1. Introduction

Knowledge of the grammatical category membership of words is an important and essential part in language development, since it is a prerequisite to know how to use words in the language and produce grammatically correct sentences. Without an accurate categorization of words, language acquisition cannot proceed. But how do children learn the basic grammatical categories of their language?

Several studies have already highlighted the usefulness of distributional information for the accomplishment of several tasks during language development (cf. Elman et al., 1996; Plunkett, 1995). For example, distributional information has shown to be a very powerful cue that assists young language learners when they face the task of segmenting the continuous speech stream into word-like units (Brent & Cartwright, 1996; Johnson & Jusczyk, 2001; Saffran, Aslin & Newport, 1996; Saffran, Newport & Aslin, 1996). In word segmentation tasks, distributional cues take the form of complex statistical information whereby sounds found within words will be strongly correlated (since they are part of the same word), while sounds found across word boundaries will be weakly correlated (since they occur one after the other only by chance). Saffran et al. (1996a) suggest that if children can learn words by recording frequent sound sequences, they might learn grammar in the same way.

Thus, when turning to the task of categorizing word-like units into their grammatical category, the same type of information might still be useful. In this case, the context of a word with respect to other words in the same sentence might provide indications about the category of that word in English. For example, English nouns are typically preceded by determiners and followed by nominal morphology, while verbs are typically preceded by auxiliaries or strong subject pronouns and followed by verbal

morphology. In this way, while in word segmentation the syllable *ba* is a high predictor of the syllable *by* (as in the word *baby*) and other syllables are not, in word categorization a determiner like *the* is equally a high predictor of a nominal element (as in the phrase *the baby*) and other words like *have* are not. In what follows, we shall examine the possibilities that distributional cues offer for word categorization.

2. Distributional cues and word categorization

As Mintz (2003) proposed, distributional information of the kind that can be found in the co-occurrence of patterns of words in sentences could provide a great deal of information relevant to the grammatical categories to which words belong. Studies on computer simulations have provided evidence for the usefulness of distributional and positional information for an initial categorization of words in the absence of semantic or referential information (Brent, 1996; Cartwright & Brent, 1997; Redington, Chater & Finch, 1998). Such distributional information appears to be available not only to adult speakers but also to young language learners (Gerken, Wilson & Lewis, 2005; Hsu, Tomblin & Christiansen, 2014; Mintz, 2002; Mintz et al. 2002; Mintz, Hao Wang, Li, 2014; Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007; Saffran et al., 1996b).

What type of distributional information is especially useful, and what kinds of distributional cues infants and young children are sensitive to and use in categorizing words? Authors distinguish between two different contexts: bigrams (Mintz et al., 2002; Redington, Chater & Finch, 1998) and frames (Mintz, 2003; Monaghan, Christiansen & Chater, 2007).

Bigrams are defined as pairs of elements where the categorizing element would either precede or follow the word to be categorized. In the case of English, for example, a word such as *come* could be successfully categorized as a verb on the basis of the bigram *they come* (where the strong subject pronoun would be the predictor of a coming verbal element) or on the basis of the bigram *coming* (where the verbal morpheme *-ing* that follows the target word would be the element that yields categorization).

Frames, on the other hand, are defined as "ordered pairs of words that frequently co-occur with exactly one word position intervening (occupied by any word)." (Mintz 2003: 93). For example, the sequence *they come here* could be considered a frame, where both the strong subject pronoun (i.e. *they*) as well as the following adverb (i.e. *here*) would be strong indicators that the intervening word in the middle is a verb.

Empirical evidence shows that bigrams and frames behave differently in terms of word categorization, as each distributional context gives very different results. Thus, while frames appear to be particularly good categorization contexts in terms of accuracy, those higher accuracy scores using frames are obtained at the expense of losing completeness strength (i.e. fewer elements are captured by frames than by bigrams). Therefore, frames provide high accuracy but weak completeness scores, while the opposite pattern is true of bigrams: they give higher completeness scores but less overall accuracy. In fact, one of the problems of distributional contexts that has been pointed out is that, when using frames for word categorization tasks, several grammatical categories might emerge, instead of a single adult-like one (Mintz, 2003).

A further limitation that a learner who relies on distributional information is likely to encounter is that of non-immediate adjacency. As noted by some authors (e.g.

Chomsky, 1975; Pinker, 1987), distributional regularities in English are not always local, but can occur over a variable distance, as shown in the examples below:

(1) a. The house has two floors.

b. The lovely big old white wooden house has two floors.

Patterns of lexical adjacency are variable in English. Thus, in the case of nouns and determiner-noun adjacencies, there can be a variable number of intervening modifying elements between the determiner and the noun. A learner who relies on strictly local distributional information and categorizes only from fixed positions could get to the wrong generalization that, for instance, *lovely* in (1b) above is a noun, and not an adjective.

How does the learner know which environments are important (as the one in (1a)), and which ones should be ignored? What kind of distributional cues are there in the linguistic input that might help children group individual lexical items into a larger and more general grammatical category? How reliable are such cues? To which extent could children work out the category for nouns in English accurately and on the basis of distributional information alone? The present paper aims at examining the kind and the amount of distributional cues which are present in the input addressed to English-learning children and which would allow the successful categorization of nouns in English. In what follows, the details of our analysis and the results obtained will be discussed.

