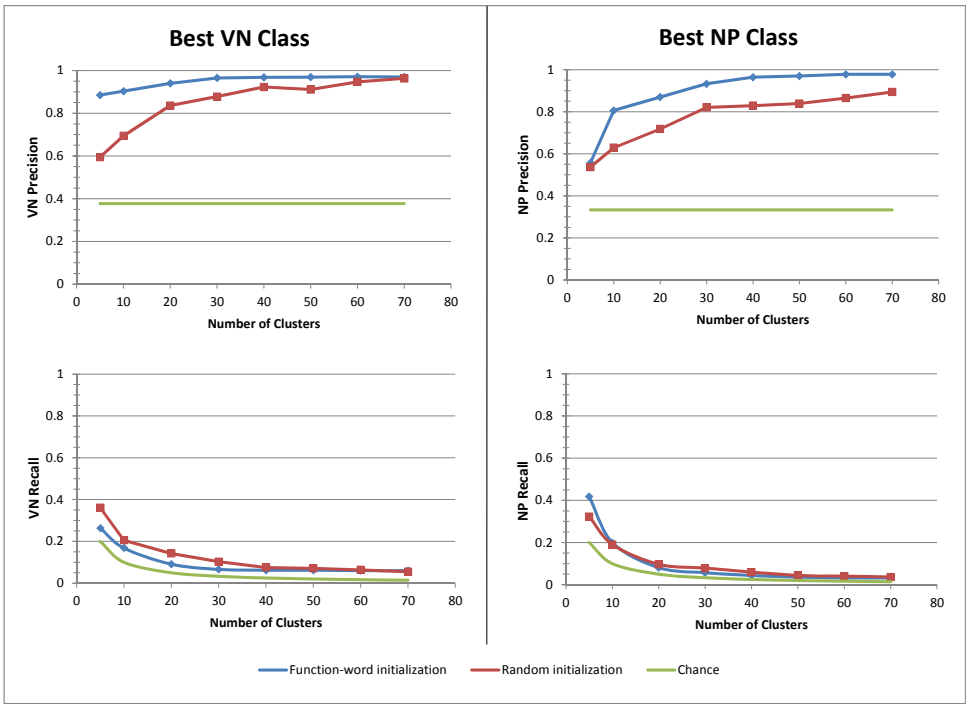

Figure 2: Precision and recall measures for the best VN and NP classes, as a function of the number of classes, in the different models. Standard error bars are not shown as they represent less than 0.1 points.

2.3.1 10 classes

We further analysed the behaviour of the model with 10 classes, the smallest number of classes that yielded good results for both VNs and NPs.

The results of the 10-class model on the entire corpus (rather than on a sub-corpus, as above) are presented in Figure 3. Table 4 gives the purity measure of the output classes, sorted by descending purity. The name of each class indicates the initial L-word from which it was created, while the growth column provides the ratio between the final and the initial class sizes (in other words, it provides an indication of how many phrases were added to the class in the EM learning process).

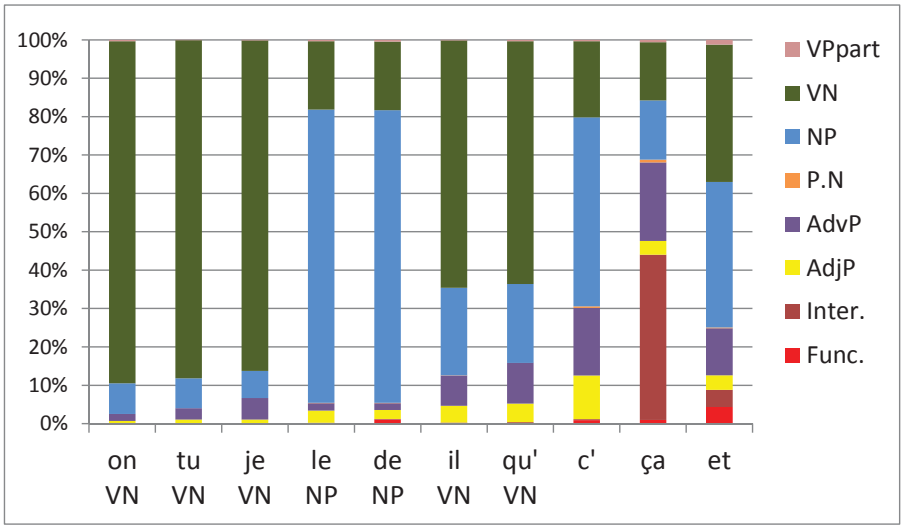

Figure 3: Results of the model with 10 classes. Every vertical bar represents a class, and the coloured regions describe the proportions of the different phrasal categories in each class. Note in particular the light blue which corresponds to NPs and the dark green which corresponds to VNs (the other categories are VPPart = Participal Verbal Phrases, P.N = Proper Nouns, AdvP = Adverbial Phrases, AdjP = Adjective Phrases, Inter. = Interjections, Func. = Functional words appearing alone). The labels in lower case correspond to the class name, while the labels in capitals are manually marked and signal classes with high purity values as well as their majority category (VN or NP): Thus, the classes labelled by the determiners *le* and *de* predict NPs while those labelled by pronouns (*tu, on, je, il*) as well as the relativizer *qu'* predict VNs.

