Developmental Science

Developmental Science 12:3 (2009), pp 396–406 DOI: 10.1111/j.1467-7687.2009.00825.x

SPECIAL SECTION: COMPUTATIONAL PRINCIPLES OF LANGUAGE ACQUISITION

Categorizing words using 'frequent frames': what cross-linguistic analyses reveal about distributional acquisition strategies

Emmanuel Chemla,[1] Toben H. Mintz,[2] Savita Bernal[1] and Anne Christophe[1,3]

1. Laboratoire de Sciences Cognitives et Psycholinguistique, DEC-ENS/EHESS, CNRS, Paris, France
2. University of Southern California, Los Angeles, USA
3. Maternité Port-Royal, AP-HP, Paris, France

Abstract

Mintz (2003) described a distributional environment called a frame*, defined as the co-occurrence of two context words with one intervening target word. Analyses of English child-directed speech showed that words that fell within any frequently occurring frame consistently belonged to the same grammatical category (e.g. noun, verb, adjective, etc.). In this paper, we first generalize this result to French, a language in which the function word system allows patterns that are potentially detrimental to a frame-based analysis procedure. Second, we show that the discontinuity of the chosen environments (i.e. the fact that target words are* framed *by the context words) is crucial for the mechanism to be efficient. This property might be relevant for any computational approach to grammatical categorization. Finally, we investigate a recursive application of the procedure and observe that the categorization is paradoxically worse when context elements are categories rather than actual lexical items. Item-specificity is thus also a core computational principle for this type of algorithm. Our analysis, along with results from behavioural studies (Gómez, 2002; Gómez and Maye, 2005; Mintz, 2006), provides strong support for frames as a basis for the acquisition of grammatical categories by infants. Discontinuity and item-specificity appear to be crucial features.*

Introduction

Grammatical categories, such as noun, verb and adjective, are the building blocks of linguistic structure. Identifying the categories of words allows infants and young children to learn about the syntactic properties of their language. Thus, understanding how infants and young children learn the categories of words in their language is crucial for any theory of language acquisition. In addition, knowledge of word categories and the syntactic structures in which they participate may aid learners in acquiring word meaning (Gleitman, 1990; Gleitman, Cassidy, Nappa, Papafragou & Trueswell, 2005; Landau & Gleitman, 1985).

In their introductory text on syntactic theory, Koopman, Sportiche and Stabler (2003) describe the main concepts that allow linguists to posit syntactic categories: 'a category is a set of expressions that all "behave the same way" in language. And the fundamental evidence for claims about how a word behaves is the distribution of words in the language: where can they appear, and where would they produce nonsense, or some other kind of deviance.' These observations were at the core of the notions behind structural linguistics in the early 20th century (Bloomfield, 1933; Harris, 1951), namely, that form-class categories were defined by co-occurrence privileges. Maratsos and Chalkley (1980) advanced the proposal that children may use distributional information of this type as a primary basis for categorizing words. In the past decade, a number of studies have investigated how useful purely distributional information might be to young children in initially forming categories of words (Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport & Bever, 2002; Redington, Chater & Finch, 1998). Employing a variety of categorization procedures, these investigations demonstrated that lexical co-occurrence patterns in child-directed speech could provide a robust source of information for children to use to correctly categorize nouns and verbs, and to some degree other form-class categories as well.

Address for correspondence: Emmanuel Chemla, Laboratoire de Sciences Cognitives et Psycholinguistique, 29 rue d'Ulm, 75005 Paris, France; e-mail: emmanuel.chemla@ens.fr

Invited target article for B. McMurray and G. Hollich (Eds.) 'Core computational principles of language acquisition: Can statistical learning do the job?' Special section in *Developmental Science*.

One challenge in forming categories from distributional cues is to establish an efficient balance between the detection of the especially informative contexts and the rejection of the potentially misleading ones. For example, in (1), that *cat* and *mat* both occur after *the* suggests that the two words belong to the same category. However, applying this very same reasoning to example (2) would lead one to conclude that *large* and *mat* belong to the same category (see Pinker, 1987, for related arguments).

(1) *the cat is on the mat*
(2) *the large cat is on the mat*

To address the problem of the variability of informative distributional contexts, the procedures developed by Redington *et al*. (1998) and Mintz *et al*. (2002) took into account the entire range of contexts a word occurred in, and essentially classified words based on their distributional profiles across entire corpora. Whereas in (1) and (2) the adjective *large* shares a preceding context with *cat* and *mat*, in other utterances it occurs in environments that would not be shared with nouns, as in (3). Many misclassifications that would occur if only individual occurrences of a target word were considered turned out not to result when taking into account the statistical information about the frequency of a target word occurring across different contexts.[1]

(3) *the cat on the mat is large*

Mintz (2003) took a different approach. Rather than starting with target words and tallying the entire range of contexts in which they occur, the basis for his categorization is a particular type of context that he called *frequent frames*, defined as two words that frequently co-occur in a corpus with exactly one word intervening. (Schematically, we indicate a frame as [A x B], with A and B referring to the co-occurring words and x representing the position of the target words.) For example, in (3), [*the* x *on*] is a frame that contains the word *cat*; it so happens that in the English child-directed corpora investigated by Mintz (2003), this frame contained exclusively nouns, leading to a virtually error-free grouping together of nouns. Examining many frames in child-directed speech, Mintz demonstrated that, in English, frames that occur frequently contain intervening words that almost exclusively belong to the same grammatical category. He proposed that frequent frames could be the basis for children's initial lexical categories.

One critical aspect of frequent frames is that the framing words – for example, *the* and *on* in the example above – must frequently co-occur. Arguably, co-occurrences that are frequent are not accidental (as infrequent co-occurrences might be), but rather arise from some kind of constraint in the language. In particular, structural constraints governed by the grammar could give rise to this kind of co-occurrence regularity. It is not surprising, then, that the words categorized by a given frequent frame play a similar structural role in the grammar, that is, they belong to the same category.