3. Objectives

Several authors have already pointed out the usefulness of syntactic or distributional information for grammatical categorization (e.g. Cartwright & Brent, 1997; Mintz, 2003; Redington et al., 1998). Most of the studies that have been carried out to date are based on English data. Due to the limited number of English inflectional morphemes, distributional contexts or frames that have proved to be effective in the categorization of nouns in English are often made up of closed class items (e.g. pronouns, prepositions determiners) as well as open class words (e.g. lexical verbs or adjectives). As a consequence, a considerable number of possible categorizing frames emerge, and they are different across different corpora or different environments. This entails that, before facing the task of categorizing words, the child needs to work out which are the particular contexts that are relevant and useful for categorization in her own environment and which are not.

Furthermore, as mentioned earlier, considering a wide range of frames for word categorization brings about the risk of having language learners forming several different grammatical categories for elements which, in fact, have the same syntactic properties. When taken individually, frames exhibit low completeness scores, which results in more than one frame-based category for the single grammatical category of nouns.

In order to avoid that problem, in the present study, distributional contexts for noun categorization will be defined as the set of bigrams and frames which include determiners, interrogative elements or prepositions to the left of the intervening element (i.e. the element to be categorized) and English plural morphemes *–s* or *–es* to the right of the intervening element. Some researchers suggest that very young children might

not be able to attend to these elements, given the fact that they are not phonologically salient, neither in terms of stress nor in terms of word length. However, recent empirical evidence shows that children are well aware of the presence or absence of functional morphemes and words long before they start producing them (e.g. Gerken, 2001; Jusczyk, 2001). Therefore, they are very likely to be sensitive to the kind of words that make up the distributional contexts selected in this study.

The initial prediction is that a number of distributional contexts will emerge which will be smaller to those obtained in previous studies. Besides, there will also be considerable overlap across different corpora, in the sense that the same contexts will be useful for different children exposed to different kinds of input.

Furthermore, a type/token comparison analysis will be carried out in order to test the degree to which the same nominal type is captured by several different frames or bigrams in the form of different nominal tokens. If a considerable degree of lexical overlap between different distributional contexts is found, it might be possible that English-speaking children can generalize across different contexts and they might ultimately be able to come up with a single unifying and more abstract grammatical category.

As mentioned earlier, a further problem that has been identified with distributional analyses of the input is the fact that not all grammatical relationships are local, especially in the case of the English *Determiner + Noun* relationship, where there are often modifying elements like adjectives which are placed between determiners and nouns (Mintz, 2003; Pinker, 1987). As the present analysis only considers extremely local *Determiner + Noun* relationships, a further analyis will be carried out in order to test other open class words (i.e. verbs, adjectives and adverbs) against the same set of

frames and bigrams with which nouns were tested. Our prediction is that, while it is true that sometimes English nouns are not immediately adjacent to determiners, the categorization scores obtained by nouns will be higher than those obtained by all other elements together. Such regularities and consistencies in the environment would ultimately ban children from drawing the wrong conclusions about the grammatical categorization of certain non-nominal elements, despite their accidental appearance within nominal distributional contexts.

4. Methodology

4.1. Data source

The data used for the analysis comes from the Manchester corpus (Theakston et al., 2001), available from the CHILDES database (MacWhinney, 2012). This corpus consists of transcripts from a longitudinal study of twelve English-speaking children between the ages of approximately two and three years. In particular, from these twelve children, a subsample of four children was randomly selected for the analysis (i.e. the corpora corresponding to Aran, Carl, Anne and Becky).

Within the individual corpus of every child, only those transcriptions in which the child was 2;6 years old and younger were selected. All those transcriptions in which the children were 2;7 years old or older were discarded, unlike in other studies with similar goals, in which all the samples of child-directed speech were extracted, regardless the age of the child to whom these utterances were addressed (i.e. Monaghan et al., 2007; 2005). Previous research on the acquisition of nominal elements by children shows that, by the age of 2;6, children have already formed a grammatical category for nouns, as they have been shown to generalize and apply nominal

morphology productively to novel nouns they have not heard before (Tomasello & Brooks, 1999; Tomasello & Olguin, 1993). Since the aim of the study is to test the accuracy and reliability with which the nominal category is represented in children's input, we believed it necessary, therefore, to select from the corpora all the transcriptions in which the child is potentially young enough so as to fulfill the requirement of not having formed a noun category yet.

For the present study, only adult language was taken into account, and utterances spoken by the child were not analysed. This was due to the fact that the interest of the study lies on the kind of language that children are exposed to before they form a grammatical category for nouns, and not the kind of language that children themselves actually produce.