Class		Purity	Growth
on	'we'	0.89	1.76
tu	'you (sg.)'	0.88	1.09
je	'I'	0.86	2.07
le	'the (m.)'	0.76	4.86
de	'of'	0.76	6.50
il	'he'	0.65	2.31
qu'	'which'	0.63	2.77
c'	'this (is)'	0.49	1.09
ça	'that'	0.43	8.72
et	'and'	0.38	4.58
Avg. Purity		0.67	
Rand. Avg. Purity		0.50 ± 0.01	

Table 4: Purity measures of the 10-class model, for each class and on average. The average purity measure is also given for the random initialization model, run 10 times.

The results show that 5 classes have an excellent purity measure, above 0.75, and that an additional two classes have good purity of 0.63–0.65. All these classes are good predictors of the NP or VN phrasal categories (see Figure 3). While the remaining 3 classes do not serve as predictors for these classes, they may still reveal some structure. For instance, the *ça* 'that' class captures 94% of all interjection phrases. The random initialisation model, on the other hand, provides on average only 2.2 ± 0.25 classes of purity larger than 0.60 (range: 1–3 in our test).

An interesting observation about the good classes is the negative correlation between the growth rate and the purity measures. Considering the 5 best classes together, we see that the higher the purity measure, the lower the growth level. The verbal classes *tu*, *on* and *je* in particular tend to have a very high purity rate (85% and above) and a relatively low growth rate. In other words, these classes are initially very good (i.e., the L-word initialisation provides homogeneous classes), but the algorithm succeeds only mildly in generalizing them to more data points (in contrast to the nominal classes).[8] This corroborates the hypothesis that relying on function words is highly informative for the classification process.

[8]The remarkably low growth rate of the *tu* class, only 9%, can likely be attributed to the use of an orthographically transcribed corpus, as 2nd person singular French present tense verb forms are written differently from other present tense singular verbal forms, thus making it difficult for the learning algorithm to generalize over orthographically different (but phonetically identical) verbal forms, such as *(tu) manges* '(you) eat' and *(je) mange* '(I) eat'.

2.4 Discussion

The present model tested the hypothesis that the edge words of a prosodic phrase provide useful information regarding the category of that phrase. The model was initialized with a limited number of classes that contained all prosodic phrases starting with a certain left-most word. The number of classes was varied parametrically from 5 to 70: the k most frequent left-most words were selected to build k classes. When the model contained 10 classes, all these words were function words, and the model exhibited a good average purity, of approximately 0.65, much higher than that of a model starting with random classes. Hence, relying on frequent words appearing at the left edges of prosodic phrases provided the model with useful information to categorize these phrases.

However, despite the quality of the classes produced by this method, it establishes classes based on single function words rather than generalized grammatical classes, such as VN or NP. As a result, VNs and NPs are distributed over several classes. A similar problem was identified in several distributional word-categorisation approaches (Mintz 2003; Chemla et al. 2009), and no straightforward way to merge classes post-hoc could be identified (but see Parisien et al. 2008). To overcome this problem, we chose to initialize the model with semantically-based classes.

3 Experiment 2

In the second experiment, we incorporate the possibility that early on, the child manages to learn the meaning of a few frequent nouns and verbs. These words often refer to concrete objects and agentive actions, and can thus constitute a seed for the prototypical "noun" and "verb" grammatical categories. For example, if the child knows the words *voiture* 'car' and *jouet* 'toy', she would be able to associate the two prosodic phrases *la voiture* 'the car' and *le jouet* 'the toy' to the same phrasal category related to physical objects, which we call NP. The idea that children group together words referring to physical objects on the one hand and words referring to actions on the other hand on the basis of semantics is in line with experimental data showing that children have a separate representation for agents and artifacts (for a review, see Carey 2009) and for causal actions (Saxe & Carey 2006). Indeed, such words are plausible candidates to be among the first words learnt by a child.[9]

[9] The idea that semantic classes can serve as a basis for syntactic classes is not new. Pinker (1984, 1989) proposed the *semantic bootstrapping hypothesis* in which children are hypothesized to group words into universal meaning categories, such as agent, patient, transitive verb, and so on. In his account, they would furthermore use innate linking rules to map such semantic categories onto the corresponding syntactic categories.

To model this initial semantic knowledge, we provide our clustering algorithm with a *semantic seed*, i.e. a short list of known words, which are explicitly associated with the VN and NP categories.

3.1 Material

We used the same input corpus, tagged with prosodic phrase information, as in Experiment 1. Additionally, a limited prior word knowledge, the semantic seed, is fed into our clustering algorithm. The size of the semantic seed is varied parametrically in order to observe how the size of the vocabulary can influence categorization. Following Brusini et al. (2011), we defined five semantic seeds ranging from a very small set of 6 nouns and 2 verbs (6N, 2V) to a larger set of 96 nouns and 32 verbs (96N, 32V). The *n* words chosen for the semantic seed correspond to the *n* most frequent nouns and verbs in the corpus.[10] For example, the smallest semantic seed (6N, 2V) contains the 6 most frequent nouns in the corpus, *doudou* 'stuffed toy', *bébé* 'baby', *livre* 'book', *chose* 'thing', *micro* 'microphone', *histoire* 'story', and the 2 most frequent verbs, *aller* 'go' and *faire* 'do'.

3.2 Method

As in Experiment 1, we used the Expectation-Maximization algorithm, with a modified initialization stage.

