UNIVERSITAT DE
BARCELONA

Learning from the input: syntactic, semantic and phonological cues to the noun category in English

Sara Feijóo Antolín

LEARNING FROM THE INPUT:

SYNTACTIC, SEMANTIC AND PHONOLOGICAL CUES TO THE NOUN CATEGORY IN ENGLISH

Tesi doctoral presentada per

Sara Feijóo Antolín

com a requeriment per a l'obtenció del títol de

Doctora en Filologia Anglesa

Programa de Doctorat: *Lingüística Aplicada*
Bienni 2005-2007

Departament de Filologia Anglesa i Alemanya
Universitat de Barcelona
Setembre 2010

Directores:
Dra. Carme Muñoz Lahoz
Dra. Elisabet Serrat Sellabona

ACKNOWLEDGEMENTS

During the three years I have spent writing this Ph.D. dissertation I have been lucky enough to count on the help of many great people. They deserve my most sincere gratitude and I would like to dedicate these very first lines of my dissertation to thank all of them.

First of all, I would like to express my deepest thanks and appreciation to my supervisors, Dr. Carme Muñoz, from the English and German Philology Department at the University of Barcelona, and Dr. Elisabet Serrat, from the Psychology Department at the University of Girona. Dr. Muñoz's expertise in the field of language acquisition has been fundamental for the successful completion of this thesis. I very much appreciate her insightful comments and feedback, as well as her support. She has been extremely fast, efficient and very encouraging, despite all her other academic and personal commitments.

Dr. Serrat's excellent command of research tools for the study of child language has been a key element to carry out this study. This thesis would not have been possible without her help with the CHILDES database as well as the statistical tools, not to mention her magnificent support, her guidance and her patience at all stages of this thesis. I am grateful to Dr. Serrat not only for her professional involvement in the fulfillment of this academic requirement, but also for her personal implication. From the very beginning, she has been incredibly supportive, both personally and emotionally, and it is certainly impossible to account for everything that she has offered me in just a paragraph. I also thank Ada and Biel, for answering the phone so readily, and Miguel, for the printing.

My thanks are extended to the members of both research groups at both universities. Thank you to all members of the *Llenguatge i Cognició* group at the University of Girona, especially to Anna Amadó, Jesica Serrano and Dr. Francesc Sidera, for their cheerfulness and their positive comments, especially during the last stages of this thesis, and for making me smile at times when I thought there was nothing to smile about.

Many thanks to all the members of the *GRAL* research group at the University of Barcelona as well. I am especially indebted to Dr. Elsa Tragant, for encouraging me to start doing research back in 2002, and for her help, her time and her patience, especially during the writing of the last chapters of this dissertation. I am also grateful to Dr. Teresa Navés, for her support and her willingness to collaborate, and to Dr. Imma Miralpeix, for her useful contributions regarding semantics and vocabulary acquisition, and for her encouraging remarks.

Having to combine the writing of this thesis with my teaching duties was not always easy. Thus, I would also like to express my gratitude to all the professors from the English and German Philology Department at the University of Barcelona with whom I have been able to collaborate. My most sincere acknowledgements to Dr. Isabel Verdaguer, Núria Gavaldà, Emilia Castaño, Dr. M. Àngels Llanes, Dr. Raquel Serrano and Dr. Roger Gilabert, for making things far easier. I also wish to thank the administrative staff of the Department, Olga Borderas and Mercedes Padial, for their support and their efficiency.

My job at the English and German Philology Department has also given me a number of rewarding experiences. One of the most important ones has been the opportunity to meet my dear friends Elisabet Comelles and Dr. Natàlia Judith Laso. I am grateful to both of them for their kindness and their friendship. Eli's invaluable help and encouragement has granted me the strength I needed to go on with my work and my research. I very much appreciate her support at all times, even when things were not any easy for her either. And what to say about Natàlia? I simply cannot find the words to express my gratitude for her advice, her support, her comforting remarks, and her incredibly useful comments on the last draft of my thesis, which she produced so carefully and in her most beautiful handwriting.

I would also like to thank the organizers of the 2007 Summer School on First Language Acquisition in Basel, Switzerland. Thank you to Dr. Annelies Häcki Buhofer and Dr. Heike Behrens, for allowing me to participate in the workshop. Attending those sessions gave me the inspiration I used as the seed for the present dissertation. I am especially thankful to all the professors that participated in the seminar, Dr. Melissa Bowerman, Dr. Shanley Allen, Dr. Eve V. Clark, Dr. Elena Lieven and Dr. Michael Tomasello, for their impressive teaching at the workshop, their useful comments and their ideas.

From outside the field of Linguistics and Psychology, many thanks must go as well to Dr. Antònia Agulló and Dr. Jaume Vernet. It has been more than a pleasure to be able to meet them and I am certainly grateful for the great moments we shared together, their encouragement with my thesis and their wholehearted support. Their wise comments and friendly remarks have comforted me during many sleepless nights.

I also have to acknowledge all my friends who have helped me and have cheered me up during all these years, especially during the last hectic year of my thesis work. Thank you for making the experience a bit more pleasant for me. A big hug to all of you. My best friend Inma Palomar deserves not only a special line here but also a monument. She has always been there for me through thick and thin and I owe almost everything to her. Thank you for being such a good listener twenty-four hours a day and seven days a week, and thank you for being the greatest person and the best friend one could ever have.

I would also like to say thank you to all the members of my family, both in Catalonia and outside. Many thanks especially to my parents, for their patience, their help and all their sacrifices. Thank you very much to my dear sister Núria, whom I admire deeply. Her talent and her gorgeous pictures have cheered me up and accompanied me on many occasions. Many thanks also to Jaume Sarret, for his time, his support, for the great moments and his wonderful cooking. And of course, many thanks to seventeen-month-old little Júlia, my dear niece, for sharing her *lacasitos*, her dancing, her *ninis*, and for her sweet little words, which gave me another reason to believe in the usefulness of morphosyntactic frames for word categorization and language development.

Last and foremost, I would like to express my most sincere gratitude to Dr. Joseph Hilferty, with whom I have learned so much. The experience I got from him and my achievements under his guidance will certainly stand out as the most important milestones in the whole of my scientific career. He was the professor of one of the first courses in Linguistics that I took during my B.A., back in 1998, and the one who influenced me the most. His wonderful teaching was what made me become interested in the study of human language. After that, during my Ph.D. studies, it was thanks to him and under his guidance that I discovered the impressive empirical work on usage-based linguistics and language acquisition by Michael Tomasello, as well as the statistical learning experiments by Jenny Saffran and her colleagues. I would never have even started writing this thesis if it had not been for him. I am truly grateful for all his work, his inspiration and his helpful comments on my work at all times during these years. And I am certainly grateful for his invaluable friendship. This whole dissertation is dedicated to him. I only hope one day he feels as proud of me as I am of him.

Sant Feliu de Llobregat, 20th September 2010.

To Joe

ABSTRACT

Over the past years, mainstream linguistic theories have described first language acquisition and development as a process involving innate knowledge about the grammatical properties of language on the part of infant language learners. Such accounts of syntactic development assume the primary linguistic input to be too poor for language learning to take place (i.e. the so-called *Poverty from the Stimulus Argument*). Innate linguistic knowledge would thus provide children with the information they need to become fully competent speakers of their native language. The present dissertation aims at challenging such a view by showing that children get to syntax from the information available in the speech signal.

One of the most important linguistic tasks that children must accomplish throughout their development is that of categorizing words into their corresponding grammatical categories. Knowledge of the grammatical category membership of words is an essential part of language development, since it is a prerequisite to know how to use words and produce grammatically correct sentences. How do English-learning children know that "*table*" is a noun, "*eat*" is a verb, and "*kiss*" can be both a noun and a verb? Under nativist accounts, abstract grammatical categories are innate, that is, children have the *a priori* knowledge of the fact that the language to which they are exposed contains such things as nouns, verbs or adjectives. Their task is simply to learn the lexical items of their native language's lexicon and to map these items into their categories. However, recent empirical research shows that infants possess a series of learning mechanisms which are more powerful than previously assumed and which enable them to extract regularities from their linguistic environment and use them to acquire basic grammatical information. Consequently, if evidence is found that grammatical categories are reliably and consistently represented in the speech addressed to young language learners, given the fact that learners are able to track these regularities, there will be no need of postulating neither a presumably impoverished linguistic input nor any kind of innate knowledge about the nature of grammatical categories.

The present study examines the reliability with which the noun category is represented in the speech addressed to English-learning children up until the age of two and a half years old. A close examination is undertaken at a series of phonological, distributional and semantic criteria which most English nominal elements have in common and which would potentially make up the English noun category. The results from this study reveal that each of the above-mentioned criteria alone can account for a considerable proportion of the whole inventory of nouns to which children are exposed. More interestingly, when combined, all those three sources of information subsume most of the nouns under consideration.

Therefore, on the one hand, this study provides evidence for the fact that phonology, syntax and semantics are not clear-cut dimensions of linguistic description, but they are rather interconnected and interact with one another in a systematic way. On the other hand, the study presents compelling evidence for the fact that the usefulness of linguistic experience has been underestimated under nativist accounts of language development. Thus, the results show that the environmental speech stream to which very young language learners are exposed contains information which is sufficient and reliable enough so as to form an abstract grammatical category for nouns on the basis of experience alone and without any *a priori* linguistic knowledge.

RESUM

Durant els darrers anys, les línies centrals de la teoria lingüística han descrit l'adquisició i el desenvolupament de primeres llengües com un procés basat en el coneixement innat de les propietats gramaticals de la llengua per part dels infants aprenents. Aquestes descripcions del desenvolupament sintàctic assumeixen que l'entorn lingüístic a què els nens estan exposats és massa limitat com per donar lloc a l'aprenentatge de la llengua (és a dir, l'anomenat *argument de la pobresa de l'estímul*). D'aquesta manera, és el coneixement lingüístic innat el que permetria als infants esdevenir parlants competents de la seva llengua nadiva. Aquest estudi es planteja l'objectiu de qüestionar aquesta visió tot mostrant que els nens poden assolir coneixements gramaticals a partir de la informació que reben de l'entorn.

Una de les tasques més importants que els nens han de resoldre al llarg del seu desenvolupament és la d'agrupar paraules a les seves categories gramaticals corresponents. Saber a quina categoria gramatical pertany cada paraula és un component essencial del desenvolupament lingüístic, ja que aquest fet constitueix un requisit bàsic per tal de saber com utilitzar les paraules i arribar a formar frases gramaticalment correctes. Com saben els infants aprenents d'anglès que "*table*" és un nom, "*eat*" és un verb, i "*kiss*" pot funcionar com a nom i com a verb? Segons una visió innatista, les categories gramaticals són innates, és a dir, els nens ja saben *a priori* que la llengua a què estan exposats conté coses com ara noms, verbs o adjectius. La seva tasca simplement consisteix a aprendre les peces lèxiques del vocabulari de la seva llengua materna i relacionar-les amb la seva categoria corresponent. Això no obstant, estudis empírics recents demostren que els infants posseeixen una sèrie de mecanismes d'aprenentatge molt més eficients del que s'havia assumit prèviament. Aquests mecanismes els proporcionen la capacitat d'extraure les regularitats del seu entorn lingüístic i fer servir aquesta informació per aprendre les propietats gramaticals bàsiques d'una llengua. Així doncs, si les categories gramaticals estiguessin representades de manera consistent i fiable en la parla adreçada als infants, donat que els infants són sensibles a aquest tipus d'informació, no hi hauria cap necessitat de postular teòricament un entorn lingüístic empobrit o un coneixement innat pel que fa a les categories gramaticals.

Aquest estudi examina la fiabilitat amb què la categoria nominal està representada en la parla adreçada a infants aprenents d'anglès de fins a dos anys i mig d'edat. Es realitza un estudi detallat d'una sèrie de trets fonològics, distribucionals i semàntics comuns a la majoria d'elements nominals en anglès i que constituirien els criteris a partir dels quals es forma aquesta categoria. Els resultats d'aquest estudi demostren que cadascun dels esmentats criteris descriu acuradament bona part de l'inventari de noms a què els nens estan exposats. A més a més, quan tots tres criteris es combinen entre ells, la correcta classificació dels noms analitzats augmenta.

Així doncs, per una banda, aquest estudi evidencia el fet que la fonologia, la sintaxi i la semàntica no són dimensions de descripció lingüística nítidament separades l'una de l'altra, sinó que totes tres interactuen i estan relacionades entre sí de manera sistemàtica. D'altra banda, aquest estudi presenta evidència convincent que apunta cap al fet que, dins dels criteris innatistes del desenvolupament del llenguatge, l'utilitat de l'experiència lingüística dels nadons s'ha subestimat. Així doncs, els resultats demostren que la llengua de l'entorn a què els infants aprenents estan exposats conté informació prou rellevant i fiable com per formar una categoria gramatical nominal abstracta a partir de l'experiència i sense previ coneixement lingüístic innat.