4.2. Corpus preparation

For the corpus of every child, lists of words with their corresponding token frequencies were obtained by using the FREQ utility of the CLAN program. The main goal of the present analysis is to work out the accuracy with which a given set of cues categorize English nouns and do not overcategorize other non-nominal elements. While previous studies focus on the accuracy with which a distinction between open vs. closed class items in general can be drawn (e.g. Monaghan et al., 2007; 2005), or on the accuracy with which nouns and verbs can be grouped into different grammatical categories (e.g. Monaghan et al., 2007; 2005; Mintz, 2003; Mintz et al., 2002), the focus of the present analysis is to test the likelihood with which nominal elements are successfully categorized by a given set of input-driven distributional cues, and to

measure the risk that the same cues might be too general as to wrongly categorize any other open class word within the noun category. Thus, from the list of words obtained, all lexical items were considered (i.e. nouns, verbs, adjectives and adverbs) and grammatical items such as determiners or modal auxiliaries were not considered in the list.

Words were then classified into two different categories to be analysed separately. One category included all nouns, and the other category, which was labelled as "other", included all verbs, adjectives and adverbs. For dual-class words, that is, English words that can, for instance, be both classified as nouns and verbs (e.g. *kiss*, *call*, *brush*), the KWAL utility of the CLAN program was used in order to work out the exact number of tokens that were nouns and the number of tokens that were used as verbs in every transcript.

4.3. Cue derivation

For the present analysis, a total of fifteen different distributional cues were considered, each corresponding to a different syntactic context. Distributional contexts were defined either as bigrams or frames. Bigrams included both combinations of a noun plus a plural morpheme (*–s* or *–es*), or the combination of an initial element (e.g an article or a possessive determiner) and a singular noun. The set of frames was made up of combinations of an initial element (e.g. an article or a possessive determiner) followed by a noun with plural morphology (*–s* or *–es*).

Following previous studies (Mintz, 2003; Mintz et al. 2002; Monaghan et al., 2005), the set of English articles, demonstrative determiners, possessive determiners,

quantifiers, prepositions and *wh*- interrogative elements were considered as categorizing items in the "*x + noun*" bigrams or the "*x + noun + -(e)s*" frames. Unlike Mintz (2003), this study only considered these closed-class words as the first element within a bigram or frame, and plural nominal morphology as the second categorizing element, and not just any independent word which would correlate with nouns. This was done in order to reduce the number of frames or possible distributional contexts and, therefore, the number of sources for noun categorization obtained by Mintz (2003). That would give a greater homogeneity to the resulting nominal category, or reduce the risk of having children create several different categories for the same elements. Besides, it would also eliminate variability in the environment of every child and make frames more similar across different corpora, which would give distributional contexts greater consistency. Furthermore, results would be more likely to be generalized to other infant English learners.

For the present analysis, only extremely local distributional contexts were considered, unlike other previous studies (cf. Mintz et al., 2002; Redington et al. 1998 St Clair, Monaghan, & Christiansen, 2010). Thus, while *boy* would be taken as a successfully categorized element in a bigram like *the boy*, it would not be so in a context like *the big boy*, where the adjective *big* would interfere between the categorizing element in the bigram *the* and the target word *boy*. Such an approach would allow the possibility of capturing and measuring the risk at which certain non-nominal elements might be miscategorized as nouns on the basis of their accidental occurrence within the set of established nominal distributional contexts (e.g. the case of the adjective *big* in the example above).

A list of the fifteen different distributional contexts was generated and every target word was analysed to see whether its context matched any of the fifteen established. Words scored 1 if they appeared in a syntactic context that matched the description of a given distributional cue and they scored 0 if their context did not match. Table 1 shows a summary of the set of distributional contexts that were considered.

Table 1. Distributional contexts obtained in the corpora.

Syn0	**a**	{∅} + x (e.g. *Peter, dog*)
	b	{∅} + x + *-(e)s* (e.g. *book**s**, house**s***)
Syn1	**a**	{*a, an*} + x (e.g. ***a** dog, **an** apple*)
Syn2	**a**	{*the*} + x (e.g. ***the** man, **the** children*)
	b	{*the*} + x + *-(e)s* (e.g. ***the** bus**es**,. **the** toy**s***)
Syn3	**a**	{*this, that, these, those*} + x (e.g. ***this** doll, **those** men*)
	b	{*this, that, these, those*} + x + *-(e)s* (e.g. ***these** girl**s**, **those** house**s***)
Syn4	**a**	{POSSESSIVE} + x (e.g. ***my** book, **their** food*)
	b	{POSSESSIVE} + x + *-(e)s* (e.g. ***her** book**s**, **your** kiss**es***)
Syn5	**a**	{QUANTIFIER} + x (e.g. ***much** milk, **many** people*)
	b	{QUANTIFIER} + x + *-(e)s* (e.g. ***many** stamp**s**, **some** sausage**s***)
Syn6	**a**	{PREPOSITION} + x (e.g. ***on** time, **for** lunch*)
	b	{PREPOSITION} + x + *-(e)s* (e.g. ***at** weekend**s**, **with** glass**es***)
Syn7	**a**	{WH- ELEMENT} + x (e.g. ***which** story, **what** colour*)
	b	{WH- ELEMENT} + x + *-(e)s* (e.g. ***whose** books, **which** trick**s***)

The same contexts were used for the analysis of nouns as well as for that of verbs, adjectives and adverbs of every corpus (i.e. if an adjective, for example, was

found immediately after a definite article, it would also score 1). Thus, *x* in the table stands for any intervening element within the bigram or frame.