During initialization, the final word (or R-word) of each phrase was examined; if it was one of the known words from the semantic seed, the phrase was assigned to the V (Verbal) or N (Nominal) classes (according to the category of the known word). The remaining phrases were assigned to the U (Unknown) class (see Table 5 for examples). The first maximization phase was then conducted on the N and V phrases together with a similarly sized random sample of U phrases (so that the prior probability of the U class would be similar to those of the N and V classes). The remainder of the EM algorithm proceeded as before. Note that under this initialization condition there is no flexibility regarding the number of classes: there are exactly three (N, V or U). The percentage of phrases which were assigned to the N or V classes in the initialization phrase for each semantic seed level ranged between 4.5% (6N, 2V) to 23% (96N, 32V).

As in Experiment 1, the learning algorithm relies on the variables L_{-1}, L_0, L'_0, and R_0 (see Table 2).

[10] In order to construct the semantic seed, the full corpus was taken into consideration, including the one-word utterances which were excluded from the actual modeling.

Phrase		Assigned Category
vas-y	'go! (sg.)'	Unknown
tu vas apprendre	'you (sg.) will learn'	Unknown
je vais prendre	'I will take'	Verbal
le bain	'the bath'	Unknown
au bébé	'to the baby'	Nominal
et le crocodile	'and the crocodile'	Unknown

Table 5: Examples of prosodic phrases with their initial semantic category, with a semantic seed of 48 nouns and 16 verbs (48N, 16V).

3.2.1 Evaluation measures

Ideally, the resulting N and V classes should correspond to the NP and VN syntactic categories, respectively. Thus, we can easily calculate their precision and recall levels, as defined in equations 4 and 5 above.

As we have five levels of semantic seed used in the method, we can compare these measures across various levels of initial knowledge. Moreover, the results are compared to two baselines: first, we compare them to a uniform random clustering into 3 classes. Such classes will, by definition, have a recall level of $1/3$, and a precision level equivalent to the relative proportion of NPs and VNs in the corpus. These are the "chance" results. Second, we compare the results to a "zero-knowledge" model, which is modelled by running the random initialisation EM with 3 classes, which are *a posteriori* labelled as N or V classes in order to obtain maximal precision measures (specifically, among the classes with a majority of VNs, we take the one with the highest VN purity as the V cluster, and subsequently we take the class with the highest NP purity as the N cluster).

As in Experiment 1, we divided our corpus into 10 sub-corpora to estimate the variability of our results across different runs.

3.2.2 Discriminatory power

To investigate which variables are the most important ones in the learning process, we used a measure called "discriminatory power". For a given data point with its predicted category, we can calculate the additional contribution a variable adds to the likelihood in comparison to its average contribution when predicting other categories. When we average this measure over all data points, we get the discriminatory power. Formally, it can be computed as follows (i runs over the n data points, while j runs over the k classes):

$$\text{disc}(F) = \frac{1}{n}\Sigma_{i=1}^{n}\frac{1}{k}\Sigma_{j=1}^{k}\log P(F = f_i, \phi = \text{Cl}_{\text{best}}) - \log P(F = f_i, \phi = \text{Cl}_j) \quad (6)$$

A higher measure indicates a higher contribution of a variable. While we expect all variable measures to be positive, the absolute discriminatory value of a variable is not interpretable. We are rather interested in the relative magnitudes of these values.

3.3 Results

Figure 4 presents the precision measures as a function of the different sizes of the semantic seed, compared to the random baselines. Precision is very high, between 75% and 85%, and varies very little with the size of the semantic seed.

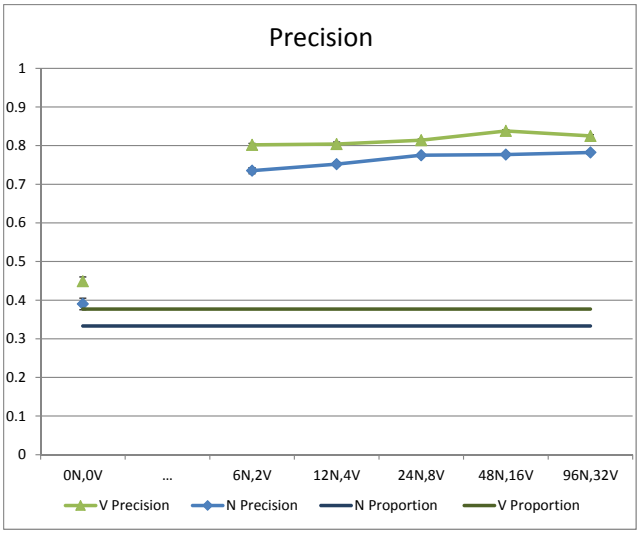

Figure 4: Precision as function of the semantic seed level (given as the number of known nouns (N) and verbs (V). The dark lines represent the chance baselines (related to the proportion of NPs and VNs in the corpus). The standard error is less than 0.01 for all conditions except for 0N,0V.

Figure 5 presents the recall measure for each semantic seed level. Again, recall

is much higher compared to the baseline recalls, and relatively stable across the variation in semantic seed size.