Thus, in the frequent-frames approach, the important computational work involves identifying the frequent frames. Once identified, categorization is simply a matter of grouping together the words that intervene in a given frequent frame throughout a corpus. In contrast, in other approaches (Mintz *et al*., 2002; Redington *et al*., 1998) the crucial computations involved tracking the statistical profile of each of the most frequent words with respect to all the contexts in which it occurs, and comparing the profiles of each word with all the other words. Thus, an advantage of the frequent-frames categorization process is that, once a set of frequent frames has been identified, a single occurrence of an uncategorized word in a frequent frame would be sufficient for categorization. Moreover, it is computationally simpler, in that fewer total contexts are involved in analysing a corpus.

In addition to research showing the informativeness and computational efficiency of frequent frames (in English), several behavioural studies suggest that infants attend to frame-like patterns and may use them to categorize novel words. For example, Gómez (2002) showed that sufficient variability in intervening items allowed 18-month-old infants to detect frame-like discontinuous regularities, and Gómez and Maye (2005) showed that this ability was already detectable in 15-month-olds. This suggests that young infants have the resources required to detect frequent frames. Second, Mintz (2006) showed that English-learning 12-month-olds categorize together novel words when they occur within actual frequent frames (e.g. infants categorized *bist* and *lonk* together when they heard both words used in the [*you* x *the*] frequent frame).

Although frequent frames have been shown to be a simple yet robust source of lexical category information, the analyses have been limited to English. One goal of the present paper was to start to test the validity of frequent frames cross-linguistically. To this end, in Experiment 1 we tested the validity of frequent frames in French, a language that presents several potentially problematic features for the frame-based procedure.

An additional goal was to characterize the core computational principles that make frequent frames such robust environments for categorization. To this end, in Experiment 2, in both French and English, we compared frames with other types of contexts that are at first sight very similar to frames in terms of their intrinsic informational content and structure: [A B x] and [x A B]. Interestingly, despite the similarity of these contexts to frames, they yielded much poorer categorization. The results of this experiment suggest that co-occurring context elements must *frame* a target word.

[1] Mintz *et al*. and Redington *et al*. also incorporated more distributional positions into their analysis than just the immediately preceding word; for example, the following word, words that were two positions before or after, etc. However, the addition of contexts does not, *a priori*, make the potential for misclassifications go away.

Finally, in Experiment 3 we investigated the consequences of a recursive application of this frame-based procedure, again with French and English corpora. Specifically, we performed an initial analysis to derive frame-based categories, and then reanalysed the corpus defining frames based on the *categories* of words derived in the initial analysis. A somewhat counterintuitive finding was that the recursive application of the frame-based procedure resulted in relatively poor categorization. This finding suggests that computation based on specific items – actual words as opposed to categories – is a core principle in categorizing words, at least initially.

Experiment 1: French frequent frames

This first experiment investigates the viability of the frequent-frames proposal for French. Several features of the language suggest that frequent frames may be less efficient in French than in English. For example, English frequent frames rely heavily on closed-class words, such as determiners, pronouns and prepositions. In French, there is homophony between clitic object pronouns and determiners, *le/la/les*, which could potentially give rise to erroneous generalizations. For instance, *la* in 'la pomme' (*the apple*) is an article and precedes a noun, whereas *la* in 'je la mange' (*I eat it*) is a clitic object pronoun and precedes a verb. There are also a greater number of determiners, which could result in less comprehensive categories. For instance, French has three different definite determiners, *le/la/les*, varying in gender and number, that all translate into *the* in English. Finally, constructions involving object clitics in French exclude many robust English frame environments; for example [*I* x *it*], a powerful verb-detecting frame in English, translates into [*je le/la* x] in French, which is not a frame. Do French frequent frames nevertheless provide robust category information, as in English?

Material

Input corpus

The analysis was carried out over the Champaud (1988) French corpus from the CHILDES database (MacWhinney, 2000). This corpus is a transcription of free interactions between Grégoire (whose age ranges between 1;9.18 and 2;5.27) and his mother. Only utterances of the mother were analysed, comprising 2,006 sentences. This is the largest sample available to us for which the age of the child is in the range of the English corpora analysed by Mintz (2003). Those corpora contained on average 17,199 child-directed utterances, so the present corpus is an order of magnitude smaller. Thus, this experiment provides a test of the robustness of the frequent-frames approach, in addition to a test of the cross-linguistic viability.

The corpus was minimally treated before the distributional analysis procedure was performed: all punctuation and special CHILDES transcription codes were removed.

Table 1 *Distribution of the syntactic categories across the French corpus investigated*

Categories	No. types	% corpus	No. tokens	% corpus
wh-word	3	0.1	12	0
Interjection	16	0.7	226	1.2
Conjunction	20	0.8	954	5.1
Adjective	281	12.3	1132	6
Preposition	29	1.2	1223	6.5
Determiner	12	0.5	1515	8.1
Adverb	111	4.8	1898	10.2
Verb	789	34.7	4253	22.8
Noun	953	41.9	2901	15.5
Pronoun	61	2.6	4485	24.1
Total	2275		1,8599	

Tagging the corpus

We ran Cordial Analyseur over the corpus. This software, developed by Synapse Développement (http://www.synapse-fr.com), maps each instance of a word with its syntactic category relying on supervised lexical and statistical strategies. The resulting categorization of words was used as the standard for evaluating the categories derived using frequent frames.