Table of contents:

Appendices

List of figures:

List of tables:

Chapter 1
Introduction

Current theories in evolutionary biology describe structures that aid a given organism as either inherited or specific adaptations. Inherited adaptations are said to derive from the organism's ancestors. Such traits are general in two important ways: first, many species share them (i.e. they are not species-specific) and, second, they serve relatively general purposes (i.e. they are not domain-specific). Specific adaptations, on the other hand, may have suited the organism's specific needs (Kelly & Martin 1994: 105; see also Deacon, 1998; Tomasello, 1999).

As for human beings, some of the cognitive, perceptual and motor skills that we possess, or the biological structures that underlie them, have certainly been inherited from our ancestors. Thus, such behavioral competences are shared across different species (e.g. *bipedalism*, the ability to walk on our legs). By contrast, other faculties might be unique to our own evolutionary history. Over the years, mainstream linguists have claimed that language is one such faculty. Our ability for speech and grammar has been described as both species-specific, in the sense that it is unique to human beings, and domain-specific, since linguistic competence is seen as independent from other systems of human perception and cognition (Chomsky, 1975).

Under such accounts, learning is accomplished by mechanisms which are particular to the domain in question. Thus, for linguistic nativists, language learning is a task which involves mechanisms that are specific to language and that only human beings possess.

> The study of language and cognition during the past several decades has given increasing credibility to the view that human knowledge of natural language results from –and is made possible by– a biologically determined capacity specific both to this domain and to our species. (Anderson & Lightfoot 1999: 698)

In particular, nativists claim that language acquisition is accomplished by means of an innate capacity or *language faculty*. The initial state of the language faculty is called *Universal Grammar* (UG). It is defined as a set of parametric grammatical principles common to all natural languages. Since most parametres are generally binary in nature, they allow for very little variation and ban children from considering an infinite number of hypotheses during the process of language development. Given such a framework, much, if not most, of what we know of our own language is claimed to be part of our genetic endowment and, therefore, present at birth, rather than inductively learned from observation of the language around us. "The functional properties of this capacity develop along a regular maturational path, such that it seems more appropriate to speak of our knowledge of our language as growing rather than as being learned." (Anderson & Lightfoot 1999: 698).

However, an important part of the language acquisition problem has been shown to involve learning from the environment. Recent research from laboratories indicates that children solve many complex language-learning problems during their first year of life, as they attune their perceptual skills to the structure of their native language (Saffran et al., 1996a; Kuhl, 2000).

Thus, one of the most difficult issues within this nature-nurture debate is to actually figure out what exactly is innate and what is learned. Exposure to any form of natural language during the first years of life is doubtlessly essential for language acquisition to take place. Besides, everybody would agree with the fact that human beings are somehow biologically prepared to acquire a natural language. The problem is then how to identify the exact nature of this preparation and whether or not the ability to acquire a language involves the inheritance of specific linguistic structures (i.e. both specific to the linguistic domain and to the human species).

Arguments for a more general kind of linguistic nativism assume that what is innate about language is the manner and the procedures by which we process linguistic information. According to this view, human beings are assumed to have innate learning mechanisms which are unique and specific to language. Such mechanisms are, therefore, independent from other cognitive operations and procedures:

> [N]o features that are characteristics of only certain natural languages, either particulars of syntax, or phonology, or semantics, are assumed... to be innate. However, there are many reasons to believe that the processes by which the realized, outer structure of a natural language comes about are deep-rooted, species-specific, innate properties of man's biological nature. (Lenneberg 1967: 394; in Clark 2007: 373)

Other arguments, however, assume linguistic nativism not only regarding the learning mechanisms that are involved in the acquisition of natural languages, but also regarding the categories and the structures to be learned (Chomsky, 1965; 1981; Crain & Lillo-Martin, 1999). Not only that, but they further assume that the essence of language is grammar, which is understood as being completely independent from other levels of linguistic analysis as well as from other cognitive systems (Marantz, 1995;

Newmeyer, 1998). In this line, Newmeyer (1998) has described the autonomy of syntax using three interconnected ideas:

> the autonomy of syntax, which holds that there exists a cognitive system of nonsemantic and nondiscourse-derived syntactic elements whose principles of combination make no reference to system-external facts; the autonomy of linguistic knowledge from use, which postulates a system embodying knowledge of language that is characterizable independently of language use; and the autonomy of grammar as a cognitive system, namely the idea that there is a cognitivve system exclusively dedicated to language. (Newmeyer 1998: 94; in Hilferty 2003: 40)

Hence, it is easily assumed that, on the one hand, grammar is independent from the rest of linguistic domains (i.e. from phonology, the lexicon, pragmatics, etc.) and, on the other hand, that grammar is innate and part of our phylogenetic inheritance (Chomsky, 1965). Thus, within this representational innateness approach, the innate linguistic module does not contain things like innate learning mechanisms specific to language, but it includes actual grammatical content and structure.

One of the problems about theories of representational innateness is that they are very vague in their descriptions of what this innate grammar consists of, what exactly belongs to UG, or even the number and the kind of parametres which govern linguistic structure and which are part of our genetic endowment. In one of the few detailed descriptions of what constitutes UG, O'Grady (1997) provides a list of syntactic categories (i.e. both lexical and functional categories) which are understood as core elements of what is assumed to be part of human beings' innate grammatical knowledge. Within lexical categories, he distinguishes the categories of Noun (N), Verb (V), Adjective (A), Preposition/postposition (P) and Adverb (Adv). Within the framework of generative grammar, following the binary nature of most parametre settings, innate lexical categories are described as having different values for the two binary features ±N and ±V (Jackendoff, 1977; in Baker, 2003). Thus, elements like nouns are defined as [+N, -V], while elements like verbs are defined as [-N, +V].

One of the most prevalent arguments for linguistic nativism is the so-called *Poverty of the Stimulus Argument* (Anderson & Lightfoot, 2002; Chomsky, 1993; Crain & Pietroski, 2001; Laurence & Margolis, 2001; 2002; Pinker, 1994). According to this argument, the samples of linguistic data to which children are exposed during their first years of life are too impoverished and inconsistent, and the core aspects of grammatical structure that need to be acquired are only scarcely represented. Therefore, given the fact that language development cannot be accomplished with such impoverished linguistic representations, and that the input alone cannot account for all the linguistic aspects that need to be mastered, it follows that children must be born with innate *a priori* grammatical structure that might guide their processing of linguistic information and ultimately make language development a feasible task to complete. In Chomsky's words,

> adult speech to young children [offer] at best a degenerate version of a language – such speech [is] full of errors, hesitations, breaks in construction, retracings, pauses, and other disfluencies, repairs to vocabulary, to pronunciation, and so on, to the extent that children would necessarily have great difficulty both in learning what might be

> systematic in a language and in discerning what the structures were. (Chomsky in Clark 2009: 23)

The present dissertation aims at testing the empirical validity of such an argument. In particular, I will examine the way in which the *Poverty of the Stimulus Argument* translates into the word categorization problem and the nature of grammatical categories. As we shall see, one of the main problems regarding language universals and the binary nature of most grammatical elements within generative grammar is that they do not account for the linguistic diversity found across different natural languages.

Thus, for example, regarding primitive lexical categories, which are claimed to be innate, the framework assumes that all natural languages have such things as verbs, adjectives or nouns. While many typologists question such an assumption (e.g. Croft, 2003; Culicover, 1999), an aspect which is broadly accepted is that all natural languages have nouns. If this is true, then it follows that the noun category is part of our innate grammatical knowledge. Thus, from birth, children are supposed to have the knowledge that the language they are exposed to contains such things as nouns, and they only have to learn the particular lexical entries that belong to that *a priori* innately specified grammatical category. Furthermore, within nativist lines of argumentation, the noun category must be necessarily innate, given that with the limited linguistic experience to which children are exposed, it is impossible for them to build a grammatical category on the basis of experience alone.

The present dissertation will examine the extent to which such theoretical assumptions are plausible and valid. Chapter 2 will provide a general description of categorization in general, and the problem of categorization when it comes to linguistic elements. The main focus of the chapter will be to analyze nativist arguments for innate categories as well as the problems that such arguments bring about, ranging from the problem of cross-linguistic diversity and typology, to the assumption that both the input and children's learning mechanisms are too limited to account for the whole language acquisition task.

The end of chapter 2 will link the *Argument of the Poverty of the Stimulus* to the grammatical categorization problem, and this will link chapter 2 with chapter 3. One cannot assume that the linguistic samples that children hear are too impoverished for the acquisition of grammatical categories without closely examining the kind of linguistic input that children are exposed to. Chapter 3 will then review previous empirical work on the availability of cues in the input with which grammatical categories can be inferred. Furthermore, such cues appear to be not only exclusively syntactic in nature, but phonology and semantics also seem to converge and group individual words into large word classes.

Chapter 4 will present the main objectives of the present dissertation and the predictions that have been formulated. A first prediction is formulated in connection to the nature of individual sources of information for grammatical categorization. A second prediction will then be devoted to the kind of behaviour that such sources of information exhibit once they are combined and interact with one another.

In chapter 5, the methodological procedure undertaken in the present study will be described. The main way in which the linguistic input to which children are exposed can be examined is by analyzing language samples spoken by mothers and caregivers when addressing children. Chapter 5 offers a description of the corpora that have been considered for the purpose of the empirical test, as well as the decisions taken regarding the predictor variables to be considered in each of the tests and the statistical tools that have been used.

In chapter 6, the results obtained in the present study will be described. The chapter will be structured according to the research questions and the predictions established in chapter 4. Thus, an initial section with general descriptive data will be followed by the results obtained from the tests using individual sources of information for grammatical categorization. The rest of the chapter will describe the results obtained from the tests where two or three different sources of information were combined with one another.

The results presented in chapter 6 will finally be discussed in chapter 7. Once again, the structure outlined for chapter 4 will be kept. Thus, an initial discussion will deal with the behaviour and the distribution of each of the three sources of information for word categories in isolation. Then, the second part of the chapter will be devoted to the discussion and the conclusions that can be drawn from the evidence provided by the interaction of linguistic cues.

Finally, chapter 8 will offer general conclusions and the way in which the findings of the present study fit within the nature-nurture debate. The main objective is not to find the exact nature of children's learning capacities, or human beings' innate endowment for linguistic abilities, but to clarify one of the main problems that children have to face when learning a language (i.e. the problem of grammatical categorization) and to test the assumptions that have been broadly made in the generative grammar literature, namely the *Argument of the Poverty of the Stimulus*.

Chapter 2
Word categorization: Nature or nurture?

2.1. Categorization and cognition

Harnad (2005) defines categorization as "any systematic differential interaction between an autonomous, adaptive sensorimotor system and its world." (Harnad 2005: 21). Thus, categorization is defined as a *systematic* process, in the sense that it is not an arbitrary interaction between a system and its environment, but it is rather patterned. It is also a process which is performed by a sensorimotor system or being which is *adaptive*, since it develops and changes internally across time. And most importantly, categorization is also *differential*, since while categorization and recategorization processes are performed, the same input will never bring about exactly the same output across different categorization stages. In Harnad's words, "categories are *kinds*, and categorization occurs when the same output occurs with the same *kind* of input, rather than the exact same input" (Harnad 2005, 22). Therefore, abstraction is a general property of any categorization process.

Considering human beings as autonomous and adaptive sensorimotor systems in interaction with their environment, Harnad (2005) suggests that most of our categories are the result of learning and developmental processes. A learning process is described as the result of generating outputs in response to some kind of input on the basis of trial and error. However, a mere unsupervised and unconstrained categorization process that would cluster elements together based on their structural similarities will not work, since every member in each category is made up of a high number of features, which yields an infinite number of possible combinations and groupings:

> Every category has both an "extension" (the set of things that are members of that category) and an "intension" (the features that make things members of that category rather than another category). Not only are all things the members of an infinite number of different categories, but each of their features and each combination of features is a potential basis (…) for assigning the thing to still more categories. (Harnad 2005: 28)

So, categorization is a matter of taking all relevant features together correctly, depending on the demands of particular contexts and situations. In order to sort out this problem, supervision and corrective feedback are also necessary for successful categorization to take place. Thus, as long as there is enough data, as well as feedback and time, supervised learning can help categorization.