In order to obtain the different distributional contexts for every word, the COOCCUR utility of the CLAN program was used to generate a list of every target word (i.e. every noun, verb, adjective and adverb) plus the word which occurred immediately before every target, as well as the overall token frequency of every obtained pair. Clusters of words containing nouns were analysed in order to determine the number of nouns within every corpus that was successfully categorized by any of the distributional cues. Other clusters of words containing verbs, adjectives or adverbs were analysed in order to determine the likelihood of finding in the input a word which is not a noun within a distributional syntactic context that would prototypically describe nouns. This was done in order to work out the degree of miscategorization and the possibility of finding non-nominal elements within nominal distributional contexts, since that would lead children erroneously to assume that such non-nominal elements are also nouns. As said before, such situations would include accidental adjacencies (e.g. *I like* ***those, give*** *them to me*) as well as non-immediate adjacencies within noun phrases (e.g. ***a*** *beautiful brown* ***dog***).

5. Results

5.1. Descriptive data

As mentioned earlier, four of the twelve children from the Manchester corpus were randomly selected for the present study. In particular, only transcriptions in which the children were 2;6 years old or younger were selected. From the list of words in each

corpus obtained with the FREQ utility of the CLAN program, a subsample of words was selected, from which all function words were discarded, and open class words were classified into two different groups: nouns, on the one hand, and all other open class words (i.e. verbs, adjectives and adverbs) on the other. Table 2 shows the total size of the corpus from all transcriptions that fell under the span of time under consideration before the target items were selected, the total number of open class words once function words had been eliminated, and the total amount of nouns and other open class words that resulted from the classification. These numbers include all the words from the corpora of all four children together.

Table 2. Total size of all corpora and total size of subsample selected for the present analysis.

	Total Types	**Total Tokens**	**Type/Token Ratio**	**Proportion of Types**	**Proportion of Tokens**
Total corpus	10,681	364,196	0.029	--	--
Total selected	9,621	139,624	0.069	0.90	0.38
Total nouns	5,388	51,577	0.104	0.56	0.37
Total other	4,233	88,047	0.048	0.44	0.63

Once all the open class words had been classified in either the “Noun” group or the “Other” group, words in both groups were tested against the set of distributional contexts described above (see table 1). The total number of nouns that was classified by each of the distributional contexts established is shown in table 3, while table 4 shows the results obtained from the equivalent analysis with the rest of open class words. As shown in the tables, the total number of noun types and tokens that were found in all the nominal distributional contexts under analysis (i.e. from *Syn0b* to *Syn7b*) is higher than

the corresponding type and token totals of all other open class words. On the other hand, the only distributional cue that described absence of categorizing syntactic context (i.e. *Syn0a*, which grouped all words which neither were preceded by a determiner nor were followed by plural nominal morphology) displays the opposite results, that is, there are less nouns and more other open class words that are found is this kind of syntactic context.

Table 3. Total of nouns found in distributional contexts.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	2,438	17,405	0.140	0.45	0.34
{∅} + x + -(e)s (Syn0b)	817	3,088	0.265	0.15	0.06
{a, an} + x (Syn1)	1,459	6,414	0.227	0.27	0.12
{the} + x (Syn2a)	1,707	9,314	0.183	0.32	0.18
{the} + x + -(e)s (Syn2b)	487	1,501	0.324	0.09	0.03
{DEMONSTRATIVE} + x (Syn3a)	728	2,292	0.318	0.14	0.04
{ DEMONSTR.} + x + -(e)s (Syn3b)	167	382	0.437	0.03	0.01
{POSSESSIVE} + x (Syn4a)	874	3,572	0.245	0.16	0.07
{POSSESSIVE} + x + -(e)s (Syn4b)	246	982	0.251	0.05	0.02
{QUANTIFIER} + x (Syn5a)	767	2,263	0.339	0.14	0.04
{ QUANTIFIER } + x + -(e)s (Syn5b)	454	1,210	0.375	0.08	0.02
{PREPOSITION} + x (Syn6a)	585	1,971	0.297	0.11	0.04
{ PREPOSITION } + x + -(e)s (Syn6b)	195	317	0.615	0.04	0.01
{WH- ELEMENT} + x (Syn7a)	222	769	0.289	0.04	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	54	101	0.535	0.01	0.00

Table 4. Total of other open class words found in distributional contexts.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	3,955	79,032	0.050	0.93	0.90
{∅} + x + -(e)s (Syn0b)	0	0	--	0.00	0.00
{a, an} + x (Syn1)	346	2,009	0.172	0.08	0.02
{the} + x (Syn2a)	228	1,112	0.205	0.05	0.01
{the} + x + -(e)s (Syn2b)	0	0	--	0.00	0.00
{DEMONSTRATIVE} + x (Syn3a)	433	1,615	0.268	0.10	0.02
{ DEMONSTR. } + x + -(e)s (Syn3b)	0	0	--	0.00	0.00
{POSSESSIVE} + x (Syn4a)	142	329	--	0.03	0.00
{POSSESSIVE} + x + -(e)s (Syn4b)	0	0	--	0.00	0.00
{QUANTIFIER} + x (Syn5a)	528	1,682	0.314	0.12	0.02
{ QUANTIFIER } + x + -(e)s (Syn5b)	0	0	--	0.00	0.00
{PREPOSITION} + x (Syn6a)	340	1,651	0.206	0.08	0.02
{ PREPOSITION } + x + -(e)s (Syn6b)	0	0	--	0.00	0.00
{WH- ELEMENT} + x (Syn7a)	85	617	0.138	0.02	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	0	0	--	0.00	0.00

5.2. Tests of significance

5.2.1. Types. In order to test the significance of distributional cues among types, a Mann-Whitney U-test was performed on the difference between the means of the

5.388 noun types and the 4.233 types from other open class words obtained from the corpora of all four chidren together. The results of the tests are found in table 5.