Syntactic categories included: *noun*, *pronoun*, *verb*, *adjective*, *preposition*, *adverb*, *determiner*, *wh-word*, *conjunction* and *interjection*.[2] (The word *group* designates a set of words that are grouped together by the distributional analysis.) Table 1 provides details about the distribution of the categories across the corpus. In Table 1 and throughout this paper, we use *type* to refer to a particular word and *token* to refer to a specific instance of the word in the corpus.

Method

Distributional analysis procedure

Every frame was systematically analysed from the corpus, where a *frame* is an ordered pair of words that occurs in the corpus with an intervening target word (schematically: [A x B], where the target, x, varies). Utterance boundaries were not treated as framing elements, nor could frames cross utterance boundaries. The frequency of each frame was recorded, and the intervening words for a given frame were treated as a frame-based category. The frame-based categories were then evaluated to determine the degree to which they matched actual linguistic categories, such as noun and verb.

[2] Another set of analyses relied on a set of categories in which pronouns and nouns were collapsed into a single category, as in previous distributional investigations; results were very similar.

Evaluation measures

In order to obtain a standard measure of categorization success, comparable to prior studies, we computed *accuracy* and *completeness* scores. These measures have been widely used for reporting in other studies (e.g. Cartwright & Brent, 1997; Mintz, 2003; Mintz *et al.*, 2002; Redington *et al.*, 1998). Pairs of analysed words were labelled as Hit, False Alarm or Miss. A Hit was recorded when two items in the same group came from the same category (i.e. they were correctly grouped together). A False Alarm was recorded when two items in the same group came from different categories (i.e. they were incorrectly grouped together). A Miss was recorded when two items from the same category ended up in different groups (i.e. they should have been grouped together but were not).

As equation 1a shows, accuracy measures the proportion of Hits to the number of Hits plus False Alarms (i.e. the proportion of all words grouped together that were correctly grouped together). Completeness measures the degree to which the analysis puts together words that belong to the same category (as equation 1b shows, it is calculated as the proportion of Hits to the number of Hits plus Misses). Both measures range from 0 to 1, with a value of 1 when the categorization is perfect.

$$Accuracy = \frac{Hits}{Hits + FalseAlarms}, \tag{1a}$$

$$Completeness = \frac{Hits}{Hits + Misses}. \tag{1b}$$

Two scoring conditions were available for each measure, depending on whether word *tokens* or word *types* were considered. By default, we will report results for the *type* condition.

Departing from Mintz (2003), we elected first to evaluate *all* frames and their corresponding word categories, even if the frames were relatively infrequent. In subsequent analyses, like Mintz, we then established a frequency threshold to select a set of frequent frames and corresponding word categories to evaluate.

Comparison to chance categorization

For each set of frame-based categories, 1000 sets of random word categories were arbitrarily assembled from the corpus; these random categories were matched in size and number with the actual frame-based categories they were to be compared with. The mean accuracy and completeness obtained from these 1000 trials provided a baseline against which to compare the actual results and were used to compute significance levels, using the 'bootstrap' or 'Monte Carlo' method. For instance, if only 2 out of 1000 trials matched or exceeded the score obtained by the algorithm, that score was said to significantly exceed chance level, with the probability of a chance result being $p = .002$ (2 out of 1000).

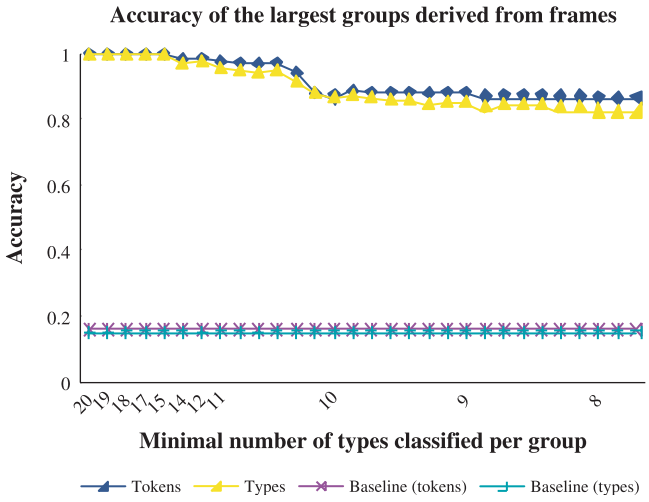

Figure 1 *Accuracy for the largest groups obtained from frames. From left to right, accuracy is reported for the largest group, the set composed by the two largest groups, the set composed by the three largest groups and so on; numbers on the horizontal axis represent the minimal number of types classified for each group included in the result.*

Results

Global results

Frame-based categories contained mainly nouns and verbs. Specifically, in the largest frame-based categories – the 20 categories containing at least 10 different types – 48% of the types were nouns, and 41% were verbs. This is not a surprise, as nouns and verbs constitute 75.6% of the types in the corpus. Interestingly, the frame statistics are similar even if calculated in terms of tokens: although nouns and verbs together constitute only 38.3% of the tokens in the whole corpus, 37% of the tokens captured by the frames were nouns and 46% were verbs.

Rather than applying an *a priori* threshold to select a set of frequent frames to evaluate (Mintz, 2003), we first evaluated performance iteratively on a successively larger number of frame-based categories. That is, we first assessed categorization by evaluating the largest frame-based category (by type), then the two largest categories, then the three largest categories, etc. Essentially, at each successive iteration we relaxed the criterion for determining whether or not a given frame defined a category.[3] Figure 1 reports the accuracy for such sets of groups: from left to right the number of groups increases as the criterion for category size is relaxed. Figure 2 reports completeness

[3] Although category size is not directly based on frame frequency, the number of types occurring within a frame is correlated with the frequency of the frame. We chose to organize the presentation of the evaluation metrics by category size simply for clarity. Below, we analyse categorization using a specific frame-frequency threshold.