On linguistic grounds, one of the arguments that nativists put forward to claim that grammatical categories must be unlearned (and therefore innate and part of our genetic endowment) is precisely the impoverished representation of linguistic features in the environment as well as the lack of supervision and corrective feedback in the process of first language acquisition. According to nativists, in spite of that, target grammatical categorization is always successfully accomplished by children during a relatively short period of time.

Certainly, other researchers have pointed out the idea that there are in fact some human general learning abilities that need to be accounted for. To assume that something is "innate" is to assume that it is unexplainable. Nevertheless, if human beings are able to successfully build categories, then there must be a cognitive basis for that ability and there must also be enough data in the input in order for categorization to take place. Thus, one should not reject the notion of learning out of hand without previously examining the reliability with which features are represented in the input as well as human beings' abilities to cope with those features and use them to build abstract groups and categories.

Furthermore, despite there being an open-ended number of features in every to-be-categorized element to be taken into account, one should also consider the possibility that categorizing organisms might weight some features more heavily than others. Thus, it is not true that all available features are equally important: some features are more prominent than others, some are more salient than others, and some features might not even be detected at all.

Working memory and processing cognitive abilities might also constrain the kind of features that are considered and the span of features to be considered at every stage. Miller's (1956) classical paper on our limitations in processing information states that if we had to remember a whole sequence of *zeros* and *ones* and retrieve it correctly, a string of up to seven items in a row would be about our limit as far as working memory is concerned. But if we learned to recode individual elements into groups of three and give each group a new name (e.g. 001 would be called *one*, 010 would be called *two*, 011 would be called three, etc.), and we learned that code so that we could read strings automatically using the new code, then we could be able to remember three times as many individual items (i.e. seven complex items made up of three individual items each). Thus, for example, the sequence 110 would no longer be three items, but it would be a "chunk" that would work as an individual item. Then, new emerging chunks would be subject to the seven-item constraint in our memory. For Miller, recoding elements by grouping them in bigger chunks is a way of enhancing working memory.

Therefore, developing organisms might initially draw their attention to small individual elements and learn to group them in bigger chunks which, in turn, would bootstrap the emergence of newer and bigger strings. I will now turn to categorization in natural languages and the way in which the general cognitive categorization procedures that have been sketched so far can fit into the formation of grammatical categories.

2.2. Grammatical categories in natural languages

Grammatical categories are the groups to which individual linguistic elements are assigned. There are three different criteria under which elements are grouped into one category or another:

a) A morphological criterion: elements of the same category can all take the same common inflections. Thus, for example, all nouns in English can take plural inflection, or all verbs can take past tense inflection, but nouns do not take past tense inflection.

b) A syntactic criterion: elements of the same category can occupy similar syntactic slots and can be combined with certain elements. For example, all English nouns can perform the grammatical function of subject or direct object, and they can be preceded by determiners, while English verbs cannot.

c) A semantic criterion: elements of the same category have similar denotational meanings. For example, English nouns can refer to objects, while English verbs can refer to actions or events.

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. But how do children learn the basic grammatical categories of their language? How do they know that "*table*" is a noun, "*eat*" is a verb, and "*kiss*" can be both a noun and a verb? There are two possible answers to those questions. The first one is that grammatical categories might be innate. According to this assumption, our linguistic abilities to generate grammatical strings of words are neither the product of learning nor the product of evolution, but they are rather inherent properties of the structure of our brains. That is, from the moment children face the language learning problem, they already "know" that in the input they are exposed to individual items can be grouped around abstract grammatical categories like Noun or Verb. Under this assumption, the language learner's task is simply to acquire lexical items and to map those items to their corresponding word class.

A second explanation claims that grammatical categories might be learned. Under this assumption, the child might use information available from the environment to cluster similar words together and form some kind of rudimentary word classes that would later develop into adult-like grammatical categories. In what follows, both possibilities are analyzed in depth.

2.3. The innateness of grammatical categories

Mainstream nativist theories on language acquisition claim that children have innate constraints that make language development possible. Some theorists argue that such innate constraints include a complete grammar of human language in the form of universal principles that define common structural properties of all natural languages, together with a finite set of parametric variations that define the structural differences between languages (Baker, 2001; Chomsky, 1975; Crain & Lillo Martin, 1999).

From this perspective, the entire grammatical machinery of human language is considered to be innate. Hence, the set of possible grammatical categories (i.e. nouns, verbs, adjectives, etc.) must be similarly innate. Within such an approach, the child's task is simply to learn which lexical items belong to which syntactic categories.

The main explanation to support linguistic nativism states that, during the language acquisition process, highly complex linguistic forms develop extremely rapidly. However, the language input available to the young language learner is said to be both incomplete and sparsely represented, compared to children's eventual linguistic abilities. In sum, the kind of linguistic knowledge that children ultimately attain is "perfect" and it is impossible to extract perfect knowledge from the imperfect data of everyday language usage. Therefore, since it is impossible to learn language from the environment, language must be innate.

> Even if children could have learned that the adult grammar has U (given the right data), it does not follow – and it may not be at all plausible – that they acquired a grammar with U by learning. The requisite data may be unavailable to children (or anyone but trained linguists); available data might not be utilizable (by children); or the data might not be sufficiently ubiquitous to account for the knowledge of property U by all normal children, especially with respect to aspects of grammar where very young children exhibit adult linguistic competence. (Crain & Pietroski 2001: 151)

This argument is best-known as the *Poverty of the Stimulus Argument.* It mainly highlights the fact that, on the one hand, the input is too impoverished and children's linguistic knowledge extends far beyond the sample of utterances to which they are exposed (i.e. positive evidence is scarcely represented in the input). On the other hand, at the same time the input is said to be too rich, since it affords incorrect inductive generalizations which children never make (i.e. the input lacks sufficient negative evidence).

As stated in section 2.1., in order for successful categorization to take place, there must be (a) sufficient data which is reliable enough and from which elements can be categorized; and (b) corrective feedback and supervision that might guide the categorization process and ban wrong or misleading feature selections. For nativists, the linguistic environment to which language learning children are exposed does not contain data which is reliable enough (i.e. parental utterances are often incomplete and inconsistent) and it does not provide any form of corrective feedback either (i.e. parents never correct their children).

Therefore, grammatical categories must be innate. Within this approach, children might then be born with the innate intuition that there are such things as nouns and verbs in the language they are learning, and their task would simply consist in mapping individual lexical items into their corresponding grammatical category.

2.4. Problems to nativist approaches to grammatical categories

There are several empirical problems to such lines of argumentation, since neither cross-linguistic empirical studies nor research conducted with language learning infants wholly support such nativist views. In what follows, I will review some of the main counter arguments to nativist theories of language acquisition.

2.4.1. Grammatical categories and universals

If we assume that grammatical categories are innate and part of human beings' genetic endowment, then it might follow that all natural languages have the same grammatical categories (i.e. all natural languages have nouns, verbs, determiners, prepositions, etc.). However, current research in linguistic typology shows that not all possible grammatical categories occur in all natural languages (e.g. Croft, 2003, Culicover, 1999; Dixon, 1977). Thus, the child still needs to figure out which grammatical categories are realized in the language.

For example, while English has a single grammatical category for adjectives, Japanese distinguishes two different categories that would be equivalent to English adjectives (i.e. two distrinct grammatical categories contain words that denote properties). On the contrary, there are languages in which there are no adjectives, and properties are expressed by means of nominal elements, or even by verbs, as shown in the examples below from Quechua (1a) and Mandarin Chinese (1b) (Schacther 1985: 17-18).

(1) a. Rikashka: alkalde-kuna-ta ('I saw the mayors')
I-saw mayor-pl-acc

b. Liaojie de nühaizi ('a girl who understands')
understand rel girl

In Malagasy[1], most verbs are derived from nonverbal roots. The examples in (2) show that some verbs can be derived from adjectives and nouns by adding the verbal prefix *man-* to the nonverbal root (Hanitriniaina & Travis, 1998; quoted in Travis, 2005):

(2) a) ADJECTIVES:

lany ('used up')	→	*mandany* ('to use up')
madio ('clean')	→	*manadio* ('to clean')

b) NOUNS:

lavaka ('hole')	→	*mandavaka* ('to pierce')
doka ('flattery')	→	*mandoka* ('to flatter')

Also, according to Demirdache & Matthewson (1995), grammatical categories in St'át'imcets[2] seem to be underspecified. As seen in (3) and (4), the same lexical roots can appear in a verbal context (3a-c) or a nominal context (4a-c):

(3) a. **qwatsáts**-kacw ('you left/leave')

leave-2sing.subj

b. **smúlhats**-kacw ('you are a woman')

woman-2sing.subj

[1] A Western Malayo-Polynesian language spoken in Madagascar.

[2] A Salish language spoken in British Columbia.

c. **xzúm**-lhkacw ('you are big')

big-2sing.subj

(4) a. qwatsáts ti **smúlhats**-a ('the woman left')

leave DET woman-DET

b. smúlhats ti **qwatsáts**-a ('the one who left is a woman')

woman DET leave-DET

c. qwatsáts ti **xzúm**-a ('the big one left')

leave DET big-DET

Furthermore, besides cross-linguistic evidence from other oral languages, research shows that in some sign languages, several pairs of nouns and verbs which have a similar meaning also have the same formal features (i.e. same configuration, contour of movement, orientation and place of articulation) (Bouchard et al., 2005). Some examples are shown in (5):

(5) HELP / TO-HELP

TEACHING / TO-TEACH

INTERPRETER / TO-INTERPRET

Identical pairs are also frequently found in other categories. When such thing happens, the only way to determine the category to which each element belongs is by analyzing the context in which it is used:

(6) COMPETENCE / COMPETENT (identical nouns and adjectives)

TO-CONTAIN / IN (identical verbs and prepositions)

ORAL / ORALLY (identical adjectives and adverbs)

The basic distinction between nouns and verbs does not appear to be universal even by considering English data. We can also find some examples of multicategorial words in English (e.g. *hug*, *kiss*, *drink*, etc.), where noun-verb pairs have identical roots, and they only differ in the kind of syntactic contexts in which they occur and the kind of inflectional morphology that they take.

2.4.2. Learners' abilities

A second problem to nativist arguments is that, even if there were any innate *a priori* knowledge about specific grammatical categories, simply knowing that such categories exist does not solve the grammatical categorization problem, since children still have to determine which items in the input belong to which category. This leads us back to our previous question: how do children know that "*table*" is a noun and "*eat*" is a verb?

An assumption shared by nativist models is that analyses that lead to the formation of grammatical categories rely on input that has previously been segmented into words. The task of the language learning infant is simply to learn such word-like units from the input and classify them into their appropriate innate grammatical category.

However, recent research with infants in laboratories shows that children must perform complex word segmentation tasks before they are able to come up with word-like units and use them productively in grammatically correct strings. Such word-segmentation tasks require sophisticated perceptual and/or computational capacities on the part of infant language learners (Morgan, Shi & Allopenna 1996: 264). Being able to perform such tasks shows that infants are not as poor learners as the nativist framework assumes them to be, and that grammatical categorization might not be a matter of simply picking up pre-established units and cluster them within pre-given grammatical categories, after all. What follows is a review of infants' initial analyses of input aimed at solving the word segmentation problem.

2.4.2.1. Early word segmentation

The linguistic mainstream has typically modelled the way infants acquire knowledge of their language based on input understood as strings of words, rather than strings of sounds. However, accurate representations of strings of words are certainly very difficult for children to achieve, and yet without accurate identification of words or morpheme boundaries, syntax acquisition cannot proceed.

Speech lacks consistent physical markers of the location of word boundaries: words in fluent speech are not separated by pauses, or signaled by any other consistent feature occurring only at word onsets or endings (Cole & Jakimik, 1980; Aslin et al., 1996). Thus, the assumption that children begin language acquisition with accurate strings of words grossly misrepresents the information available in the input.

Identifying word boundaries in fluent speech is certainly one of the first and most complex tasks infants have to face when learning their native languages. Furthermore, it must also be an input- or data-driven task, at least in large part, since the structure of words and the constraints on what constitutes a syllable are so variable across languages that it is difficult to imagine how innate knowledge could solve this problem. In this sense, the most an infant could be born with would be the expectation that words actually exist, that is, that communicative messages can be delivered by means of linguistic chunks made up of smaller discrete units. Other than that, there can be no specific expectations as to the actual structure of words. Thus, infants at the preverbal stage cannot have any *a priori* set of words or word templates on which to build a lexicon.

Be this as it may, the word-segmentation problem must be solved, at least in part, by an interaction between infants' learning mechanisms and the particular language input the child is exposed to. It is the child's task, therefore, to tune in those cues which are relevant for word identification in the input they hear, and which perceptual features correlate with boundaries of words in their language (Jusczyk 1997).