Table 5. Mann-Whitney U-test for the 15 distributional cues with all types.

Distributional cues	Nouns	Other	Z	Significance
{∅} + x (Syn0a)	0.45	0.93	-49.714	0.000
{∅} + x + -(e)s (Syn0b)	0.15	0.00	-26.483	0.000
{a, an} + x (Syn1)	0.27	0.08	-23.523	0.000
{the} + x (Syn2a)	0.32	0.05	-31.938	0.000
{the} + x + -(e)s (Syn2b)	0.09	0.00	-20.074	0.000
{this, that, these, those} + x (Syn3a)	0.14	0.10	-4.869	0.000
{these, those} + x + -(e)s (Syn3b)	0.03	0.00	-11.554	0.000
{POSSESSIVE} + x (Syn4a)	0.16	0.03	-20.383	0.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.05	0.00	-14.083	0.000
{QUANTIFIER} + x (Syn5a)	0.14	0.12	-2.513	0.012
{QUANTIFIER} + x + -(e)s (Syn5b)	0.08	0.00	-19.347	0.000
{PREPOSITION} + x (Syn6a)	0.11	0.08	-4.666	0.000
{PREPOSITION} + x + -(e)s (Syn6b)	0.04	0.00	-12.504	0.000
{WH- ELEMENT} + x (Syn7a)	0.04	0.02	-5.851	0.000
WH- ELEMENT} + x + -(e)s (Syn7b)	0.01	0.00	-6.531	0.000

As the table indicates, the fifteen distributional cues that were used in the analysis were highly significantly different in terms of their means for nouns and other open class words. As far as types are concerned, nouns are more likely than other open

class words to be preceded by articles (*Syn1* and *Syn2*), demonstrative determiners (*Syn3*), possessive determiners (*Syn4*), quantifiers (*Syn5*), prepositions (*Syn6*) as well as *wh-* interrogative elements (*Syn7*). Nouns are also more likely than other open class words to be followed by morpheme *–(e)s*. On the other hand, other open class words are more likely to be found in the kind of syntactic contexts in which no determiner precedes the target word and no morphological marker follows the word (*Syn0a*). In other words, absence of distributional contexts of the kind "*determiner + x*" is more typical of verbs, adjectives and adverbs than it is of nouns.

5.2.2. Tokens. The same significance test used for types was also performed on the 51,577 noun tokens and the 88,047 tokens from the "other" group from the corpora of the four children. The results of the Mann-Whitney U-tests are shown in table 6. As with types, the significance analyses with tokens revealed highly significant differences for the fifteen distributional cues as far as nouns and other open class words are concerned.

According to the results, nominal tokens are more likely to be found either before morpheme *–s*, or after a determiner, or between these two elements. On the contrary, other open class words are more likely to be found in a syntactic context with no such elements preceding or following the target word (i.e. context *Syn0a*).

Table 6. Mann-Whitney U-test for the 15 distributional cues with all tokens.

Distributional cues	Nouns	Other	Z	Significance
{∅} + x (Syn0a)	3.23	18.67	-48.073	.000
{∅} + x + -(e)s (Syn0b)	0.57	0.00	-26.452	.000
{a, an} + x (Syn1)	1.19	0.47	-23.528	.000
{the} + x (Syn2a)	1.73	0.26	-31.936	.000
{the} + x + -(e)s (Syn2b)	0.28	0.00	-20.066	.000
{this, that, these, those} + x (Syn3a)	0.43	0.38	-4.846	.000
{these, those} + x + -(e)s (Syn3b)	0.07	0.00	-11.554	.000
{POSSESSIVE} + x (Syn4a)	0.66	0.08	-20.467	.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.18	0.00	-14.081	.000
{QUANTIFIER} + x (Syn5a)	0.42	0.40	-2.575	.010
{QUANTIFIER} + x + -(e)s (Syn5b)	0.22	0.00	-19.341	.000
{PREPOSITION} + x (Syn6a)	0.37	0.39	-4.713	.000
{PREPOSITION} + x + -(e)s (Syn6b)	0.06	0.00	-12.503	.000
{WH- ELEMENT} + x (Syn7a)	0.14	0.15	-5.826	.000
WH- ELEMENT} + x + -(e)s (Syn7b)	0.02	0.00	-6.531	.000

5.3. Tests of diagnosticity

The significance differences revealed by the Mann-Whitney U-tests reported in the previous sections indicate that most of the distributional cues that were selected for the analysis contribute towards the classification of English nouns and distinguish them

from other open class words. However, how successful are those distributional cues in diagnostic tests for discriminating nouns from other open class words?