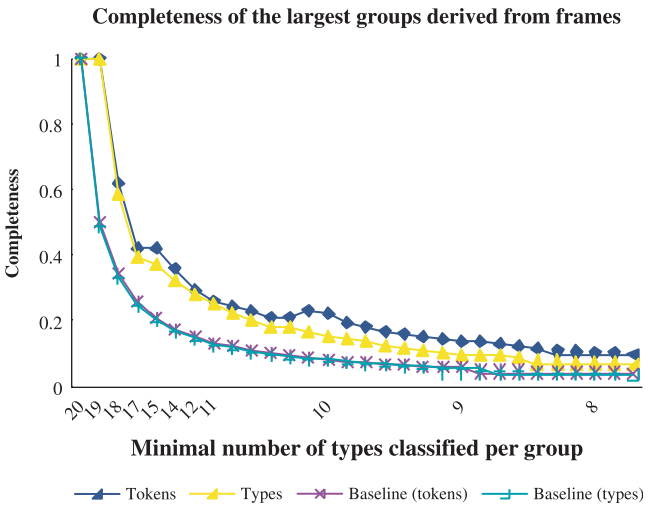

Figure 2 *Completeness for the largest groups obtained from frames (see Figure 1 for details about the groups selected).*

for the same sets of groups (the set with only one group being trivially complete).

Accuracy remains at ceiling for groups classifying 15 different types or more and is overall significantly better than chance for every set of groups represented here ($p < .01$, see section 'Comparison to chance categorization' above for an explanation of how this probability was estimated). Behavioural data from Gómez (2002) suggest that internal variability makes distant dependences easier to recover for 18-month-old infants. As a result, these groups, classifying many different types, may well also correspond to the most salient frames psychologically.

Unsurprisingly, completeness decreases quickly as the number of groups taken into account increases: for instance, because several frames consistently capture nouns, each of them leads to an independent group of nouns and completeness suffers from this situation (e.g. many pairs of items coming from different groups happen to be pairs of nouns, which adds to the number of misses). Nevertheless, completeness is overall significantly above chance ($p < .01$).

Example of frequent frames

The previous analysis showed accuracy and completeness over a range of possible frame-based categorizations, as a function of the size of the resulting categories. We then analysed a more limited set of frames based on a frame-frequency criterion, similar to in Mintz (2003). We selected frames that both grouped together more than .5% of the types present in the corpus (i.e. 11 types) and accounted for more than .1% of the tokens (i.e. occurred more than 18 times). This led to the six frequent frames described in Table 2; overall they classified 172 tokens from 99 types, which accounted for 2236 tokens of the corpus. Thus, this sample of frames accounted for 31% of the nouns and verbs of the corpus (12% of all the tokens).

Table 2 *Groups obtained from the most frequent frames. Numbers in parenthesis indicate the number of times each type occurs within the given frame (when it occurs more than once). The first three frames classify nouns. Each starts with one of the forms of the definite determiner (i.e.* la*: feminine singular;* le*: masculine singular;* les*: plural and unspecified for gender) and ends with the genitive particle* de *(of). The last three frames classify verbs. They involve the split French negation, which is in full form '*ne *Verb* pas*' (last frame);* ne *is reduced to* n' *when the verb starts with a vowel (frame 5) and can be dropped entirely in colloquial speech (frame 4 translates into [*he *x* not*])*

[*la* x *de*] (35 occurrences, 20 types): cabine(2), casquette, coupe, couronne(2), disposition, fin(3), langue(3), main, maison(4), maman, photo, place(2), pomme, porte(2), salle(3), soupe(2), tringle, trompe, télé(2), tête;
[*les* x *de*] (21 occurrences, 17 types): adverbes, aiguilles, bras, casquettes, chaussons, chaussures(2), cheveux, feutres, lâcher, oreilles, palmes(2), photos, pommes(3), pyjamas, talents, yeux, échanges;
[*le* x *de*] (20 occurrences, 18 types): bas, bateau, cadre, chalet, champ, chapeau(2), cran, cri, discours, droit, fils, hangar, jardin, pied, processus, puzzle(2), sens, stade.
[*il* x *pas*] (40 occurrences, 16 types): a(7), connaît, dit(2), est(14), exprime, fallait(2), faut(3), peut, pousse, produisait, tombera, utilise, va(3), voit, voulait;
[*n'* x *pas*] (29 occurrences, 12 types): a(6), aimais, aimes, allait, as(3), avaient, avait(3), emmènerait, est(9), ira, iras, était;
[*ne* x *pas*] (27 occurrences, 20 types): connaissent, coupe, dirai, distingue, déchire, fais(2), frappes, jette, jettes, montes(2), peux(2), porte, recommence(2), reproduisent, sais(2), sont(2), tire, vas(2), veux(2).

This classification was perfectly accurate: each frame selected words from only one category. Although framing elements were function words, they categorized target words: specifically, three groups of nouns, and three groups of verbs. Token and type completeness was .34 and .33, respectively (see Table 4). Again, these scores are significantly above chance level ($p < .01$). Interestingly, each group of nouns formed by these frames corresponds to a different subcategory (feminine, plural and masculine nouns respectively); this is not surprising given that the frames categorizing nouns included determiners, which, in French, mark grammatical gender. Although it may be tempting to conjecture that frames could capture finer-grained categories more broadly, none of the frames capturing verbs favoured a particular sub-type of verb (such as transitive or intransitive).