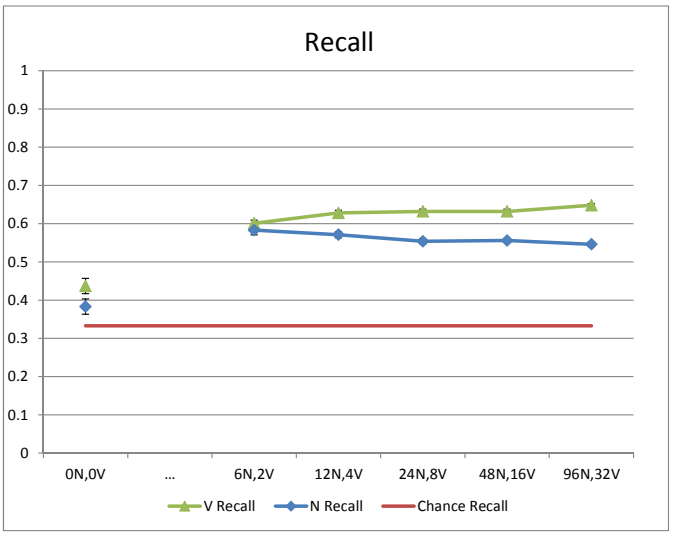

Figure 5: Recall as a function of the semantic seed size (given as number of known nouns and verbs). The red line represents the chance baseline. The standard error is at most 0.012 for all conditions except for 0N,0V.

The high precision levels are further illustrated in Figure 6, which presents the content of the classes obtained using the smallest seed. Considering that the smallest seed permits an initial classification of only about 4.5% of the prosodic phrases of the corpus, the final classes capture the verbal and nominal phrases extremely well, while other phrasal categories fall mainly in the U class. Using a larger semantic seed results in a similar picture.

Although the semantic seed model is based on an initial clustering according to content words (R-words), the classification process ultimately relies on the function words (L-words). Indeed, as Figure 7 shows, the most prominent variables for the classification are L_0 and L'_0. Note that while the R_0 variable becomes somewhat more prominent with the larger seeds (reflecting the larger initial semantic knowledge), it is still less important than the phrasal L-words. Not surprisingly, the L_{-1} variable, which represents the previous phrase's L-word, contributes least to the classification, in part because this variable is empty (thus not truly informative)

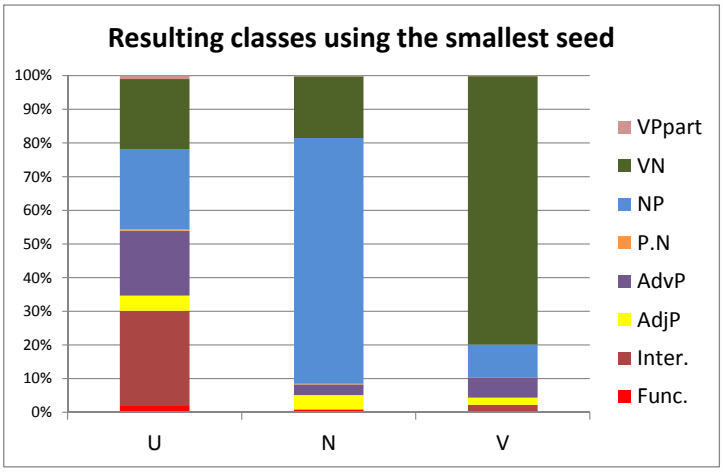

Figure 6: Results of the semantic seed model using the smallest seed (6N, 2V). Every vertical bar represents one class (Unknown, Nominal or Verbal), and the coloured regions indicate the proportions of the different phrasal categories in each class. Note in particular the light blue which corresponds to NPs and the dark green which corresponds to VNs.

whenever the first prosodic phrase of an utterance is considered. We can conclude that even though the model starts its classification on the basis of examining content words given in a semantic seed, it "learns" that a good classification should instead be based on the examination of function words. In other words, ultimately relying on function words leads to a more accurate classification of prosodic phrases.

Further support for this claim arises from examining how the model fares with one-word prosodic phrases, which for the large part consist only of a content word (such as an interjection, or an imperative verb). For these phrases the results are far from satisfactory: using the largest semantic seed, for example, the precision levels for these phrases is only of 46% (N) and 50% (V), with recall levels as low as 6% (N) and 37% (V). By contrast, for phrases with at least two words, which normally contain a function word, the precision levels are 79% (N) and 86% (V) with a recall level of 62% and 69%, respectively. Clearly, phrases containing more than one word are easier to classify correctly, and these phrases often contain a function word.

Looking closer at the performance on phrases of at least 2 words, we observe that the length of prosodic phrases differentially affects the quality of the N and V classes. The N class fares best with phrases of exactly 2 words – typically consisting of a determiner + noun – and captures longer nominal phrases less well. For

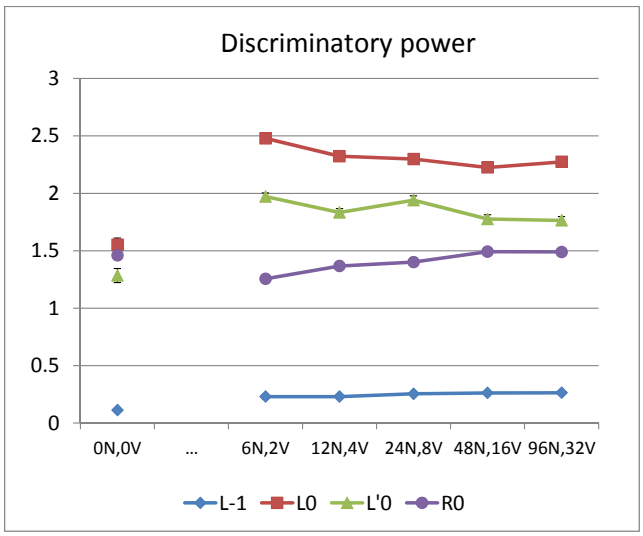

Figure 7: Discriminatory power of the variables used in the semantic seed EM algorithm. Since the variance is consistently low, standard error bars are too small to be visible in this figure.