The combined contribution of all the distributional cues towards correct classification of words was further tested using a multivariate linear discriminant analysis. Discriminant analysis provides a classification of items into categories based on a set of predictor variables. The chosen classification maximizes the correct classification of all members of the predicted groups. The baseline for classifying into two categories is 50%. Therefore, correct classification over 50% means that there is useful information in the predictor variables (Meyers et al., 2006). The results from such analyses are shown below.

5.3.1. Types. Correct classification of all types from the four corpora was assessed for the fifteen distributional cues together. When all cues were entered simultaneously, 72.8% of nouns and 88.8% of other open class words were correctly classified. Overall, 79.9% of types were correctly classified. This was highly significant (Wilks $\lambda = 0.611$, $\chi^2 = 4734.621$, $p < 0.001$). Thus, as far as types are concerned, the results show that categorization using distributional cues was very successful, both in terms of accuracy (i.e. the amount of other open class words correctly categorized as "other") as well as completeness (i.e. the amount of nouns that were correctly categorized as nouns).

A further validation of these results was carried out using 50% of the sample as a holdout group to perform cross-validation, given that the size of the sample was large enough so as to allow for such validation to be done without subsamples affecting significance results. With this method, 50% of the total number of types are selected at

random. Then, the standard discriminant analysis is performed using exclusively that first subsample which has been randomly selected. Once this is done, the second half of the sample which was left as a holdout group is then tested against the system for classification obtained from the discriminant function with the first half of the sample.

It is believed that this cross-validation is the closest estimation that can be performed to the task that language learners undertake. That is, a learner with no prior knowledge about neither the makeup of each grammatical category nor the category membership of each of the words encountered, would start the grammatical categorization task by working out the set of features that define each of the grammatical categories on the basis of a set of linguistic subsamples obtained from the environment. Learners would then use that system they have come up with to further classify new elements as they find them in their linguistic input. Thus, if the results from this cross-validation are still successful, this entails that the learner's task at classifying new elements is not at risk of missclassifying words into wrong categories.

The results obtained from this cross-validation using 50% of the sample gave a total of 71.8% of nouns and 88.7% of other open class words correctly classified. This was highly significant: Wilks $\lambda = 0.602$, $\chi^2 = 2419.459$, $p < 0{,}001$. Thus, although there is a slight drop in accuracy both for correct classification of nouns as well as other open class words, the selected variables are still valid for categorization purposes, since most items are correctly classified.

5.3.2. Tokens. Correct classification of all tokens from the four child corpora was assessed for the fifteen distributional cues together in the same way as with types. As far as tokens are concerned, when all cues were entered simultaneously, 44.6% of

nouns and 96.5% of other open class words were correctly classified. Overall, 67.4% of tokens were correctly classified. This was highly significant (Wilks $\lambda = 0.921$, $\chi^2 = 790.324$, $p < 0.001$).

As can be seen from the results obtained, compared to types, the analysis of tokens reveals a higher accuracy score, since the elements from the "other" group are more accurately classified into their corresponding category, and there are less other open class words that are misclassified as nouns. However, completeness scores in the token analysis are much lower than those of the analysis with types, since there were many more nominal elements that fell out of their corresponding category and were misclassified as other open class words (i.e. only 44.6% of nouns were mapped into their corresponding grammatical category).

A cross-validation analysis using 50% of the cases chosen at random was also performed, as with types. Again, the results obtained from the second cross-validation were very similar to those obtained with the standard discriminant analysis: correct classification reached a total of 45.7% with nouns, and 97.1% with other open class words, resulting in a total of 68.4% of items which were correctly classified into their corresponding category. This was also highly significant, Wilks $\lambda = 0.919$, $\chi^2 = 401.987$, $p < 0.001$.

Following previous studies (Monaghan et al., 2005; 2007), in order to explore whether word frequency had an impact on accuracy in word classification, a subset of high-frequency tokens were selected and tested against the fifteen distributional variables. This selection included words with a token frequency of 10 or higher, and it resulted in a total of 1,168 nouns and 1,105 other open class words. When the same discriminant analysis was performed with these high-frequency words, a total of 70.2%

of nouns and 94.8% of other open class words were correctly classified, Wilks λ = 0,713, χ^2 = 727,759, p < 0,001. Thus, while accuracy scores were kept, completeness scores improved significantly. Overall, within high-frequency words, a total of 82.8% of tokens were correctly classified.

A further cross-validation with 50% of randomly selected cases was also performed with this subset of high-frequency tokens. The results found from this cross-validation were again very similar to those found in the original analysis (overall correct classification of 82.8% of high-frequency words, Wilks λ = 0,702, χ^2 = 387,876, p < 0,001), indicating that the set of predictor variables is robust and valid as a classification method.

6. Discussion

The main objective of the present study was to analyze whether the distributional information available in child-directed speech was sufficient for English-learning infants to group nouns in their corresponding grammatical category. As seen earlier, the type of distributional contexts that were considered in the present study were either bigrams (i.e. one categorizing element plus one intervening element to be categorized, either to the left or to the right), or frames (i.e. two categorizing elements plus one intervening element to be categorized in the middle). Furthermore, nominal distributional contexts for the present analysis were designed as including only the most prototypical elements that make up English noun phrases, namely determiners and nominal morphology, as elements that act as context words in bigrams or frames. This differs from previous studies with similar goals and methodology, which also included

other categorizing elements such as interjections (Monaghan et al., 2005), or just any kind of context word (Redington et al, 1998; Mintz, 2003).