Discussion

This analysis extends previous results described by Mintz (2003) for English child-directed speech: here we showed that frequent frames delimit accurate groups of content words in French. It is striking that these results hold for French, which has a more varied and ambiguous system of function words than English. Recall the

Table 3 *Accuracy and completeness for groups derived from frequent frames, back and front contexts from the French and the English corpora*

Environments		Frames [A x B] (Expt 1)		Front [A B x] (Expt 2)		Back [x A B] (Expt 2)	
Scoring condition		Types	Tokens	Types	Tokens	Types	Tokens
French corpus	Accuracy	1.	1.	.30	.30	.25	.35
	(Baseline)	(.13)	(.13)	(.17)	(.16)	(.17)	(.16)
	Completeness	.33	.34	.050	.10	.057	.11
	(Baseline)	(.16)	(.16)	(.072)	(.093)	(.067)	(.090)
English corpus	Accuracy	.90	.95	.52	.46	.29	.41
	(Baseline)	(.18)	(.18)	(.20)	(.18)	(.19)	(.18)
	Completeness	.047	.057	.056	.071	.046	.060
	(Baseline)	(.031)	(.038)	(.048)	(.062)	(.050)	(.075)

Table 4 *Accuracy and completeness for groups derived from frequent frames and their different recursive applications discussed from the French and English corpora*

Environments		Frames [A x B] (Expt 1)		Recursive frames (Expt 3)		Asymptotic step (Expt 3)	
Scoring condition		Types	Tokens	Types	Tokens	Types	Tokens
French corpus	Accuracy	1.	1.	.37	.60	.42	.46
	(Baseline)	(.13)	(.13)	(.16)	(.16)	(.20)	(.16)
	Completeness	.33	.34	.048	.084	.11	.16
	(Baseline)	(.16)	(.16)	(.043)	(.048)	(.057)	(.067)
English corpus	Accuracy	.90	.95	.25	.53	.46	.58
	(Baseline)	(.18)	(.18)	(.21)	(.18)	(.21)	(.18)
	Completeness	.047	.057	.032	.060	.056	.10
	(Baseline)	(.031)	(.038)	(.028)	(.035)	(.029)	(.041)

potential difficulty introduced by the fact that all definite determiners are, in French, homophonous to clitic object pronouns (*le*/*la*/*les*). Whereas determiners typically precede nouns or adjectives, object clitics typically precede verbs. To estimate the extent of the problem, we tallied the number of times that *le*/*la*/*les* occurred as determiners versus clitic objects: we observed that *le*, *la* and *les* occurred as determiners 802 times in our corpus and 145 times as object clitics. These numbers indicate the ambiguity faced by any simple mechanism attempting to categorize a content word on the basis of the immediately preceding function word. In contrast, the frequent-frames mechanism did not suffer at all from the ambiguity: *le*, *la* and *les* appeared as left-framing elements of the three most frequent noun-detecting frames, and those frames did not contain verbs (see Bernal *et al.*, in press, for related experimental results with 2-year-olds). Thus, the additional constraint of co-occurrence with the right-framing element efficiently disambiguated the ambiguous function words. Note that, in principle, a verb *could* occur in many successful noun-detecting frames, such as [*le* x *de*] in (4).

(4) Je [le vois de] mon balcon.
I [it-clitic see from] my balcony.
I see it from my balcony.

However, this type of construction, though grammatical, is absent from the corpus and arguably extremely rare in child-directed speech. This illustrates the strength of the co-occurrence restriction imposed by the frame structure that reduces dramatically the syntactic ambiguity of individual function words.

In addition, a significant portion of the corpus was accounted for by a very restricted sample of frames. This result strongly argues in favour of the plausibility of the mechanism very early on: identifying even a few of the most frequent frames may allow infants to categorize many of the nouns and verbs they encounter.

One explanation of why words from categories other than noun and verb were not captured by frequent frames here is that the corpus we analysed was relatively small. For example, the corpora analysed in Mintz (2003) contained, on average, over 14,000 child-directed utterances; some frames in those analyses contained other classes, such as adjectives, determiners and prepositions. We expect that if larger corpora are analysed, frame-based categories will successfully group together words from other syntactic categories as well.

Overall, then, the potentially problematic characteristics of French do not appear to be problematic in practice for a frequent-frames approach to early word categorization.

Experiment 2: Discontinuity

In this experiment, results from the frequent-frames analysis are compared with results from analysing similar environments: *front contexts* and *back contexts*. Front contexts are ordered pairs of contiguous words that

categorize following words [A B x], and back contexts are ordered pairs of contiguous words that categorize preceding words [x A B]. In many respects, these environments are similar to frames: they all involve trigrams in which the co-occurrence of two words serves as the context for categorizing the third. The main difference is that frames are discontinuous and categorize intervening words, whereas front and back contexts feature two contiguous contextual elements. This experiment thus provides an indication of the importance of the *particular* context selected by a frame, versus contexts that are formally and computationally similar on all other dimensions. As we mentioned earlier, some English frames found by Mintz (2003) translate into one of these new environments: for example, [*I* x *it*] translates into [*je le/la* x] in French. It could be that front contexts are just as useful in French as corresponding frame contexts in English. Alternatively, it could be that the discontinuity inherent in the frame context is important for capturing category regularities cross-linguistically. Experiment 2 addresses this question, using both French and English corpora.

Material and method

Procedures and analysis methods similar to those in Experiment 1 were used here. The primary difference was that, in Experiment 2, categories were formed on the basis of two new types of environments – front contexts of type [A B x] and back contexts of type [x A B]. The analysis was run over the same French corpus as in Experiment 1, and over the English corpus, Peter, from the CHILDES database (files Peter01.cha to Peter12.cha; Bloom, Hood & Lightbown, 1974; Bloom, Lightbown & Hood, 1975). This corpus is one of the largest corpora investigated in Mintz (2003), and provides syntactic labelling for evaluating the analysis outcome.

Results

Groups resulting from the alternative contexts contain many more types than frame-based groups: for instance, in French, 19 front contexts and 14 back contexts classify more than 20 types whereas no frame does. This already suggests that these contexts are qualitatively different.