In a pioneering study, Jusczyk and Aslin (1995) performed several experiments in order to test whether English-learning infants between six and nine months of age were able to recognize words from their language in fluent speech. Results showed that, by 7½ months, infants already display some abilities to extract familiar words from the speech stream. Moreover, infants' abilities to detect words from fluent speech also have important implications for other aspects of language acquisition. For example, the ability to recognize particular words in sentential contexts could facilitate a distributional analysis of the speech stream. Thus, if learners know that a particular word appears in combination with some other words, or if they know that a particular word can appear in certain locations within a sentence, but not in others, then learners might also be able to grasp the grammatical category of this word, or the syntactic function it plays within the sentence (Jusczyk & Aslin, 1995).

However, what are the exact cues that infants need to be familiar with in order to be able to parse the speech stream into word-like units? In particular, what sources of information might English-learners draw on when they begin to find word boundaries in English? In what follows, I will discuss several proposals that have been put forward in order to account for the kinds of perceptual cues that infants are most likely to attend to when they locate word boundaries in fluent speech.

2.4.2.2. Infants' sensitivity to prosodic cues

Over the past two decades, researchers have put forward the so-called *Phonological Bootstrapping Hypothesis*. This approach, also known as prosodic bootstrapping, contemplates the possibility that learners might extract directly from the speech signal information about regularities at other linguistic levels (i.e. morphology or syntax).

As far as word segmentation is concerned, the phonological-bootstrapping approach postulates that the prosodic structure of English words might serve as a cue to word structure. In particular, researchers have proposed a Metrical Segmentation Strategy (MSS), according to which English speakers would identify strong syllables with the initial syllables of content words (Cutler, 1990; 1996).

In order to test how successful English speakers might be when segmenting the speech stream using the MSS, Cutler & Carter (1987) examined the prosodic structure of words in an English corpus[3]. Over a corpus of 33,000 entries, they found that the most common word type in English was a polysyllabic word with stress on the initial syllable. Moreover, only about 17% of lexical tokens were polysyllabic words with weak initial syllables (Cutler & Carter, 1987; quoted in Cutler 1990: 106-107).

Furthermore, Cutler & Carter tested these data against a natural speech corpus, the *Corpus of English Conversation*, which consists of about 190,000 words of

[3] The MRC Psycholinguistic Database, which is a lexicon of about 98,000 words of which 33,000 entries have phonetic transcriptions. It is based on the *Shorter Oxford English Dictionary* (Cutler, 1990).

spontaneous British English conversation. Cutler & Carter found that only 10% of the lexical words in the whole corpus were iambic (i.e. polysyllabic words with initial unstressed syllables). The predominant pattern in most polysyllabic words in the corpus was trochaic (i.e. with primary stress on the first syllable).

Thus, based on the evidence provided by these data, the conclusion can be drawn that any listener encountering a strong syllable in English is very likely to find that syllable to be the onset of a word. On the contrary, weak syllables in English correspond to either the final syllable of polysyllabic words, or to monosyllabic grammatical words. Therefore, the MSS, that is, the strategy by which strong syllables are identified with word onsets, predicts quite an accurate segmentation of the speech stream in English.

Jusczyk et al. (1993) investigated the possibility that English-learning infants might also segment words from fluent speech on the basis of the location of strong syllables in utterances. In a series of experiments, Jusczyk et al. (1993) found that nine-month-old English-learning infants show a significant preference for trochaic words (i.e. words following a strong/weak stress pattern) over iambic words (i.e. words following a weak/strong stress pattern). No such preference was found with a younger group of six-month-old English-learning infants, suggesting that the preference for trochaic words develops as a result of infants' increasing familiarity with the prosodic features of the language in their environment.

A similar experiment was performed by Echols (2001). She familiarized English-learning infants with three-syllable nonsense sequences which contained stress on the medial syllable. Then, at test trials, she presented infants with the same stimuli as in the familiarization phase, but this time sequences contained a pause either before or after the medial stressed syllable. The results from this study show that nine-month-old infants extract trochaic sequences from the input more readily than iambic sequences, since they preferred those sequences which contained pre-stress pauses over those sequences which contained post-stress pauses.

What do 7½ -month-old infants do, however, when they are exposed to words that conform to the weak/strong pattern instead? The MSS predicts that English-learning infants will have difficulties with weak/strong patterns, since they identify word onsets with strong syllables, and would therefore missegment a weak/strong word when presented with it.

To test the possibility that 7½ -month-old English-learning infants might missegment words with the weak/strong pattern, Jusczyk et al. (1999) familiarized infants with weak/strong words (e.g. *guitar*, *device*, *surprise*, *beret*), and then tested infants on passages containing these words. Contrary to the results from the previous experiment, but consistent with the MSS hypothesis, after familiarization with weak/strong words, infants showed no significant listening preferences for the passages which contained the familiarized weak/strong targets over the passages which did not.

However, in an additional experiment, infants were familiarized with just the strong syllable of the weak/strong words (e.g. *tar*, *vice*, *prize*, *ray*), and then they were tested on the same passages as in the previous experiment. After familiarization with just the strong syllables of weak/strong words, infants showed a significant listening preference for the passages that included the weak/strong words whose strong syllables matched the syllables at the familiarization phase. Therefore, 7½-month-old English-learning infants indeed appear to be missegmenting weak/strong words at the strong syllable boundary (Jusczyk et al., 1999).

On the basis of the evidence provided, we can conclude that (1) the input contains a number of prosodic cues useful for word segmentation, (2) that infants are sensitive to those prosodic cues, and (3) they can make use of them to perform one of the first and most fundamental linguistic tasks they face: that of identifying words in speech (Echols, 2001; Jusczyk et al., 1999; Swingley, 2005). However, although prosody appears to be a valuable cue for segmenting the speech stream into word-like units, it cannot be the only cue available to infants. As the available empirical evidence indicates, the problem of correctly determining the boundaries of weak/strong words, or many other words with more than two syllables, still remains.

Besides, although the MSS predicts successful segmentation of the English speech stream, not all languages provide reliable prosodic cues to word boundaries. The MSS for English is based on the opposition between strong and weak syllables, which is an important feature of English phonology. In other languages, however, the contrast between strong and weak syllables might not be a relevant feature for word structure and segmentation. Spanish and Catalan, for example, are not defined as prosodically stress-timed languages like English, but they are syllable-timed languages. Syllabic segmentation is, then, much more relevant in these languages than stress-based segmentation. Japanese, on the other hand, is a mora-timed language. Thus, sensitivity to moras rather than strong syllables is required so that Japanese-learning infants could properly segment the speech stream of their language.

Moreover, as pointed by Saffran et al. (1996b), in order to learn language-specific prosodic patterns and be able to apply any parsing strategy like the MSS, the learner should have some additional source of information about word boundaries independent of prosodic cues. Without such information, there would be no way for the infant to determine that strong syllables correspond to word onsets, and not to the middles or ends of words. In sum, while prosody is likely to be quite helpful for infants attempting to locate word boundaries in the speech stream, it may not always be present in the input or accessible to the learner.

There is, however, another type of information which is always present in the input, but which researchers have traditionally considered to be too complex for language learners to use: distributional cues. In fact, if we examine the results from the study by Jusczyk et al. (1999) closely, a peculiar pattern emerges: for the strong/weak words, previous familiarization with just the strong syllables of these words did not lead to infants' preference for the passages containing the strong/weak words. However, for the weak/strong words, previous familiarization with just the strong syllables did lead to infants' preference for the passages containing the weak/strong words.

This pattern of results can be accounted for in terms of the distributional properties of the input. That is, in addition to information about the strong syllables to locate word boundaries in fluent speech, English-learners might be using distributional cues to determine word onsets and ends. Thus, English-learning infants seem to be identifying strong syllables with the onsets of new words in fluent speech. This leads them to missegment words which begin with weak syllables.

However, infants also seem to be attending to the distributional properties of the input in order to determine where a word is likely to end. Learning about distributional cues may then provide the infant learner with additional information that would override the MSS. The following section examines some recent findings from research which suggest that distributional information is indeed available in the input and it might be accessible to language learners in their attempts to parse the speech stream into word-like units.

2.4.2.3. Infants' early distributional analyses

Words in any language can also be defined in distributional terms. Each word is a lexical unit which consists of a sequence of phones in a fixed order. These units are typically unbreakable, that is, pauses or other words generally occur between words, rather than in the middle of them. Thus, language learners also experience complex statistical information in the input. This statistical information will take the form of strong correlations between sounds found within words (since they are part of the same lexical unit) and weak correlations between sounds found across word boundaries (since they are adjacent to one another only accidentally).

Any learner of English would perceive the spoken phrase "*yummy cookie*" as a continuous string "*yummycookie*". How is the learner going to parse this string into its two constituent words? Over a corpus of English, word-internal syllable pairs (e.g. "*yu*" + "*mmy*", "*coo*" + "*kie*") will occur more frequently than word-external syllable pairs, which are relatively unconstrained (e.g. "*mmy*" + "*coo*"). So, given the syllable "*yu*", we can calculate the likelihood with which the syllable "*mmy*" will appear. If the probability of occurrence of "*mmy*" given "*yu*" is high, then the sequence "*yu+mmy*" is likely to be a word. This is what experts have called *transitional probability* (Saffran et al., 1996b). The transitional probability of element *Y*, given element *X*, is computed as shown in (7):

(7) $$\frac{\text{Frequency of pair } XY}{\text{Frequency of } X}$$

A high transitional probability (i.e. a high coeficient obtained from (7)) indicates that the presence of *X* strongly predicts the presence of *Y* immediately after. When applying this to syllables in words, we would expect a different transitional probability in (8a) and in (8b):

(8) a. $\dfrac{\text{Frequency of “}\textit{yu}\text{” + “}\textit{mmy}\text{”}}{\text{Frequency of “}\textit{yu}\text{”}}$ b. $\dfrac{\text{Frequency of “}\textit{mmy}\text{” +“}\textit{coo}\text{”}}{\text{Frequency of “}\textit{mmy}\text{”}}$

A high transitional probability in (8a) would indicate that there is no word boundary between “*yu*” and “*mmy*”, that is, that “*yummy*” is a word. On the contrary, transitional probabilities are lower for syllable pairs across word boundaries. Thus, a low coeficient in (8b) would indicate that “*mmycoo*” is not an English word, that is, that there is a word boundary between “*mmy*” and “*coo*”.

Some of the advantages of the statistical learning approach, as compared to the phonological bootstrapping approach, is that information about transitional probabilities will be equally available in the input to all learners. Moreover, statistical cues to word boundaries are present in all natural languages, regardless their acoustic or prosodic features.

Corpus-based models of speech processing (e.g. Brent & Cartwright, 1996) have shown that this kind of statistical information is sufficient *in principle* for word segmentation. Besides, this type of information is consistently available throughout all kinds of words. Given a learner who is able to calculate transitional probabilities over co-occurring syllables, the input contains sufficient information from which the boundaries of words can be calculated. Are human beings such learners?

In order to test adult subjects' ability to use transitional probabilities as cues to word boundaries, Saffran et al. (1996b) designed an artificial language learning task for adult subjects in which transitional probabilities were a key element for the extraction of word-like units from the speech stream. The artificial language that was designed for the test consisted of English phonemes which were combined to make up 12 different syllables with a “consonant + vowel” structure. These syllables were used to generate six trisyllabic words (i.e. *babupu*, *bupada*, *dutaba*, *patubi*, *pidabu*, and *tutibu*). The transitional probabilities within the words themselves varied, ranging from 0.31 to 1.0. The transitional probabilities between syllables spanning a word boundary were lower (range 0.2 to 0.1).

For the familiarization phase, these words were concatenated to create speech strings. The same word never occurred twice in a row, to prevent that such repetitions might ease the computation of transitional probabilities. All word boundaries were removed from the text. The stimuli were read by a synthesizer which did not insert any acoustic cues to word boundaries and which produced equivalent levels of coarticulation between all syllables. The only indication to word boundaries in the speech strings were, therefore, the different transitional probabilities of syllables within and across words.

For the test phase, half of the adult subjects were given a forced-choice test between words from the language and nonwords. Nonwords consisted of syllables from the language in a novel order, which had not been exemplified in the speech stream (not even across word boundaries). Thus, the transitional probabilities between each of the syllables in the nonwords were zero. The other half of the adult subjects received a forced-choice test between words and part-words. Part-words were made up of two syllables from a word plus an additional syllable.