As a result, the number of distributional contexts that emerged in the present study was smaller than in all previous studies, as only fourteen distributional contexts emerged here (i.e. six frames and eight bigrams). There was an additional distributional variable which was used as control and described absence of any context word, either to the left or to the right of the intervening word (i.e. *Syn0a*). This yielded a total of fifteen distributional variables. A further advantage to considering only determiners and nominal morphology as context elements was that there were minimal qualitative differences in terms of the make up of distributional contexts across the four different corpora under analysis.

Significance tests showed that differences between nouns and other open class words obtained from the corpora under consideration were statistically highly significant in terms of distributional variables. Furthermore, for most of the variables that subsumed distributional contexts where nouns were expected, mean scores among nouns were higher than mean scores among other open class words.

Regarding sucessful classification, the initial prediction was that fewer distributional contexts would increase the explanatory force of each individual frame or bigram, but at the expense of getting lower completeness scores in diagnosticity tests than in previous studies. This prediction was born out: previous analyses which are directly comparable, as they use similar methodology and similar measures, obtained completeness scores ranging from 53% of correctly classified nominal tokens (Redington et al., 1998) to 62.4% of correctly classified tokens (Monaghan et al., 2005). In the present analysis, when all tokens were considered simultaneously with all fifteen

distributional variables, a total of 44.6% of nominal tokens were correctly classified. Nevertheless, when correct classification of nominal elements was assessed in interaction with word frequency, completeness scores improved significantly, with a correct classification of 70.2% of high-frequency nouns.

Furthermore, the analysis with all fifteen distributional cues using all the types from the corpora gave very high completeness scores as well, with 72.8% of correctly classified nominal types. Evidence from previous research studies in language processing as well as first language acquisition (Bybee, 1995; Maratsos, 2000; Marchman & Bates, 1994) suggests that regular morphosyntactic patterns from words are generalized once patterns exhibit a relative type frequency. As far as language development is concerned, the studies suggest that, in terms of extracting morphosyntactic patterns, children appear to be more attentive to type frequency than to token frequency. Thus, they are more likely to generalize from distributional contexts that appear on many stems than those that appear on only a few stems, even when the token instantiations of those fewer stems have an overall higher frequency. High token frequency is useful to keep an irregular form, but does not make a paradigm productive. On the other hand, type frequency helps language learners to identify productive paradigms (Clark, 2009). Given the considerable lexical overlap between types and tokens, children might be able to extract a rote-learned distributional pattern from a few instantiations and apply it productively to other units in a rule-based way. In similar lines, Reeder et al. (2013) have shown that language learners can generalize the distributional properties of a grammatical category to a new word that shares just one context with the other category members. However, when a new word does not show any overlap in terms of distributional contexts, learners become more reticent to

generalize regarding the distributional properties of that novel word. According to Reeder et al. (2013), these results show that learners use distributional information in a systematic way to determine when to generalize and when to keep the lexical irregularities of specific items.

Criticisms to distributional analysis approaches (e.g. Chomsky, 1975; Pinker, 1987; 1997) have pointed out the risks of postulating a model based on mere linear contiguity between two given elements. They emphasize the fact that human language has a hierarchical and not linear structure, and that learning the syntax of a language is a matter of acquiring a complex structure, and not simply learning a mere chain of elements, one immediately adjacent to the other. In this line, they suggest that a pure distributional analysis of linguistic input in which only exclusively local dependencies are considered (e.g. Mintz, 2003; Monaghan et al., 2005 or the present study) might lead to wrong inferences on the part of the learner, since not all relationships between linguistic elements are exclusively local.

Thus, for example, a noun categorization model based on the assumption that a distributional context such as "{*the* + x}" (i.e. variable *Syn2a* in the present analysis) will subsume a great proportion of English nouns does not account for the fact that the position of "x" in this distributional context can be occupied by other elements which are not nouns in English (e.g. *the white door*, *the brown dog*). In those cases, the English-learning child is at risk of taking the wrong assumption that *white* or *brown* are nouns, since they are found in contexts where nouns should be found according to a strict distributional analysis of the input based on local dependencies.

In order to analyse the degree to which the selected distributional contexts in the present study would be overpermissive so as to miscategorize other non-nominal

elements as nouns, accuracy measures were taken, besides the completeness measures mentioned above. Categorization accuracy measures were obtained out of testing the list of other open class words (i.e. verbs, adjectives and adverbs) against the same set of distributional variables that were used with nouns. Since the kind of distributional variables that were considered in the present study include context words which are typically associated with the syntactic contexts of noun phrases, the prediction regarding miscategorization was that accuracy scores would be high. In other words, the percentage of other open class words which would be wrongly misclassified as nouns on the basis of the selected distributional variables would be low.