Applying the previous frequency threshold, frequent contexts in the French corpus were defined as the contexts occurring at least 18 times and classifying more than 11 types. For the English corpus, we kept the 45 most frequent contexts, just as in Mintz (2003).[4] Results are given in Table 3: in French, accuracy for frequent front and back contexts is .30 and .25; in English, it is .52 and .29. These scores are significantly above chance level ($p < .01$) although far below results from frames. Completeness is .050 and .057 for frequent front and back contexts in French; .056 and .046 in English. Except for front contexts in English, these results are at chance.

Discussion

Front contexts lead to slightly better results than back contexts. This asymmetry may reflect the fact that both French and English are *right* recursive languages so that function words generally precede the words they control. This explains why the co-occurrence of two adjacent words imposes a stronger syntactic constraint to the following words than to the preceding words.

Crucially, for English and French, frames yield much better categorization than the two continuous environments. The superiority of frames may be symptomatic of a syntactic consequence of the discontinuity that characterizes their structure. Specifically, we propose that the frequent co-occurrence of two words is indicative of a syntactic pattern that the two words together regularly exemplify, and thus the words are likely to be relatively close structurally in such situations. This would strongly constrain the structural relationship of the intervening target word and the framing elements. In contrast, in the case of front or back contexts, the preceding or following positions would be relatively unconstrained with respect to the target word. To illustrate this idea, Figure 3 shows a range of structural relationships between a target word and front or back contexts, as well as the two basic structural schemas for frames. The possible structures are much restricted for frames compared to front/back contexts. Intuitively, this should result in greater uniformity in the target position's syntactic category for frames compared to the other contexts.

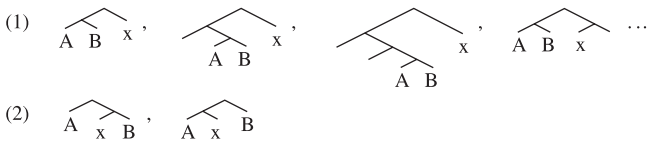

Figure 3 *A crucial role for discontinuity. Trees in (1) illustrate the fact that an adjacent pair of words in close syntactic relation could be embedded into virtually infinitely many structures (many more structures could be constructed if we lifted the constraint that A and B are immediate sisters, but we suppose that they are in close syntactic relationship since they co-occur frequently). Thus, the following – or preceding – syntactic position is not constrained very much: this may explain the low accuracy results of adjacent contexts (Experiment 2). Trees in (2), in contrast, show that when A and B are not adjacent (but are still syntactically close), only two positions remain theoretically available for an intervening word. This may account for the fact that discontinuity appears to be an essential feature of the success of the frequent-frames algorithm (even though adjacent contexts appear to be computationally equivalent, at first sight).*

[4] We considered other thresholds to make the new results match in number of tokens or types categorized, or number of groups obtained with groups obtained from frequent frames; results were similar.

Thus, the advantage of frames over the alternative contexts examined in this experiment could be explained by the types of syntactic structures that are likely to be involved. These results suggest that discontinuity is a crucial property of frequent frames for the purpose of categorization

In the next experiment we test the effects of recursively applying the frequent-frames analysis to a corpus. Specifically, we use the *groups* formed by an initial application of the procedure as the new framing elements, rather than specific words. This natural generalization may allow us to capture more abstract regularities; for example, *verbs are found between two pronouns*, as opposed to *verbs are found between 'he' and 'it'*.

Experiment 3: Item-specificity and recursivity

The frequent-frames mechanism as investigated so far could yield initial category knowledge that could serve as a basis for detecting new frame-like contexts with the re-application of the categorization procedure. For instance, if the words *I* and *you* have been categorized together, it may be reasonable to consider them as equivalent in terms of their role in defining frames and to obtain a single group from the frames [*I* x *it*] and [*you* x *it*]. This would be a highly desirable outcome, as it could consolidate separate frame-based groups belonging to the same linguistic category (for example, the frames [*I* x *it*] and [*you* x *it*] both contain verbs), thus making frame-based categories even more informative linguistically (but see Mintz, 2006, for an alternative consolidation proposal).

Material and method

The same French and English corpora as used in Experiment 2 were analysed here. In the first phase of the procedure, we performed a frequent-frames analysis as described in Experiment 1. In the second phase, we ran the procedure again, but allowed the groups produced in the first phase to participate in *recursive frames.* For example, suppose that the first application of frequent frames grouped together *I* and *you* in one frame (F1), and *this* and *that* in another (F2). Then, the utterance *I saw this* would trigger the categorization of the intervening word *saw* in three different recursive-frame-based groups: [F1 x *this*] (which groups together any word intervening between a word previously categorized in F1 and *this*), [*I* x F2] (*idem*) and [F1 x F2] (which groups together words intervening between a word categorized in F1 and a word categorized in F2).

Results

The results from the first phase are reported in Experiment 1 and Mintz (2003). Thus, we focus on the recursive application of the procedure.

Recursive frames

The groups derived from this procedure contain a large number of tokens and types. Furthermore, they capture words from a wider variety of categories, including function words. Establishing a frequency criterion for frames as before,[5] accuracy for frequent recursive frames is .37 in French and .25 in English: these scores are much lower than those obtained from simple frames (see Table 4), although still better than chance.

Asymptotic results

These lower results may be caused by some artifact in the recursive application of the frequent-frames mechanism. First, there were a few miscategorization errors within groups derived from item-based frames: these errors may add noise to the recursive step. Second, the actual groups provided by the first application of the mechanism may not be optimal for the purposes of recursive categorization – for instance because they capture mainly content words, which may impose fewer constraints on their neighbours than function words do. To evaluate the asymptotic performance of the recursive-frames mechanism, we assumed perfect categorization in the previous steps, and used actual syntactic categories to establish recursive frames. We call this analysis 'asymptotic' because it corresponds to analysing the limit of the performance of recursive frames, given an ideal prior categorization of framing elements.