As for the results, in the nonword test condition, the mean score was 27.2 of a possible 36 (76%). All words were learned at a level significantly better than would be expected by chance. The adult subjects' performance on the part-word test was slightly worse. The mean score was 22.3 out of 36 (65%). Worse results in the part-word condition were already expected, since part-words were more confusable with words than were nonwords. However, adult subjects in this more difficult condition still performed significantly better than would be expected by chance. Moreover, performance was superior precisely on the words containing higher transitional probabilities. Thus, taken together, these data support the hypothesis that the strength of the statistical relationship between different pairs of syllables affects adult subjects' ability to learn word-like units from the speech stream (Saffran el al. 1996b: 613-614).

A further piece of evidence for this hypothesis lies in the distribution of errors that subjects in the part-word condition made. In the part-word test, subjects false alarmed (i.e. they chose a part-word as a word incorrectly) when part-words contained the final two syllables of words in 48% of the cases. However, when part-words contained the initial two syllables of words, subjects only false alarmed for 29% of all the test items. So, subjects confused words with part-words which looked like the ends of words more often than part-words which looked like the beginnings of words. It appears that the ends of words were easier for adult subjects to learn.

Why would the ends of words be easier for learners? Logically speaking, any cue which signals the end of a unit also signals the beginning of the following one. Why is information about word boundaries used to learn about the end of the last word more readily than the beginning of the next word?

Several other lines of research have hypothesized that the ends of words are more prominent or salient to language learners. In particular, it has been claimed that final syllables tend to be lengthened in English, and this final lengthening may be especially exaggerated in infant-directed speech. Because these syllables are salient, they are readily extracted from fluent speech and stored as part of the initial representation of words (Echols & Newport 1992; Echols 2001).

In order to test the effects of the salience of final syllables in word segmentation, Echols (2001) familiarized English-learning infants with trisyllabic nonsense words which were stressed either on the medial or final syllable. After familiarization, infants were tested on two types of stimuli: one with a change in the medial syllable and one with a change in the final syllable. For example, if an infant heard *mobúti* at the familiarization phase, she would then hear two types of stimuli: *modúti* and *mobúpi*. The prediction was that infants would show interest in those stimuli which contained noticeable changes, given the assumption that changes are noticeable when they occur in salient syllables (i.e. final syllables). Results from this experiment show that infants attended significantly longer to stimuli containing changes in final syllables. Thus, the hypothesis that final syllables are attended to because they are prosodically salient is borne out. However, in Echols' (2001) study, unstressed final syllables were always part of a trochaic sequence, that is, they were part of a strong/weak stress pattern, which is typical of most English content words. As I pointed out above, empirical evidence supports the idea that English-learning infants are sensitive to the typical stress pattern of words in their language in their first year of life. Consequently, in Echols' (2001) study, it is not possible to determine whether infants attended to unstressed final syllables because those syllables were final (i.e. salient) or because those syllables were extracted and stored as part of a trochaic sequence.

Another possible account of why the ends of words are more readily stored concerns the way in which transitional probabilities are worked out. The mechanism shown in (7) computes, for each syllable, the probability that the next syllable will follow. If the transitional probability from this computation is low, a word boundary is hypothesized. For example, if the frequency of *dapu* is low in relation to the overall frequency of *da*, then *dapu* is likely to cross a word boundary. While this computation tells us both that *da* is the end of a word and *pu* is the beginning of a new word, the computation has been performed in relation to the overall frequency of *da* not *pu*. It is possible that the learner keeps the information that *da* is the end of a word because it is *da* that has anchored the computations required to discover a word boundary. This would, therefore, provide additional evidence that transitional probabilities are indeed used by language learners as cues to solve the word segmentation problem. "A predisposition to pay attention to the ends of words is a logical side effect of segmentation cued by transitional probabilities" (Saffran et al. 1996b: 615).

In order to test the potential use of transitional probabilities for word segmentation in first language acquisition, Saffran et al. (1996a) carried out an additional experiment in which they tested whether eight-month-old infants could extract information about word boundaries only on the basis of the statistical information present in fluent speech. Following Jusczyk & Aslin's (1995) methodology, Saffran et al. (1996a) familiarized eight-month-old English-learning infants with two minutes of a continuous speech stream created out of the concatenation of four three-syllable nonsense words repeated in a random order. As in Saffran et al. (1996b), the only cues to word boundaries were the transitional probabilities between syllable pairs.

After familiarization, infants were first tested on the discrimination between "words" from the artificial language, and nonwords (i.e. three-syllable units with the same syllables as words, but presented in an order which was unexemplified in the stimuli used at the familiarization phase). Infants listened significantly longer to nonwords (8.85 seconds) than to words (7.97 seconds), showing a novelty effect. As in Saffran et al.'s (1996b) study with adults, infants were then tested on the distinction between words and part-words (i.e. foils made up of the final syllable of a word and the first two syllables of another word). Again, infants showed a significant discrimination between the word and the part-word stimuli, with longer listening times for part-words (7.60 seconds) than for words (6.77 seconds). Thus, given the results obtained in these two experiments, the conclusion can be drawn that eight-month-old infants are able to extract sequential statistical information from fluent speech after only two minutes of listening experience (Saffran et al., 1996a).

Since the stimuli used in Saffran et al. (1996a) was synthesized speech, it could be argued that infants' listening preferences simply correspond to statistical information in acoustic stimuli, but not in natural language. To exclude this possibility, Johnson & Jusczyk (2001) replicated Saffran et al.'s (1996a) second experiment (i.e. the word vs. part-word discrimination test) with naturally produced stimuli, and not synthesized speech. Similar results were obtained with natural stimuli: on average, infants listened significantly longer to part-words (8.25 seconds) than to words (6.49 seconds), showing a novelty effect. Thus, even if Saffran et al,'s (1996a) synthesized stimuli were not perceived as speech, Johnson & Jusczyk's (2001) study give additional evidence that infants can make use of the statistical information available in the input to locate word boundaries in the speech stream.

Therefore, it seems that distributional cues are not too complex for either adult or infant learners to exploit. Even when accompanied by no other cues or word boundary markers, the statistical differences between word-internal and word-external sequences of syllable pairs can be used by language learners. This suggests that language learners are sensitive to subtle aspects of the statistical properties of language input. This sensitivity might help learners acquire other aspects of language as well. For example, keeping track of the statistical frequency of occurrence of words like *the* and referents like *dog* might help learners acquire the grammatical category of nouns and their syntactic properties (e.g. the fact that they are preceded by determiners).

2.4.2.4. Infants' sensitivity to combined cues

The empirical evidence reviewed so far suggests that language learners can segment the speech stream into word-like units on the basis of statistics alone. However, learners never hear speech with no more cues to word boundaries than transitional probabilities. In infants' natural environment, the input contains multiple redundant language-specific cues to word boundaries. Thus, it is not clear how strong statistical information is in word segmentation, as opposed to other types of cues which might also be available.

As seen in section 2.4.2.2, young English-learning infants can segment fluent speech into trochees, since they already know that strong/weak stress patterns correlate with word structure in English. Paralelly, as seen before, English-learning infants can segment fluent monotone speech into word-like units on the basis of statistics alone. This poses a difficult chicken-and-egg problem as far as word segmentation is concerned: do infants use distributional analyses to discover prosodic regularities in their language, or do they use prosody to isolate sections of speech upon which distributional analyses are conducted?

In order to test the strength of transitional probabilities relative to prosodic cues, Johnson & Jusczyk (2001) carried out an experiment in which prosodic cues were pitted against statistics. In order to do so, they familiarized eight-month-old infants with stimuli in which prosodic and transitional cues provided contradictory information about word boundaries. Thus, the speech stream used in familiarization was similar to that used in Saffran et al. (1996a): the only cues to word boundaries were transitional probabilities. However, Johnson & Jusczyk (2001) added a conflicting cue: the first syllable of part words was stressed. Orthographically, the speech stream was something like the string in (9), where stressed syllables are capitalized:

(9) *tibudogolaTUdaropitibudodaroPIgolatu*

The speech stream in (9) contains two conflicting cues to word boundaries: transitional probabilities indicate that *TU* ends a word (i.e. the word *golatu*), whereas stress indicates that *TU* begins a new word, because stressed syllables correlate with word onsets in English. The prediction was that infants would listen longer to part-words if statistics override stress (i.e. the results from Saffran et al. (1996a) would be replicated). On the contrary, infants would listen longer to words if stress cues override statistics.

Results from this experiment indicate that, on average, infants listened longer to words (8.0 seconds) than to part-words (6.59 seconds). This difference was statistically significant. The same pattern of results emerged in a similar experiment by Thiessen & Saffran (2003), in which nine-month-old infants were exposed to an artificial language speech which was entirely iambic (i.e. all words followed a weak/strong stress pattern). Infants in this experiment also missegmented the stimuli and treated part-words as words, that is, they segmented the speech stream at stressed syllable occurrences, identifying strong syllables with word onsets. Thus, on the basis of the evidence available, the conclusion can be drawn that, at eight and nine months old, when stress and statistical cues conflict, infants tend to rely more heavily on stress cues than on statistical cues when they segment the speech stream into word-like units.

This pattern of results lead Johnson & Jusczyk (2001) claim that stress is the earliest cue used by English-learning infants for word segmentation. Infants seem to rely first on stress patterns and segment fluent speech into trochees. It is later on during development that they perform distributional analyses to correctly segment English words that do not follow the dominant strong/weak stress pattern (e.g. *guitar*, *surprise*, etc.).

However, if stress is the earliest cue for word segmentation, how can infants discover that words in English are typically stressed on their first syllable and not the last, if they are not familiar with any words yet? Infants should segment a few words first in order to become aware of the fact that words tend to be stressed on their first syllable. But how can they segment a few words first, if stress patterns are their first cues to word boundaries?

One possibility might be that infants might be using statistical cues when they first begin segmenting the speech stream. This strategy would allow them to isolate words from fluent speech and detect the dominant stress pattern of English (Thiessen & Saffran, 2003). Thus, if statistical cues are the earliest that infants can attend to when locating word boundaries, then infants should favor statistical cues over prosodic cues at the youngest age at which they can segment words from fluent speech. The weight of empirical evidence suggests that, by 7½ months of age, infants are already able to identify word-like units in the speech stream, while six-month-old infants are not able to do so (e.g. Jusczyk & Aslin, 1995; Jusczyk et al., 1999).

In order to test whether there is an age group younger than eight months old that weight statistical cues more heavily than stress cues for word segmentation, Thiessen & Saffran (2003) performed an experiment with infants between 6½ and seven months of age. These younger infants were familiarized with similar stimuli as that used with older infants in previous studies, that is, a speech stream in which all words followed the weak/strong stress pattern. An example of such a speech string is shown in (10). Stressed syllables are capitalized and word boundaries are marked as #:

(10) diTI#buGO#doBI#daPU

Thus, the artificial language created consisted of iambic words like "*diti*", "*bugo*", "*dobi*" and "*dapu*". Infants were then tested on part-words like "*tibu*". Results showed that those younger infants weighted statistical cues more heavily than stress cues. Thus, unlike their older (eight- and nine-month-old) counterparts, 6½ - and seven-month-old infants were not misled by stress cues in fluent iambic speech, and they segmented words correctly at word boundaries, extracting iambic word-like units, and therefore, relying on transitional probabilities within syllable pairs, rather than identifying strong syllables with word onsets.

Such data reveal the following developmental pattern as far as word segmentation is concerned. By the age of six months, infants are not able yet to identify words in fluent speech, although they are already familiar with a number of acoustic and phonological properties of their native languages (Jusczyk, 1997; Kuhl, 2000). At the age of 6½ and seven months, infants are already able to extract word-like units from fluent speech. In doing so, they rely primarily on transitional probabilities within syllable pairs (Thiessen & Saffran, 2003).

Their early statistical learning abilities used to segment their first words allow infants to acquire an initial lexicon which is large enough for them to learn about the correlation between stress patterns and word structure in their language. Then, knowing that strong syllables typically correlate with word onsets, 7½ - to nine-month-old English-learning infants consistently identify trochaic stress as a cue to word boundaries and rely primarily on stress cues, even when stress cues conflict with statistical information (Johnson & Jusczyk, 2001; Jusczyk et al., 1999; Thiessen & Saffran, 2003). At around eleven months, infants become less reliant on stress cues, probably because they are now aware of the existence of iambic words in English, and they favor statistical information over stress cues again (Johnson & Jusczyk, 2001). Therefore, one of the milestones in word segmentation appears to be the progression from attention to one single cue to attention to multiple cues to word boundaries.