example, the precision level of nominal phrases of at least 5 words is only 46% with 34% recall. This aligns with the predominant pattern of nominal phrases containing only 2 words. These short phrases appear to be beneficial for NP classification. By contrast, the V class is quite indifferent to phrasal length. For instance, verbal phrases of at least 5 words achieve an excellent precision level of 78% with recall of 72%. The word length distribution for V phrases is also more spread out (with comparable results for any multi-word phrases regardless of exact length). In other words, verbal phrases tend to have a larger scope than nominal phrases, and the model copes well with all these lengths.

3.4 Discussion

Initializing a model with semantically-based classes allows it to categorize initially-unclassified prosodic phrases with an excellent precision. In addition, the performance of the model remains remarkably stable with increases in the size of the semantic seed. This rather counter-intuitive result suggests that having a large vo-

cabulary is not necessary to initialize the categorization process: even a very small semantic seed (six nouns and two verbs) is sufficient. By assuming that the language learner can ground these semantically-based classes in her extra-linguistic experience – e.g. nouns typically refer to objects, and verbs to actions – we provide a plausible means of initializing syntactic categorization. In addition, the high contribution of the left-most words of the prosodic phrases to the categorization confirms the hypothesis that function words play a central role in the classification process.

4 Conclusions

In this paper, we presented two models that tested the role of phrasal prosody and edge words in the identification and classification of prosodic phrases. Both models successfully assigned syntactic labels to prosodic phrases, relying on phrasal prosody to delimit phrases, and their edge words to classify them. The two models differed only in the way classes are initially defined. The first model started out with a limited number of classes, each class being initially defined as containing all prosodic phrases starting with the same initial word. The model exhibited a good average purity level, much higher than a model starting with random classes. Thus, this model shows that relying on a small number of frequent function words is sufficient to create meaningful syntactic classes. A closer look at the behaviour of individual classes revealed that the model built a number of good VN and NP classes, as well as some classes that contained a mixture of categories. Thus, while this model confirms the intuition that paying attention to the left-most words of prosodic phrases is a good start for classifying them, it has the property that several different classes are constructed for each syntactic category.

To overcome this issue, the second model incorporated an additional piece of information, a semantic seed, that allowed the model to start with exactly three categories, one containing noun phrases, one containing verb phrases (or parts of verb phrases, corresponding to VNs), and the third one containing phrases of different categories. The size of the semantic seed was varied parametrically, from an extremely reduced semantic seed, consisting of only 6 known nouns and 2 known verbs, to a larger but still realistic one (96 nouns and 32 verbs). The results show that such an approach is highly successful: with as little initial knowledge as 4.5% of the phrases of the corpus, the algorithm manages to construct highly precise VN and NP classes, containing over 50% of the prosodic phrases in these categories. This excellent performance reveals two important features of our model. First, relying on the knowledge of a few frequent content words is sufficient for the emergence of abstract syntactic categories. Since these abstract categories (i.e., the VN and

the NP) are grounded in semantic experience (some of these words represent actions and some represent objects), no innate knowledge of the syntactic categories is *a priori* needed. Second, although the initial classes are based on content words from the semantic seed, the learning process relies ultimately on function words: The discriminatory power analysis showed that the most efficient variables are the left-most words – L_0 and L'_0 – which often correspond to function words. This can happen, since newly classified data points contribute to the learning of more structure. In other words, the knowledge of a few content words may allow the language learner to discover the role of function words.

This important role of function words is consistent with the infant literature. A number of experiments have shown that infants are sensitive to the function words of their language within their first year of life (Shafer et al. 1998; Shi et al. 2006; Hallé et al. 2008). In addition, 14- to 18-month-old children exploit function words to constrain lexical access to known words – for instance, they expect a noun after a determiner (Kedar et al. 2006; Zangl & Fernald 2007; Van Heugten & Johnson 2011; Cauvet et al. 2014). Crucially, when hearing unknown words, children of this age are able to infer the acceptable contexts for these unknown words. For instance, after hearing *the blick*, they would consider that *a blick* is possible but not *I blick* (for French: Shi & Melançon 2010; for German: Höhle et al. 2004). The present models provide a way in which infants can not only gather such information, but also use it in order to label prosodic phrases.