This prediction was also born out in light of the results obtained from the present study, since all the scores obtained from the other open class word group (i.e. accuracy scores) were very high. Thus, in the analysis taking all tokens simultaneously when the fifteen distributional variables were considered, a total of 96.5% of other open class word tokens were correctly classified (i.e. statistical discriminant analysis classified them in the "other" group and not in the "noun" group on the basis of their distributional behaviour) . This suggests that, while it is true that sometimes *Determiner + Noun* syntactic relationships are not immediately adjacent in English, they are by far the most recurrent syntactic pattern in terms of statistical discriminant functions, since only a remaining 3.5% of elements other than nouns would ever be immediately preceded by a determiner and would be at risk of being wrongly misclassified as nouns.

Accuracy scores obtained from the analyses with types or with high-frequency words were not as high as the scores obtained from the analysis with all tokens, but they were all very high proportions of correctly classified other open class words anyway. Among types, accuracy scores reached a total of 88.8%, while accuracy scores among

high-frequency words was 94.8%. Therefore, on the basis of the evidence provided by these data, the kind of distributional contexts established in the present study for the categorization of English nouns can be claimed to be accurate enough so as not to overcategorize and subsume elements other than nouns, which would ultimately avoid the miscategorization of certain elements on the part of young language learners.

Thus, for the most part, the fifteen distributional cues considered in this study have proved to successfully contribute to the correct categorization of nouns in English. Therefore, the type of linguistic input to which children are exposed during their first years of life can be claimed to contain distributional information which is consistent and reliable enough for the onset of word categorization in the process of language development.

Funding

This work was supported by the Spanish Ministry of Education [grant number FFI2013-47616-P]; and the Autonomous Catalan Government [grant number 2014SGR1089].

References

BRENT, M.R. (1996). Advances in the computational study of language acquisition. *Cognition*, 61, 1-38.

BRENT, M.R. & CARTWRIGHT, T.A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93-125.

BYBEE, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*. 10, 425-455.

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

CHOMSKY, N. (1975). *The Logical Structure of Liguistic Theory*. Plenum Press, New York.

CLARK, E.V. (2009). *First Language Acquisition*. Cambridge: Cambridge University Press.

ELMAN, J.L., BATES, E.A., JOHNSON, M.H., KARMILOFF-SMITH, A., PARISI, D. & PLUNKETT, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

GERKEN, L.A. (2001). Signal to syntax: Building a bridge. In Weissenborn, J. & B. Höhle (eds.). *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition* (pp. 147-165). Amsterdam: John Benjamins.

GERKEN, L.A., WILSON, R. & LEWIS, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32, 249-268.

HSU, H.J., TOMBLIN, J.B. & CHRISTIANSEN, M.H. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Frontiers in Psychology*, 5, 175.

JOHNSON, E.K. & JUSCZYK, P.W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 548-567.

JUSCZYK, P.W. (2001). Bootstrapping from the signal: some further directions. In Weissenborn, J. & B. Höhle (eds.). *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition* (pp. 3-23). Amsterdam: John Benjamins.

MACWHINNEY, B. (2012). *The CHILDES project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

MARATSOS, M. (2000). More overregularizations after all: New data and discussion on Marcus, Pinker, Ullman, Hollander, Rosen, & Xu. *Journal of Child Language*, 27, 183-212.

MARCHMAN, V. & E. BATES. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21, 339-366.

MEYERS, L.S., G. GAMST & A.J. GUARINO. (2006). *Applied Multivariate Research*. London: SAGE Publications.

MINTZ, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, 30, 678-686.

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

MINTZ, T.H., HAO WANG, F. & LI, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (Bigram) parts. *Cognitive psychology*, 75C, 1-27.

MINTZ, T.H., NEWPORT, E.L. & BEVER, T.G. (2002). The Distributional Structure of Grammatical Categories in Speech to Young Children. *Cognitive Science*, 26, 393-424.

MONAGHAN, P., CHATER, N. & CHRISTIANSEN, M.H. (2005). The differential contribution of phonological and distributional cues in grammatical categorisation. *Cognition*, 96, 143-182.

MONAGHAN, P., CHRISTIANSEN, M.H. & CHATER, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, 55, 259-305.

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

PINKER, S. (1997). Acquiring Language. *Science*, 276, 1177-1181.

PLUNKETT, K. (1995). Connectionist approaches to language acquisition. In P. Fletcher & B. MacWhinney (eds.). *The Handbook of Child Language* (pp. 36-72). Oxford: Basil Blackwell.

REDINGTON, M., CHATER, N. & FINCH, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 435-469.

REEDER, P.A., NEWPORT, E.L. & ASLIN, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66(1), 30-54.

SAFFRAN, J.R., ASLIN, R.N. & NEWPORT, E.L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

SAFFRAN, J.R., NEWPORT, E.L. & ASLIN, R.N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35, 606-621.

St CLAIR, M.C., MONAGHAN, P., & CHRISTIANSEN, M.H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116, 341-360.

THEAKSTON, A.L., E.V. LIEVEN, J. PINE, C.F. ROWLAND. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language*, 28, 127-152.

TOMASELLO, M. & P.J. BROOKS. (1999). Early syntactic development: A construction grammar account. In M. Barrett (ed.). *The Development of Language* (pp. 161-190). Hove Psychology Press.

TOMASELLO, M. & R. OLGUIN. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8, 451-464.