However, we do not yet know how these learning abilities for speech segmentation may generalize to the larger question of language acquisition as a whole. The problem of segmenting basic units from a large unsegmented speech stream is a very general problem, characteristic not only of initial levels of language processing, but also of general perception in other domains (Saffran et al., 1999). Nevertheless, it is also very unlikely that such complex statistical analyses of the speech stream might just serve a single purpose, namely that of segmenting the speech stream. Could such strategies simply be abandoned at all once word segmentation tasks are complete? It would be rather unnatural to assume so. In Morgan's words:

> Since perceptual analyses on the input logically must (...) succeed before syntactic or semantic analyses can proceed, we think it reasonable to ask whether such perceptual analyses might furnish additional linguistically useful information that, for example, could be used to separate words into classes that are forerunners of grammatical categories (Morgan, Shi & Allopenna 1996: 264).

2.4.3. "Poverty of the stimulus" arguments

As has been previously described, nativist accounts on language acquisition claim that the input to which language learning infants are exposed to cannot be the source of core aspects of linguistic knowledge. Language is essentially unlearnable and, therefore, it must be a kind of human instinct (Anderson & Lightfoot, 1999; 2002). The claim that children's linguistic knowledge results from a specific innate capacity rather than from inductive observation of their linguistic input is supported by the fact that children have been shown to know things that they could not have learned from observation or any plausible teaching.

Nevertheless, recent studies that analyze child-directed speech in depth have provided evidence that suggest that the linguistic envirnonment to which children are exposed has been underestimated. Thus, for instance, findings from studies of neural activity suggest that one-year-old infants attend more to child-directed speech than to adult-directed speech. This suggests that child-directed speech acts as an indicator of potentially meaningful streams of speech and it triggers brain activity in very young infants (Zangl & Mills, 2007).

Child-directed speech has also been shown to facilitate language-related tasks such as word segmentation (Thiessen & Saffran, 2005). Furthermore, adults have been reported to be more fluent when they talk to young children than in adult-to-adult interactions, as there are fewer false starts, mispronunciations or hesitations in child-directed speech (Sachs, Brown, & Salerno, 1976; Snow & Ferguson, 1977).

Nativist accounts of language acquisition as far as word categorization is concerned hold the view that knowledge about grammatical categories is innately specified. Such a view is based on the assumption that there is insufficient evidence in the child's language environment to enable these properties to be learned from the input. However, one cannot deny the informativeness of the linguistic input without exploring its potentials to its limits, and several studies have already pointed out the problems of assuming an impoverished stimulus beforehand (e.g Feijóo, 2007; Pullum, 1996;

Pullum & Scholz, 2002; Sampson, 2002; Scholz & Pullum, 2002; Scholz & Pullum, 2006).

If children's innate endowment includes not only the necessary procedures to discover the basic grammatical categories that are realized in their language, but also the relevant cues which signal grammatical category membership, such cues should be universally valid across languages. However, it is very unlikely that there exists such set of universal cues for grammatical categorization common to all natural languages.

Thus, children must be able to identify that some properties of words actually correlate with word classes, that is, that words can be grouped in grammatical categories and, depending on the category to which they belong, they can enter into particular grammatical relationships and perform certain syntactic functions. What is more, children should be able to identify such properties from the linguistic samples to which they are exposed. The following chapter provides a review of the kind of cues which, being sufficiently represented in the input, might yield accurate categorization of linguistic elements.

Chapter 3: Noun categorization: What is in the input?

3.1. The availability of cues for grammatical categorization

Any approach which describes a learning model based on cues must consider the validity of such cues (i.e. their availability and reliability in the input), as well as the strength of those cues and the cost of processing them (Kail, 2000). Studies of cues that are effective in identifying the grammatical category of words have focused on properties that are either internal or external to words. External properties are those that determine the use or function of the word from its context (i.e. distributional or semantic information). In contrast, effective information for the same purpose can also be found within the word itself. Such internal cues include phonological or prosodic information. Thus, while phonology and distributional information were useful for word segmentation, they appear to be equally useful for more complex linguistic tasks, like word categorization. Crucially, access to semantic information and word meaning might also be a powerful cue to signal category membership of words in the stream of speech. In what follows, I will revise every type of cue and its role in grammatical categorization in turn.

3.1.1. Phonological cues

Young language learning infants have no access to word meaning, especially prelinguistic infants at the first stages of development. Provided we assume no innate grammatical endowment, young language learners will have no initial access to grammar either. Thus, during the onset of the language learning process, phonological cues are the only ones from which young infants might start building a linguistic system. As seen before, phonological information is already available and useful for children's word segmentation tasks. Furthermore, a number of studies have also provided evidence for the usefulness of phonological information as a key element for the access to grammatical properties of language (e.g. Brooks et al., 1993; Cassidy & Kelly, 1991; 2001; Cutler, 1993; Gleitman & Wanner, 1982; Jusczyk, 2001; Morgan 1986; Peters, 1997). So, phonological cues might also be a useful source to help children determine the distinctions between grammatical categories.

A range of phonological cues have been proposed in the literature to correspond to particular syntactic categories in English. On the one hand, some of the proposed cues are related to distinguishing open class words (i.e. nouns, verbs, adjectives, etc.) from closed class words (i.e. prepositions, determiners, conjunctions, etc.). Such an initial broad categorization might help children establish a distinction between lexical items and function words (Monaghan, Christiansen & Chater, 2007; Monaghan, Chater & Christiansen, 2005; Morgan, Shi & Allopenna, 1996). This might help them further categorize lexical items in further subcategories more accurately. In English, among the cues that have been said to be relevant in the open *versus* closed class word distinction, the following are the most significant ones (Monaghan, Chater & Christiansen, 2005):

- *Word and syllable duration*: open class words tend to be longer and have longer syllables than closed class words. Closed class words, on the other hand, tend to be acoustically minimal, although they occur with very high frequency. Furthermore, the status of function words as free morphemes is often weakened: they can become cliticized or shortened, and therefore, they may lose syllabicity.

- *Consonant clusters*: open class words are more likely than closed class words to contain consonant clusters. In fact, the inventory of consonant phonemes found in function words is much smaller than that found in content words.

- *Vowel quality*: closed class words are more likely to contain centralized vowels than open class words. Furthermore, these centralized vowels are often reduced to *schwa*. On the other hand, open class words tend to have full low or high vowels.

- *Consonant position*: closed class words are less likely than open class words to contain consonants in word onsets. Even if they do, the kind of consonants that closed class words have in word onset position are different from the ones in open class words (e.g. /ð/ only occurs word-initially in closed class words in English).

Additionally, some other phonologial cues have been found to be relevant to distinguish some open class words from others (Kelly 1992, 1996; Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007). Furthermore, empirical studies reveal that young language learners as well as adults are aware of such cues and their correlation to grammatical categories (Cassidy & Kelly, 2001; Farmer, Christiansen & Monaghan, 2006; Fitneva, Christiansen & Monaghan, 2009; Kelly & Bock, 1988; Monaghan, Chater & Christiansen, 2003). These are the cues which are particularly relevant to the topic under consideration, since they have been reported to be especially useful in the noun *versus* verb distinction. Among the set of proposed cues for English, there are some of them which are highly reliable and have been reported to be quite accurate and consistent as far as word categorization is concerned. Others, on the other hand, while still useful and informative, are considered less consistent or only applicable to a limited range of words in the English lexicon. Among the set of highly reliable cues, the following pair are the most relevant ones:

- *Stress*: disyllabic nouns tend to be stressed on the first syllable, while disyllabic verbs tend to be stressed on the last syllable. In fact, analyses of the English noun-verb stress difference show that there is not a single noun-verb homograph in English in which the verb has first syllable stress and the noun has second syllable stress. So, if the noun and verb versions of an English word contrast in stress at all, the noun is always stressed on the first syllable and the verb is always stressed on the second syllable. Besides, a further examination of other disyllabic nouns and verbs in English (i.e. not exclusively noun-verb homographs) showed that 90% of the words with first syllable stress were nouns, while 85% of words with second syllable stress were verbs (Kelly, 1996; 1992).

  Furthermore, Kelly (1988) has also found evidence of such correlations between word class and stress pattern in examples of grammatical category conversion during the course of the diachronic evolution of the English language. Thus, over the past centuries, English nouns with second syllable stress have been more likely to develop verb uses, while verbs with first syllable stress have been more likely to develop noun uses. In fact, noun-verb homographs like *record* are historically and semantically related, and both once possessed the same stress pattern at a given period in the history of English. Later on their respective pronunciations diverged by means of a shift in the stress pattern of the verbal form (Sherman, 1975).

- *Syllables*: nouns contain more syllables than verbs. Experimental evidence suggests that trisyllabic linguistic units are typically associated with nouns, while monosyllabic linguistic units are associated with verbs (Cassidy & Kelly, 1991). In their study, Cassidy & Kelly (1991) found a very strong correlation between syllable number and grammatical category. In particular, the probabilities that one, two, three and four syllable words were nouns and not verbs were 38%, 76%, 92% and 100% respectively. Furthermore, children seem to be aware of that difference in vocabulary acquisition (Cassidy & Kelly, 2001).

  As with the stress pattern distinction, evidence was also found that the syllable number difference between nouns and verbs spans for several centuries in the history of the English language. Cassidy & Kelly (1991) examined words entering the English lexicon from the twelfth to the twentieth century and found that novel nouns contained significantly more syllables than verbs in each of the centuries examined.

Other phonological features have also been found to be correlated with the grammatical categories of words, although such correlations are weaker or they are just true of a subsample of the English lexicon (Kelly, 1992 ; 1996 ; Durieux & Gillis, 2001). Among those, the following ones are the best-known:

- *Word duration*: as said before, nouns have been described as being significantly longer than verbs as far as syllable number is concerned. The superiority of nouns compared to verbs as far as length is concerned has also been found at other levels within the word. Thus, controlling for syllable number, nouns tend to contain more phonemes than verbs (Kelly, 1996; Monaghan, Chater & Christiansen, 2005). The difference has also been found by comparing monosyllabic nouns and verbs. Even for noun-verb homonyms like *coach*, syllable duration was found to be significantly longer for the noun version than for the verb version (Kelly, 1992).

- *Vowel height and quality*: nouns tend to have more low vowels, while verbs tend to have more high vowels (Kelly, 1996; Monaghan, Chater & Christiansen, 2005). Besides, vowels in nouns tend to be back, whereas vowels in verbs tend to be front (Soreno & Jongman, 1990). However, empirical research shows that such differences only arise among words of high frequency (i.e. token frequencies of at least 250 per million).

- *Consonant quality*: nouns are more likely than verbs to have nasal consonants. Furthermore, if a word finishes in a consonant, it is more likely to be voiced if the word is a noun rather than a verb.

As said before, the individual magnitude of each cue is not the same, as some cues are weaker than others. Besides, the literature suggests that, taken individually, each phonological cue is not very powerful on its own, in the sense that each cue alone does not successfully categorize a significant sample of nouns or verbs. However, taken together, they yield an accurate and successful categorization of nouns and verbs. Thus, provided the constellation of cues outlined above is available in the input, it will significantly reflect the overall phonological shape of broad grammatical categories.

Furthermore, as initial broad categorization and analysis proceeds, the child will be able to deal with smaller linguistic samples to be analyzed, where correlations between cues and categories might be stronger. Besides, as their development proceeds, learners will be able to cope with other sources of information in combination with phonology.

3.1.2. Semantic cues

3.1.2.1. The Semantic Bootstrapping Hypothesis

Noam Chomsky (1975, 1993) maintains that the vocabulary acquisition problem is similar to the problem of the acquisition of other aspects of language. Thus, the principles of the *Poverty of the Stimulus Argument* are sometimes kept within nativist theoretical frameworks, even for word learning tasks:

> The pervasive problem of "poverty of the stimulus" is striking even in the case of simple lexical items. Their semantic properties are highly articulate and intricate and known in detail that vastly transcends any relevant experience. (Chomksy, 1993; quoted in Bloom 2000: 9).

Some nativists have even put forward the idea that concepts themselves might be innate (Fodor, 1998; Laurence & Margolis, 2001). For instance, Pinker's (1984) semantic bootstrapping hypothesis predicts that certain semantic referents are innately specified. In particular, children are said to have innately specified information in terms of nouns referring to objects, and verbs referring to actions. Were this true, then one would expect all human languages naturally map object names to grammatical nouns, actions to verbs, etc. The empirical validity of such an argument is, however, rather questionable. English, for example, has many verbs which do not directly map into neat semantic actions in the strict sense of the word (e.g. *think*, *seem*, *be*, *have*, *understand*...). I will come back to this issue later on.