Our models rest on three assumptions. First, the language learner must have access to the boundaries of intermediate prosodic phrases. As we saw in the introduction, this hypothesis seems plausible given a wealth of experimental data showing that by the end of the first year of life, infants are not only sensitive to prosodic boundaries, but are also able to exploit them to constrain lexical access (Gout et al. 2004). Second, the models rest on the assumption that words placed at the edges play an important role: the left- and right-most words of a phrase are given special status. This assumption received experimental support from several studies: Words at edges are more salient, hence easier to segment from continuous speech (Cutler 1993; Shi et al. 1998; Seidl & Johnson 2006; Endress & Mehler 2009; Johnson et al. 2014). Third, we assume that children manage to learn and group together a few frequent and concrete nouns and verbs. This, too, is a plausible assumption given recent findings that show that infants know at least some nouns at 6 months (Bergelson & Swingley 2012; Tincoff & Jusczyk 2012) and possibly even some verbs at 10 months (the "abstract words" of Bergelson & Swingley 2013).

This final assumption does, however, warrant a note of caution. While we simplistically assume that children create two broad semantic categories of physical entities (corresponding to nouns) and actions (corresponding to verbs), several studies have suggested that infants represent distinct types of physical entities

differently, for instance agents vs. artifacts (see e.g., Carey 2009), or human vs. non-humans (Bonatti et al. 2002). It is thus quite possible that children may initially have more than two categories, and could hence distinguish between phrases referring to agents and phrases referring to artifacts (in addition to those referring to actions). For example, the nouns in our smallest semantic seed could represent (at least) two distinct categories of entities: agents (i.e., *bébé* 'baby') and artifacts (*livre* 'book', *chose* 'thing', *micro* 'microphone'). While more research is needed to better understand the early conceptual representations, our model suggests that the acquisition of syntax could be responsible for the merging of these separated classes by observing that agents and artifacts can (to a certain extent) occur in the same distributional environment.[11]

As for the second assumption, note that our models are currently built with a right-left asymmetry. In the first model, the most frequent left-most words are used to initially classify phrases, while in the second model the known right-most words are used for this initial categorization. This assumption is plausible, since several lines of experimental research suggest that infants know that frequent functional items typically occur either at the left or right edges of phrases, depending on the language (Gervain & Werker 2013; Hochmann 2013; Bernard & Gervain 2012; Hochmann et al. 2010; Gervain et al. 2008). However, this assumption is not crucial for the models: The first model could very well start with a symmetrical search of frequent items at both edges, while the second model could search the known content words at both edges. The model does not need to know in advance where content and function words typically occur. We would, however, need to make our variable set symmetrical, by adding for example an R'_0 variable to equate it with L'_0.

If the language learner has access to an approximate shallow syntactic structure consisting of labelled prosodic phrases, this can help her in two important ways. First, it may allow her to gain some insight into the syntactic structure of the language. This in turn may serve as an intermediate step toward a full understanding of its syntax. Second, this *syntactic skeleton* may enable the child to infer the meaning of unknown content words. The *syntactic bootstrapping* hypothesis proposes that syntactic structure provides additional constraints to the word learning inference problem (Gleitman 1990). Thus, language learners trying to figure out the meaning of a novel word, such as *blick*, perform better when they have access to the syntactic structure of the sentence. For instance, upon hearing a sentence such as *he blicks that the dog is angry*, listeners can infer that *blick* refers to a thought or communication verb (verbs that can take a whole proposition as complement; Gillette et al. 1998). Likewise, toddlers use sentence structure to predict that a verb used in

[11] We thank an anonymous reviewer for this interesting suggestion.

a transitive sentence has a causative meaning (Naigles 1990; Yuan & Fisher 2009). The language learner could also directly exploit the label of a prosodic phrase to constrain the meaning of some of its content words; for instance, a prosodic phrase labelled as a noun phrase should normally contain a noun (referring to an object), while a verbal nucleus should contain a verb (referring to an action). This may help the child learn the meaning of new words more easily.

Finally, our model illustrates the role of synergies in language acquisition. Knowledge of some lexical items (such as the semantic seed) permits the inference of syntactic categories, through the use of prosodic phrases and function words. Subsequently, knowledge of some syntactic categories enables the learner to enrich her vocabulary, which will further expand the child's syntactic knowledge. As such, our computational model provides a formalisation of the insights gained from the psycholinguistic literature to explain the mechanisms underlying early syntactic acquisition.

Acknowledgments

This work originated in the first author's master thesis of the *Master Parisien de Recherche en Informatique*. He is grateful for the *École Normale Superieure* for providing him the scholarship enabling him to pursue it. The research was furthermore supported by the French Ministry of Research, the French *Agence Nationale de la Recherche* (grants n° ANR-2010-BLAN-1901, ANR-13-APPR-0012, ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC and ANR-10-LABX-0083 EFL), the *Fondation de France*, the DGA (doctoral grant to ID), as well as the *Région Île-de-France*. We thank all colleagues who attended our talks on the subject, as well as Marieke van Heugten, Mark Johnson, Sonia Gharbi, Inka Leidig, Amy Perfors, Frans Plank, Valerie Shafer, and two anonymous reviewers for their suggestions and help.

References

Bergelson, Elika & Daniel Swingley. 2012. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences* 109(9). 3253–3258.

Bergelson, Elika & Daniel Swingley. 2013. The acquisition of abstract words by young infants. *Cognition* 127(3). 391–397.

Bernal, Savita, Ghislaine Dehaene-Lambertz, Séverine Millotte & Anne Christophe. 2010. Two-year-olds compute syntactic structure on-line. *Developmental science* 13(1). 69–76.

Bernard, Carline & Judit Gervain. 2012. Prosodic cues to word order: What level of representation? *Frontiers in Psychology* 3(451). http://dx.doi.org/10.3389/fpsyg.2012.00451.