The asymptotic manipulation did not improve the results of the recursive analysis. Even under these highly idealized circumstances, these environments provided groups with very low accuracy (see the last column of Table 4).[6] To give an example, the most frequent environment is [Verb x Noun], and its accuracy is .14. Sentences (5) to (8) exemplify, respectively, the occurrence of a determiner, an adjective, a preposition and a noun within this context.

(5) [Finish your cookie]. (Peter05.cha)
(6) Those [are nice towers]. (Peter12.cha)
(7) Why don't you [wait til lunchtime]. (Peter12.cha)
(8) Did you [say orange juice]? (Peter12.cha)

Discussion

These results show that the frequent-frames categorization procedure does not benefit from a recursive application, even under the idealized assumption that a first application provided a complete and error-free categorization. In other words, the information captured by frequent

[5] Again we ran the analysis with different thresholds and found no relevant difference.

[6] In an attempt to improve the categorization of the asymptotic application of the recursive algorithm, we also ran it with finer-grained categories for verbs, with three subclasses: auxiliaries, finite and infinitive verbs. This manipulation did not change the results.

frames that is relevant for lexical categories is fundamentally item-specific: frequent frames provide a better categorization when they involve specific words rather than their syntactic categories.

At first sight, this might seem like a counterintuitive result. After all, if grammars are organized around categories, shouldn't the category of the target word be predicted by the surrounding categories at least as well as by the surrounding words themselves? We speculate that allowing categories to define frames eliminates one of the powerful features of frequent frames, namely, that frequently co-occurring *items* in a frame configuration are a symptom of a linguistically 'stable' local environment. In other words, the frequent co-occurrence of two given items may reflect a given syntactic structure; if the framing elements are allowed to vary among whole groups of words, it may well be that the different instances of the co-occurrences involve different syntactic relations and thus impose different constraints on the intervening words. For instance, in sentences (5) to (8), the particular properties of the framing words – either the preceding verb, which could accept different constructions, or the noun, which could occur without a determiner or as a compound – may account for the variety of grammatical categories that can intervene.

General discussion

The analyses reported in this paper extend previous results described by Mintz (2003) for English: frequent frames help in the recovery of accurate syntactic categories in child-directed French as well. This is the case despite the fact that the function word system of French offers particular challenges to categorization based on frequent frames: for example, the increased number of determiners, and homophony between determiners and object clitics. In addition, we identified two core computational principles that should be particularly useful for any mechanism relying on context to categorize words into syntactic categories. First, discontinuous frames of the type [A x B] provide a much more efficient categorization than continuous contexts of the type [A B x] or [x A B], even though the quantity of information is formally the same in both context types. Discontinuous environments may be more constraining because of general syntactic properties of languages. Thus, line (2) in Figure 3 shows that when a specific syntactic relation holds between two non-adjacent words (as in frequent frames), the intervening syntactic position is highly constrained. It is very likely that the success of the frequent-frames algorithm derives from this type of local syntactic pattern. In contrast, line (1) in Figure 3 shows that pairs of adjacent words do not constrain their surrounding environments in the same way: a wide range of syntactic structures can fit a string of two adjacent words. Second, the recursive analysis presented in Experiment 3 shows that the distributional analysis is maximally efficient when the framing elements A and B are specific items rather than syntactic categories.

Both these principles fit well within a psychologically plausible acquisition model. For instance, infants at the start of the acquisition process already have access to specific items but not yet to established categories. It is then an unexpected bonus that item-specificity leads to better categorization than an analysis in which the framing elements are syntactic categories, even perfect ones. Furthermore, we suggested that a reason for this seemingly paradoxical finding is that recursive frames defined by open-class categories can select specific contexts that reflect a variety of structures, whereas frequently co-occurring *words* in a frame configuration reflect a much more stable structure. The words grouped by the recursive frames are hence more likely to be from different categories than are words grouped by lexically based frames.

How would the frequent-frames algorithm fit within a more global view of early lexical and syntactic acquisition? To start with, the computation of frequent frames relies on a prior segmentation of the speech stream into words. There is now converging evidence that word segmentation is efficiently mastered by the age of 10 to 16 months (depending on word types, see e.g. Jusczyk, Houston & Newsome, 1999; Nazzi, Dilley, Jusczyk, Shattuck-Hunagel & Jusczyk, 2005). Infants may thus start compiling frequent frames during the first half of their second year of life. Congruent with this hypothesis, Gomez & May (2005) showed that 15-month-olds were already able to detect non-adjacent dependences of the type [A x B] (that is, the necessary computational prerequisite to frequent frames), and Mintz (2006) showed that even 12-month-olds categorize together non-words that appear within the same frequent frames (see also Höhle *et al.*, 2004, for evidence that 15-month-old German infants exploit determiners to recognize valid noun contexts for novel words).

At this point, infants would possess frame-based categories, containing words that typically 'behave the same', in that they belong to the same syntactic category. However, even when accuracy is high, there are typically several frame-based categories for each syntactic category. For instance, several different frames pick out nouns, and several others pick out verbs. Learners would thus need to merge frame-based categories to obtain more comprehensive categories. Several possible strategies can be used to that end, such as grouping together frame-based categories that share one of their framing elements as well as some of their categorized words (see Mintz, 2003, 2006, for a fuller discussion of possible merging mechanisms).