Other researchers claim that children, as well as adults, learn the meanings of words through more general cognitive capacities. These include, among others, a rich system of conceptual representations, the sensitivity to infer the intentions of others, and a natural capacity to interpret scenes, properties and events in the ambient extralinguistic world (Gleitman & Gleitman, 2001). Once acquired, the meanings of words might be used to infer the grammatical category to which words belong.

How do children acquire the meanings or words? Many studies have highlighted a noun bias in children when learning their first language, mainly because of the conceptual simplicity that nouns exhibit (Bassano, 2000; Bates et al., 1994; Caselli, Casadio & Bates, 1999; Jackson-Maldonado et al., 1993). Concreteness or imageability might be the underlying predictor of the identifiability of nouns from their observed extralinguistic contexts. Thus, learners' noun bias might be based on an implicit realization that object-reference items are the best ones that fit in a word-to-world pairing procedure. With the meaning of nouns, and the intuition that nouns relate to real-world objects or referents, children might then start building a rudimentary nominal grammatical category.

Then, in subsequent stages of the language learning process, the learned nouns might form the scaffold from which the acquisition of more abstract vocabulary items might proceed. Thus, the knowledge of the acquired nouns can serve as bootstrapping for the learning of other unknown elements like verbs (Gleitman et al., 2005). Selectional information can play a great role too: certain kinds of nouns systematically tend to occur with certain kinds of verbs (e.g. nouns referring to food elements are likely to cooccur with verbs like *eat*; a noun like *telephone* can bootstrap the learning of verbs like *ring*, *talk*, *call*, etc.).

Then, the semantic bootstrapping proposal claims that one source of information about the meaning of words is there from the beginning of the language learning process and it constitutes the basis from which learners start building their initial lexicons. This initial information source is the ability to perceive and interpret the extralinguistic world. Then, it allows the learner to acquire a subset of lexical items (i.e. nouns that refer to specific objects) which requires little linguistic knowledge and is pragmatically supported. The knowledge of this subset is what later on triggers the acquisition of other abstract elements whose discovery requires support from previously learned linguistic representations (Gleitman & Gleitman, 2001; Gleitman et al., 2005). Thus, the grounding of grammatical categories in general starts from the identification of semantic categories first, and semantic features would later on bootstrap grammar.

The strength of the semantic bootstrapping approaches lies on the fact that mappings between semantics and grammatical classes are universal (e.g. nouns denote objects while verbs denote actions in any language). On the contrary, other types of cues like phonological or distributional cues are language-specific. However, such approaches presuppose that (a) such correlations between semantics and grammatical classes are perfect mappings in any language and (b) children start the language learning process with the expectation that such mappings actually exist. The following section examines these two problems in depth.

3.1.2.2 Good birds and bad birds

One of the problems of any semantic bootstrapping approach is that it presupposes an initial identification of specific objects and an immediate pairing of those objects to 'overt' nominal elements on the part of language learners. However, this necessary straightforward concept-word pairing is somehow problematic, as some have pointed out (Bloom, 2000).

To start with, naive learners with no prior knowledge of grammatical categories in their language might even fail to perform the object referent-word mapping. Imagine a situation where both mother and child are watching a brown dog which is running around. The mother would say something like "Look at the dog." Even if child learners were ready to map the external animal referent to a particular word in the input, how does the child know which word does the "object" semantic component refer to, out of the four possible words she has heard from her mother? The mapping task becomes less clear when facing multiple-word utterances. However, such multiple-word utterances are the ones which children are most likely to encounter in the course of their linguistic development, as only a small percentage of utterances in child-directed speech have been reported to contain words in isolation (Bernstein Ratner & Rooney, 2001). Not only that, but research also shows that some mothers never or hardly ever use words in isolation, even in situations where they are explicitly teaching new vocabulary to their children (Aslin et al., 1996).

A further problem concerns the way the object-word mapping itself should proceed. It is known as *Quine's Gavagai parable*, or the problem of referential indeterminacy (Quine, 1960). In the mother-child interaction example I mentioned before, even if the visual image and the target word *dog* were immediately associated, how does the child know that the word *dog* actually refers to the dog itself and not, say, to a more general referent such as *animal* or *mammal*, or to a specific type of dog, or to a part of the dog (i.e. its legs, its tail...) or to a physical property of the dog (i.e. its colour, its fur...) or to the action of running itself? How do children know that they can equally use the word *dog* to refer to another type of dog (i.e. a sitting dog, a white dog, a dog from a different breed, etc.) and they cannot use it with other animals like a brown running cat? A mere world-to-word mapping assumption cannot account for children's choice and learning of the word *dog* and its natural referents in the real world.

A possible solution to the *Gavagai problem* is given by Tomasello (2003) with his social-pragmatic theory of word learning. According to this theory, children's word learning process is constrained by: (1) children's structured social world, full of routines and patterned interactions, and (2) children's social-cognitive abilities for participating in such structured social world (i.e. joint attention and intention-reading). Thus, on the one hand, context and routine cultural activities structure the language learner's experience and act as constraints for word learning (Bruner, 1983; Nelson, 1985). On the other hand, in their attempts to understand adults' messages and intentions, children use a series of interpretive strategies based on the pragmatic assumption that adults' utterances are relevant to the ongoing situation (Bruner, 1983; Sperber & Wilson, 1986).

Thus, under general communicative situations and in the absence of a pragmatic context that would indicate otherwise, children (and also adult speakers in general) tend to associate first-mentioned nouns to what Rosch (1978) labelled as basic level object categories[4] (i.e. the middle level of the conceptual category hierarchy that Rosch called

[4] Basic level concepts are those which have the highest degree of semantic cue validity. Thus, a category like [animal] has no cognitive visual representation, since it is too general to refer to a particular entity. On the other hand, basic categories within the superordinate [animal], (i.e. [dog], [bird], [fish]) have the necessary informational content to be categorized in terms of reference and semantic features (cf. basic level concepts and Prototype Theory (Rosch, 1978)).

the Basic Level). This would lead children interpret the meaning of most nouns as referring to basic level object categories, and not to subordinate exemplars or to higher superordinate ones.

This assumption could then solve the above-mentioned problem: children would map the word *dog* to the basic level object category of *dog*. As a result, they would use the same word to refer to other types of dogs, but not other types of animals. However, this would only solve the problem in those situations where the words children are exposed to actually belong to basic level objects. What if children were exposed to superordinate terms? Or what if they were exposed to abstract words whose meaning does not relate to a specific object or referent?

Traditional linguistic analyses postulate that the category *noun* is both a notional and a grammatical concept (Lyons, 1977). There is a central semantic concept of noun, which is present in all languages, and which includes words for persons, animals and things. All the other more abstract ontological categories denoted by nouns appear to be generalizations from this core concept.

Acknowledging that this core concept includes only a subset of all possible nouns in English gives rise to the question of how large the proportion of nouns belonging to this subset actually is. In other words, can this subset account for all the examples of nouns that children are exposed to and that they would later acquire? Previous studies have already pointed out that semantic criteria alone do not provide a reliable basis to determine the category membership of many words in English, since there are many nouns which do not denote physical objects (e.g. *an explanation*), or there are many verbs that do not refer to actions (e.g. *to seem*), etc. (Benedict, 1979; Maratsos & Chalkley, 1980; Maratsos, 1999).

This has lead some researchers (Nelson et al. 1993) to propose two distinguishable semantic classes of nouns: on the one hand, BLOCS neatly correspond to Rosch's (1978) basic level object categories; on the other hand, XBLOCS refer to all those words which would naturally fall out of the cognitive basic level, either because their extralinguistic referent is too general or because they do not refer to a specific object referent at all. Nelson et al. (1993: 71) further distinguished different types of XBLOCS taking their meaning into account:

- *Locations*: places, both indoors and out (e.g. *beach*, *kitchen*)
- *Actions*: single actions, specific or general (e.g. *kiss*, *help*)
- *Superordinate/generic*: terms that denote what are generally considered superordinate categories (e.g. *toys*, *animals*).
- *Events*: terms that refer to complex events that take place through time (e.g. *lunch*, *party*).
- *Person roles*: roles that people play in social/cultural life (e.g. *doctor*, *brother*).
- *Natural phenomena*: states, actions, events or entities (e.g. *sky*, *snow*, *clouds*).
- *Temporal entities*: e.g. *morning*, *day*.
- *Parts of objects*: e.g. *button*.
- *Quantities*: e.g. *drop*.
- *Material*: e.g. *wood*.

In their study, Nelson et al. (1993) showed that children learn and use many XBLOC words early in the language learning process, and as multi-word combinations and noun morphology were developed, noun roles were extended to these words, showing that they had been accurately categorized within the nominal group. Then, words which lack the expected semantic content that would yield their successful categorization (and which would therefore be left unclassified on the basis of semantics alone) are nonetheless accurately classified in their right word class from the early stages of the nominal category building process. Even more problematic is the case of action words (i.e. XBLOC nouns like *kiss* or *help*, which denote actions and which should therefore be assigned to the verb category). However, these are used with noun distribution and morphology, just as all other nouns.

Furthermore, in line with the findings from Nelson et al. (1993), other studies also point out that object nouns or action verbs do not necessarily dominate children's earliest lexical productions, neither in English, nor in other languages (Bassano, 2000; Gopnik & Choi, 1995). And still, those words which do not neatly map into their corresponding semantic category are correctly classified by children and they are used in a grammatical way. Thus, non-action verbs are equally found with correct verbal morphology and non-object nouns are also found in their target syntactic context.

Given the evidence provided, it seems clear that there are other types of cues at work, alongside semantics, when words are being classified into their corresponding grammatical category. Thus, just as word segmentation tasks were carried out with the help of environmental phonological and distributional information, the same might be true of word categorization tasks. Furthermore, several researchers in the language acquisition literature have often observed that the formal and syntactic properties of words might help children discover the semantic information that those words encode, rather than having semantic information triggering the discovery of formal syntactic structure (Bowerman & Choi, 2001; Slobin, 2001). What follows is a review of the role that the syntactic distributional context of words might play in the process of grammatical categorization.

3.1.3. Distributional cues

3.1.3.1. The reliability of distributional contexts

Distributional information is a very powerful cue that assists young language learners when they face the task of segmenting the continuous speech stream into word-like units. In word segmentation, 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 the same way.

Thus, when turning to the task of categorizing the obtained 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 like *tty* 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.

Thus, 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, 1994; 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 (Braine et al., 1990; Brown, 1957; Dockrell & McShane, 1990; Gerken, 1996; Gerken, Wilson & Lewis, 2005; Mintz, 2002; Mintz et al. 2002; Saffran et al., 1996; Valian & Coulson, 1988; Waxman & Booth, 2001). Connectionist models based on distributional learning mechanisms have also proved to be quite successful at performing linguistic tasks (Elman et al., 1996; Rumelhart & McClelland, 1987; Plunkett, 1995).

Within nativist frameworks, researchers claim that unconstrained distributional analyses of a language cannot lead to the correct grammatical rules of that language. One of the potential problems that a learner who relies on distributional information is likely to encounter is that of nonimmediate adjacency. As noted by some authors (e.g. Chomsky, 1965; Pinker, 1987), distributional regularities in English are not always local, but can occur over a variable distance, as shown in the examples below:

(11) 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 the example in (11b) above is a noun, and not an adjective. How does the learner know which environments are important (as the one in (11a), which correctly predicts that *house* is a noun) and which should be ignored (as the one in (11b))?

This is why, according to Pinker (1997), the statistical learning procedure for word segmentation pointed out by Saffran et al. (1996a; 1996b) is not equally applicable to higher order linguistic tasks like word categorization. For Pinker, words and grammar are different and, therefore, learning words and learning grammar are two different computational problems. The sequence of sounds making up a word is not capturable by rules but must be memorized. And because there are a finite number of words, they all can be recorded. However, the sequence of words making up a sentence is capturable by rules, and word sequences cannot be memorized, because they form an open-ended set.

Another potential difficulty of distributional analyses of the input comes from the fact that some words in English can be both nouns and verbs, and their distributional contexts will vary accordingly. As suggested by Pinker (1987), those words will also give rise to incorrect inferences on the part of the language learner who relies on distributional information as the only source of information for grammatical categorization. Consider the following examples:

(12) a. *Laurie wants some help.*

b. *Laurie wants some chocolate.*

c. *Laurie could help.*

d. **Laurie could chocolate.*

Language learners who rely on distributional information would incorrectly categorize *help* and *chocolate* together and, after being exposed to examples as the one in (12c), they would incorrectly assume that (12d) is also possible. Thus, English words which belong to more than one category will also occur in different syntactic contexts, providing misleading information to language learners. Pinker (1987) argued that the resulting wrong generalizations would be common, and would make it impossible to accurately categorize words on the basis of distributional information alone.