Bonatti, Luca, Emmanuel Frot, Renate Zangl & Jacques Mehler. 2002. The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive psychology* 44(4). 388–426.

Brusini, Perrine, Pascal Amsili, Emmanuel Chemla & Anne Christophe. 2011. Learning to categorize nouns and verbs on the basis of a few known examples: A computational model relying on 2-word contexts. Paper presented at the Society for Research on Child Development Biennial Meeting, March 31 – April 2, Montreal, Canada.

Brusini, Perrine, Ghislaine Dehaene-Lambertz & Anne Christophe. 2009. Item-based or syntax? An ERP study of syntactic categorization in French-learning 2-year-olds. Paper presented at the 34th Boston University Conference on Language Acquisition, November 6–8.

Carey, Susan. 2009. *The origin of concepts*. Oxford University Press.

de Carvalho, Alex, Isabelle Dautriche & Anne Christophe. 2013. Three-year-olds use prosody online to constrain syntactic analysis. Paper presented at the 37th Boston University Conference on Language Development, November 2–4.

Cauvet, Elodie, Rita Limissuri, Séverine Millotte, Katrin Skoruppa, Dominique Cabrol & Anne Christophe. 2014. Function words constrain on-line recognition of verbs and nouns in French 18-month-olds. *Language Learning and Development* 10. 1–18.

Chemla, Emmanuel, Toben H. Mintz, Savita Bernal & Anne Christophe. 2009. Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science* 12(3). 396–406.

Christophe, Anne, Séverine Millotte, Savita Bernal & Jeff Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and speech* 51. 61–75.

Crabbé, Benoit & Marie Candito. 2008. Expériences d'analyse syntaxique statistique du français. *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2008), Avignon (France)*. http://www.atala.org/taln_archives/TALN/TALN-2008/taln-2008-long-017.html.

Cutler, Anne. 1993. Phonological cues to open-and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research* 22(2). 109–131.

Demuth, Katherine & Annie Tremblay. 2008. Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language* 35(1). 99–127.

Endress, Ansgar D. & Jacques Mehler. 2009. Primitive computations in speech processing. *The Quarterly Journal of Experimental Psychology* 62(11). 2187–2209.

Frank, Stella, Sharon Goldwater & Frank Keller. 2013. Adding sentence types to a model of syntactic category acquisition. *Topics in cognitive science*. 495–52.

Gerken, LouAnn, Peter W. Jusczyk & Denise R. Mandel. 1994. When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases. *Cognition* 51(3). 237–265.

Gervain, Judit, Marina Nespor, Reiko Mazuka, Ryota Horie & Jacques Mehler. 2008. Bootstrapping word order in prelexical infants: a Japanese–Italian cross-linguistic study. *Cognitive psychology* 57(1). 56–74.

Gervain, Judit & Janet F Werker. 2013. Prosody cues word order in 7-month-old bilingual infants. *Nature communications* 4. 1490.

Gillette, Jane, Henry Gleitman, Lila Gleitman & Anne Lederer. 1998. Human simulations of vocabulary learning. *IRCS Technical Reports Series*. `http://repository.upenn.edu/ircs_reports/71/` (26 February, 2013).

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1(1). 3–55.

Gomez, Rebecca L. & LouAnn Gerken. 1999. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition* 70(2). 109–135.

Gout, Ariel, Anne Christophe & James L. Morgan. 2004. Phonological phrase boundaries constrain lexical access II. infant data. *Journal of Memory and Language* 51(4). 548–567.

Hallé, Pierre A., Catherine Durand & Bénédicte de Boysson-Bardies. 2008. Do 11-month-old French infants process articles? *Language and Speech* 51(1-2). 23–44.

Hatzivassiloglou, Vasileios & Kathleen R. McKeown. 1993. Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. *Proceedings of the 31st annual meeting on Association for Computational Linguistics*. 172–182.

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

Hochmann, Jean-Rémy, Ansgar D. Endress & Jacques Mehler. 2010. Word frequency as a cue for identifying function words in infancy. *Cognition* 115(3). 444–57.

Höhle, Barbara, Michaela Schmitz, Lynn M. Santelmann & Jürgen Weissenborn. 2006. The recognition of discontinuous verbal dependencies by German 19-month-olds: Evidence for lexical and structural influences on children's early processing capacities. *Language Learning and Development* 2(4). 277–300.

Höhle, Barbara, Jürgen Weissenborn, Dorothea Kiefer, Antje Schulz & Michaela Schmitz. 2004. Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy* 5(3). 341–353.

Johnson, Elizabeth K. 2008. Infants use prosodically conditioned acoustic-phonetic cues to extract words from speech. *The Journal of the Acoustical Society of America* 123(6). EL144–EL148.

Johnson, Elizabeth K., Amanda Seidl & Michael D. Tyler. 2014. The edge factor in early word segmentation: Utterance-level prosody enables word form extraction by 6-month-olds. *PloS one* 9(1). e83546.

Jusczyk, Peter W., Elizabeth Hohne & Denise R. Mandel. 1995. Picking up regularities in the sound structure of the native language. In W. Strange (ed.), *Speech perception and linguistic experience: issues in cross-language speech research*, 91–119. Baltimore: York.