Let us assume that learners successfully merged frame-based categories to obtain more comprehensive categories. Before they can use these categories to constrain their lexical and syntactic acquisition, they would need to *label* them. That is, they would need to identify which of these categories corresponded to nouns, verbs, adjectives, etc. One way to do this would be to identify

the syntactic category of a few words referring to concrete objects and events. For instance, if infants are able to acquire the meaning of a few frequent nouns referring to concrete objects (Gillette, Gleitman, Gleitman & Lederer, 1999), they may then be able to classify as nouns all the words that occur within the same distributionally defined category (even if these other nouns are not frequent themselves, or do not refer to concrete objects). On this view, then, distributional information in the form of frequent frames accomplishes the categorization work, and the first-learned words start the category-labelling process.

Such a process would be expected to occur some time before the age of 2, as recent experimental evidence suggests that infants are able to exploit the syntactic context in which a non-word occurs to infer something about its meaning. For example, 23-month-old French infants interpret a novel verb as referring to an action (Bernal, Lidz, Millotte & Christophe, 2007), and 24-month-old American infants interpret a novel preposition as referring to a relationship between objects (Fisher, Klingler & Song, 2006). We have shown that the frequent frames in speech to learners of either French or English provides distributional information that would allow them to converge on the relevant categories within this time frame.

Conclusion

This paper investigated the cross-linguistic validity of the frequent-frames mechanism for syntactic categorization in French and English. This constitutes the initial step in testing the cross-linguistic viability of this account of how children may initially categorize words. As part of this investigation, we discovered several characteristics that might make frequent frames a particularly robust context: discontinuity and item-specificity.

Future work should address its generalizability to other, typologically varied, languages. For instance, it remains to be shown that frequent frames would also be efficient in languages with more flexible word order such as Turkish. These languages are very rich in functional elements, but they appear as bound morphemes, not as words, as in English or French. Mintz (2003) has suggested that a generalized analysis that operated on morphemes rather than on words might capture the relevant regularities in languages with freer word order and richer morphology. Ultimately, a successful distributional theory of word categorization will have to consider the word (or morpheme) segmentation process that precedes it (e.g. see Christiansen & Onnis, this issue). That process essentially defines the units over which a categorization mechanism can initially operate. It could be that functional morphemes in Turkish, for example, are readily segmented by the same mechanism that segments words in English and French. A frequent-frames analysis could then operate on stems and affixes, rather than on open- and closed-class words. That level of analysis would probably result in much more stable patterns than would be available at the level of words in languages with rich morphology and more flexible word order.

On the other hand, some languages, such as Cantonese, are said to make limited use of function words. Given that frequent frames rely heavily on function words in the two languages studied so far, how would the frequent-frames analysis fare in languages like Cantonese? A preliminary analysis of Cantonese child-directed speech suggests that frequent frames still provide useful information, with an accuracy of around .80; importantly, discontinuity proved to be a crucial property, just as in French and English.

Further cross-linguistic research is necessary to address these questions, and to test further the validity of this account of early category learning. These studies will also shed light on whether the core computational principles advocated here provide the same benefits when analysing typologically different languages.

Acknowledgements

The work presented in this paper was supported by a grant from the French ministry of Research to Anne Christophe (ACI no. JC6059), by a PhD grant from the French Ministry of Research to Emmanuel Chemla and Savita Bernal, and by a grant from the National Institutes of Health (HD 040368) to Toben Mintz. The research was also supported by a grant from the Agence Nationale de la Recherche ('Acquisition précoce du lexique et de la syntaxe', no. ANR-05-BLAN-0065-01), and by the European Commission FP6 Neurocom project. We would also like to acknowledge the personal help and support we received from Emmanuel Dupoux, Christophe Pallier, Paul Smolensky and Dominique Sportiche.

References

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

Bernal, S., Dehaene-Lambertz, G., Millotte, S., & Christophe, A. (in press). Two-year-olds compute syntactic structure on-line. *Developmental Science*.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: if, when and why. *Cognitive Psychology*, **6**, 380–420.

Bloom, L., Lightbown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development*, **40**, (Serial No. 160).

Bloomfield, L. (1933). *Language*. New York: Holt.

Cartwright, T.A., & Brent, M.R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, **63** (2), 121–170.

Christiansen, M., & Onnis, L. (2009). The secret is in the sound: bootstrapping linguistic structure from phonemes. *Developmental Science, this issue*.

Fisher, C., Klingler, S.L., & Song, H.-J. (2006). What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition*, **101**, B19–B29.

Gillette, J., Gleitman, H., Gleitman, L.R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, **73** (2), 135–176.

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

Gleitman, L.R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J.C. (2005). Hard words, *Language, Learning and Development*, **1** (1), 23–64.

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science*, **13** (5), 431–436.

Gómez, R.L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, **7** (2), 183–206.

Harris, Z.S. (1951). *Structural Linguistics*. Chicago, IL: University of Chicago Press.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: the role of determiners in the syntactic categorization of lexical elements. *Infancy*, **5**, 341–353.

Jusczyk, P.W., Houston, D.M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, **39**, 159–207.

Koopman, H., Sportiche, D., & Stabler, E. (2003). *An introduction to syntactic analysis and theory*. Unpublished manuscript.

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

MacWhinney, B. (2000). The *CHILDES Project: Tools for anlyzing talk. Vol. 2: The Database* (3rd edn). Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M.P., & Chalkley, M.A. (1980). The internal language of children's syntax: the ontogenesis and representation of syntactic categories. In K.E. Nelson (Ed.), *Children's Language*, Volume 2 (pp. 127–214). New York: Gardner.

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

Mintz, T.H. (2006). Finding the verbs: distributional cues to categories available to young learners. In K. Hirsh-Pasek, & R.M. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 31–63). New York: Oxford University Press.

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

Nazzi, T., Dilley, L.C., Jusczyk, A.M., Shattuck-Hunagel, S., & Jusczyk, P.W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech*, **48**, 279–298.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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