Kedar, Yarden, Marianella Casasola & Barbara Lust. 2006. Getting there faster: 18- and 24-month-old infants' use of function words to determine reference. *Child Development* 77(2). 325–338.

Lahiri, Aditi & Frans Plank. 2010. Phonological phrasing in Germanic: The judgement of history, confirmed through experiment. *Transactions of the Philological Society* 108(3). 370–398.

Lew-Williams, Casey, Bruna Pelucchi & Jenny R. Saffran. 2011. Isolated words enhance statistical language learning in infancy. *Developmental Science* 14(6). 1323–1329.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Marchetto, Erika & Luca L. Bonatti. 2013. Words and possible words in early language acquisition. *Cognitive psychology* 67(3). 130–150. PMID: 24041871.

Mehler, Jacques, Peter Jusczyk, Ghislaine Dehaene-Lambertz, Nilofar Halsted, Josiane Bertoncini & Claudine Amiel-Tison. 1988. A precursor of language acquisition in young infants. *Cognition* 29(2). 143–178.

Millotte, Séverine, Alice René, Roger Wales & Anne Christophe. 2008. Phonological phrase boundaries constrain the online syntactic analysis of spoken sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4). 874–85.

Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90(1). 91–117.

Mintz, Toben H., Elissa L. Newport & Thomas G. Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science* 26(4). 393–424.

Morgan, James L. 1986. *From simple input to complex grammar*. MIT Press Cambridge, MA.

Morgan, James L. & Katherine Demuth. 1996. *Signal to syntax: bootstrapping from speech to grammar in early acquisition*. Routledge.

Naigles, Letitia R. 1990. Children use syntax to learn verb meanings. *Journal of Child Language* 17(02). 357–374.

Nazzi, Thierry, Galina Iakimova, Josiane Bertoncini, Séverine Frédonie & Carmela Alcantara. 2006. Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language* 54(3). 283–299.

Nespor, Marina & Irene Vogel. 2007. *Prosodic phonology: with a new foreword*. Berlin: Walter de Gruyter.

Ngon, Céline, Andrew Martin, Emmanuel Dupoux, Dominique Cabrol, Michel Dutat & Sharon Peperkamp. 2013. (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science* 16(1). 24–34.

Oshima-Takane, Yuriko, Junko Ariyama, Tessei Kobayashi, Marina Katerelos & Diane Poulin-Dubois. 2011. Early verb learning in 20-month-old Japanese-speaking children. *Journal of child language* 38(03). 455–484.

Parise, Eugenio & Gergely Csibra. 2012. Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. *Psychological Science* 23. 728–733.

Parisien, Christopher., Afsaneh Fazly & Suzanne Stevenson. 2008. An incremental Bayesian model for learning syntactic categories. *Proceedings of the twelfth conference on computational natural language learning*. 89–96.

Pate, John K. & Sharon Goldwater. 2011. Unsupervised syntactic chunking with acoustic cues: Computational models for prosodic bootstrapping. *Proceedings of the 2nd ACL Workshop on Cognitive Modeling and Computational Linguistics*. 20–29.

Pedersen, Ted. 1998. *Learning probabilistic models of word sense disambiguation*. Southern Methodist University PhD thesis.

Pinker, Steven. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT press.

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

Saxe, Rebecca & Susan Carey. 2006. The perception of causality in infancy. *Acta Psychologica* 123(1-2). 144–165.

Schütze, Hinrich. 1995. Distributional part-of-speech tagging. *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. 141–148.

Seidl, Amanda & Elizabeth K. Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science* 9(6). 565–573.

Selkirk, Elisabeth. O. 1984. *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.

Shafer, Valerie L., David W. Shucard, Janet L. Shucard & LouAnn Gerken. 1998. An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language and Hearing Research* 41(4). 874–86.

Shi, Rushen, Anne Cutler, Janet Werker & Marisa Cruickshank. 2006. Frequency and form as determinants of functor sensitivity in English-acquiring infants. *The Journal of the Acoustical Society of America* 119(6). 61–7.

Shi, Rushen & Andréane Melançon. 2010. Syntactic categorization in French-learning infants. *Infancy* 15(5). 517–533.

Shi, Rushen, James L. Morgan & Paul Allopenna. 1998. Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of child language* 25(01). 169–201.

Shukla, Mohinish, Marina Nespor & Jacques Mehler. 2007. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology* 54(1). 1–32.

St. Clair, Michelle C., Padraic Monaghan & Morten H. Christiansen. 2010. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition* 116(3). 341–360.

Strehl, Alexander, Joydeep Ghosh & Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*. 58–64.

Tincoff, Ruth & Peter W. Jusczyk. 2012. Six-month-olds comprehend words that refer to parts of the body. *Infancy* 17(4). 432–444.

Van Heugten, Marieke & Elizabeth K. Johnson. 2010. Linking infants' distributional learning abilities to natural language acquisition. *Journal of Memory and Language* 63(2). 197–209.

Van Heugten, Marieke & Elizabeth K. Johnson. 2011. Gender-marked determiners help Dutch learners' word recognition when gender information itself does not. *Journal of Child Language* 38(01). 87–100.

Yuan, Sylvia & Cynthia Fisher. 2009. "Really? She Blicked the Baby?": Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science* 20(5). 619–626.

Zangl, Renate & Anne Fernald. 2007. Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development* 3(3). 199–231.