However, despite these potential problems, empirical studies that analyze actual child-directed speech have shown that in children's input, neither dual-class words nor non-local adjacencies undermine the informativeness of distributional patterns (Mintz et al., 2002; Redington, Chater & Finch, 1998). Although problematic environments may exist, there is nevertheless enough evidence for accurate categorization in child-directed speech, compared to the noise created by the problematic environments. Thus, although such problematic environments do, in fact, exist, it seems that they are just not frequent enough to rule out the possibility of categorizing words from distributional contexts completely.

3.1.3.2. The nature of distributional contexts

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).

Bigrams are defined as pairs of elements where the categorizing element would either precede or follow the word to be categorized. In other words, suppose that element Y is to be categorized. The pairs XY or YZ would be considered bigrams, and X and Z would be the elements that make up the context that would yield the successful categorization of Y. 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 *com*_**ing**_ (where the verbal morphology that follows the target word would be the element that would yield 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). Any words that occur as the intervening word inside a given frequent frame are categorized together" (Mintz 2003: 93). In other words, the sequence XYZ could be considered a frame, where Y would be the intervening element to be categorized, and X and Z would be the elements that make up the frame and that would yield the accurate categorization of Y. For example, the sequence *they come here* could be considered a frame, where both the strong subject pronoun as well as the following adverb would be strong indicators that the intervening word in the middle is a verb.

Mintz (2003) outlines the advantage of frames over bigrams for categorization matters. Since frames require the joint occurrence of more than one contextual element, this eliminates many accidental contexts from the analysis and slightly rules out the problem of nonimmediate adjacencies captured by Pinker (1987). Thus, in his study, Mintz (2003) found out that frequent frames were extremely effective at categorizing words, as shown by the very high accuracy scores obtained in the analysis.

In general terms, then, categorizing words in child directed speech on the basis of their distributional context produces extremely accurate categories. Thus, the efficiency of distributional contexts, and the relative simplicity and accessibility of the information they provide for categorization make them very good candidates for young language learners with very little linguistic experience as well as limited memory and processing resources. Thus, distributional information might be a viable ground from which children undertake word categorization tasks (Maratsos, 1998).

However, one of the problems that the data from Mintz (2003) reveal is that several noun categories emerge, instead of a single adult-like one. Despite the high levels of accuracy obtained, the study reveals very low levels of completeness for every frequent frame, which means that many different categories would potentially emerge out of such a distributional analysis, albeit all of them relatively large and very accurate.

How could the problem of multiple categories be solved? Mintz (2003) suggested that categories obtained from a distributional analysis of the input could be unified if there were a considerable degree of lexical overlap from category to category. Then, a simple conglomeration strategy based on lexical overlap across different smaller categories could be used to join them together into a more complete unifying category. More research is needed in order to test the likelihood with which multiple categories could be successfully joined together in a broader category without risking the high levels of accuracy obtained by smaller categories.

It seems clear, then, that a distributional analysis of the input does not completely overrule the need to consider semantic information of the type outlined in the previous section. In this way, distributional categories that contain nouns could be identified and joined together based on the semantic similarity across its constituting elements. Once the noun category is identified, it can be used as the bootstrap to identify other categories, such as verbs. Thus, initially words might be clustered together by means of a distributional analysis. Then, nouns would be identified as such based on semantic correspondences. The location of nouns in sentences could then be used to guide the identification of all other categories.

Furthermore, distributional information together with prosodic information could be used to induce higher order syntactic relationships such as phrasal constituency and hierarchical structure. The following section revises some of the studies that have considered the power of several different cues in combination in word categorization tasks.

3.2. The combination of cues in the linguistic environment

Before proposing a presumably impoverished environment for children in language learning tasks, one should consider the variety of sources of information available in the speech signal which might be useful for the completion of such tasks. Several different kinds of cues might converge to the solution of a given problem. Furthermore, each cue does not need to have 100% validity, since the learning of some initial cues might trigger the discovery of more complex ones. Different cue types might be taken individually and learned independently (Rispoli, 1999), or they might reinforce each other (Onnis et al., 2005; Reali, Christiansen & Monaghan, 2003).

As for categorization, multiple cues which are available in the input can contribute to the development of accurate and useful grammatical categories. As seen above, semantic cues as well as distributional and phonological cues might reveal the grammatical category to which words belong with a certain degree of accuracy on an individual basis. More interestingly, if taken together, combined cues might still yield a greater accuracy in grammatical category assignments.

Very few studies have analyzed the interaction between several different cues and the contribution of each cue in the grammatical category assignments of words. Besides, all of them have taken two types of cues and have analyzed their relationship and the way they interact, but no study has considered the interaction between the three different types of cues available in the child's linguistic enviroment.

For example, in an unpublished manuscript Kelly & Martin (1995), quoted in Kelly (1996), examined the interaction between semantic information and the stress patterns of words. In their experiment, they found that the effects of stress patterns were not reduced or eliminated if a word had a meaning which was characteristic of its grammatical category (i.e. if nouns paired with objects and verbs with actions). In fact, the presence of the semantic cue strengthened the categorization force of the phonological cue. Thus, participants in the experiment recognized words with iambic stress patterns as verbs faster than words with trochaic stress. Furthermore, their reactions were even faster if those verbs had a prototypical "action" semantic component. The same was true of the nominal target words in the experiment. These results indicate that word categorization is faster when multiple cues converge to indicate the grammatical category to which words belong.

Monaghan, Chater & Christiansen (2005) as well as Monaghan, Christiansen & Chater (2007) have analyzed the correlation between phonological and distributional cues in the input. The analyses show that, when combined, phonological and distributional cues interact in such a way that they provide useful, and perhaps sufficient, information for the development of grammatical categories. They propose the so-called *Phonological-Distributional Coherence Hypothesis*. Its main claim is that, when distributional information is present for the categorization of a certain word, the phonological cues of that word will be less crucial and, therefore, some shifting of these cues can occur. However, when distributional information is weaker for any given word, then the phonological cues to the accurate categorization of that word will be stronger. Thus, they predict a perfect coherence between both types of cues: when one source of information is weaker at assigning category membership, then the other type of cues will be stronger. In other words, elements that are ineffectively categorized using distributional cues will be effectively categorized using phonological cues and viceversa (Monaghan, Christiansen & Chater, 2007).

In an analysis of English corpora, Monaghan, Chater & Christiansen (2005) found support for the *Phonological-Distributional Coherence Hypothesis*: both phonological cues and distributional cues were found to be accurate to determine the grammatical category membership of words. Furthermore, accuracy in categorization increased when both types of cues were combined[5]. Not only that, but they also showed that the improvement when considering cues in combination was due to the fact that words which were miscategorized by one type of cue tended to be accurately categorized by the other type of cue. Thus, phonological and distributional cues provided additive value and contributed to the categorization of words in a different way. In particular, distributional information was found to be especially powerful in the categorization of high frequency words, while phonological cues were more coherent in low frequency words.

> This provides support for the (...) claim that cues are in a complementary distribution to aid classification. For words where distributional cues are not so rich then phonological cues provide information about category, similarly for words where the phonological cues are indistinct then distributional information supports accurate classification. (Monaghan, Christiansen & Chater 2007: 291)

[5] The set of cues they used to test the accuracy of categorization of individual cues was different from the set of cues used to test the accuracy of categorization of multiple cues. See Monaghan, Chater & Christiansen (2005) for details.

Furthermore, Monaghan, Christiansen & Chater (2007) provide evidence that the *Phonological-Distributional Coherence Hypothesis* might also be true for other languages, besides English. In their analysis of English, Dutch, French and Japanese corpora, the same interaction between phonological and distributional cues was found, suggesting that the role of the different cue types might be general across very different languages.

In conclusion, these studies demonstrate the wealth of the language learner's linguistic environment, as well as the benefit of integration of information from multiple sources. To date, no study has analyzed the correlation between all the sources of information available for grammatical categorization (i.e. semantic, distributional and phonological cues). In the following chapter, I will present my own study, using all three types of sources and the accuracy with which they categorize English nouns.

Chapter 4:
Research questions

4.1. General objectives

One of the main arguments for linguistic nativism is the so-called *Poverty of the Stimulus Argument.* As stated in previous chapters, the main claim of this argument is that the language samples to which children are exposed do not contain information which is regular and consistent enough for children to acquire the grammar of their language on the basis of experience alone (Chomsky, 1975). As far as categorization is concerned, the *Poverty of the Stimulus Argument* would predict that children's linguistic input does not contain the regular, consistent and salient information necessary for children to correctly and reliably map individual lexical items into their corresponding grammatical categories.

The aim of this dissertation is to challenge such a view and to test the empirical validity of the *Poverty of the Stimulus Argument* regarding grammatical categorization by analyzing samples of English child-directed speech and examining the regularities of the language used by mothers when addressing their children. The initial prediction is that the linguistic environment to which children are exposed does, in fact, contain a series of systematic, coherent and consistent regularities that would allow children to infer an abstract noun category without the need to postulate any *a priori* innate knowledge about the nature of such category.

4.2. Research question 1: individual cues

What kind of sources 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 sources? To which extent could children work out the category for nouns in English on the basis of a single source alone?

In this dissertation, I will examine the kind and the amount of linguistic cues which are present in the input addressed to English-learning children and which would allow the successful categorization of nouns in English. By considering a series of cues that are common to all and only the lexical items that make up the nominal category in English, I will analyze the extent to which such cues reliably map English nouns into their grammatical category as well as the extent to which those cues are sufficient and necessary for the grouping of all English nouns into the same category.

4.2.1. Syntactic cues

Several authors have already pointed out the usefulness of syntactic or distributional information for grammatical categorization (e.g. Cartwright & Brent, 1997; Mintz et al., 2002; Mintz, 2003; Redington, Chater & Finch, 1998). Distributional regularities in the input appear to be consistent and reliable enough so as to account for the location and function of most of the English nouns to which children are exposed.

Most of the studies that have been carried out to date are based on English data. Due to the limited number of inflectional morphemes in English, 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 particular frames or contexts are relevant and useful for word categorization in her own environment and which are not.

In the present study, distributional contexts for noun categorization will only be defined as including determiners 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. Furthermore, such functional elements are absent from children's early linguistic productions (Radford, 1995; 1996). However, as seen in previous chapters, children are well aware of the presence of absence of functional morphemes and words long before they start producing them (e.g. Gerken, 2001; Gerken, Landau & Remez, 1990; Gerken & McIntosh, 1993; Golinkoff, Hirsch-Pasek & Schweisguth, 2001; Jusczyk, 2001; Shady, 1996; Shipley, Smith & Gleitman, 1969). 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 and. 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.

However, as a result of the smaller number of distributional contexts, lower scores of accurate categorization are expected as well. On the one hand, not all resulting contexts will take the form of frames (i.e. two categorizing elements plus one intervening element in the middle), but also bigrams (i.e. one categorizing element plus one intervening element to the right). A comparison will be drawn between the scores obtained from bigrams alone and those obtained from frames alone, in order to evaluate the risks of losing categorizing force by just considering plural morphology as the only categorizing element on the right of frames.

In fact, even considering wider definitions of frames as contexts including several kinds of words (Mintz, 2003), very low completeness scores are also obtained in previous studies when one considers individual frames. That is, a single frame only categorizes a small number of nominal elements from the whole corpus, which results in more than one frame-based category for the single grammatical category of nouns (i.e. the child might not be creating just one but several nominal categories on the basis of frames alone). Similar results are expected here in the sense that every frame or bigram will only account for the categorization of a small proportion of the total number of nouns to which every child is exposed. However, 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 then that English-speaking children can generalize across different contexts and will ultimately be able to come up with a single unifying and more abstract grammatical category.

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; Mintz et al., 2002; Pinker, 1984). Thus, some of the frames in previous analyses have been found to group together words which actually belong to different grammatical categories. In fact, the *Poverty of the Stimulus Argument* establishes that the linguistic environment to which children are exposed is both too limited (i.e. it does not give enough information about what is grammatically possible in every language), and too general (i.e. the information it gives allows overgeneralizations as it does not contain specifications on what is not grammatically possible in every language either).

As the present analysis only considers extremely local determiner + noun relationships, a further analyis will be conducted 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. My 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 prevent children from drawing the wrong conclusions about the grammatical categorization of certain non-nominal elements, despite their accidental appearance within nominal distributional contexts.

4.2.2. Phonological cues

A number of studies have also pointed out the importance of phonology and sound cues for the grammatical categorization of nouns and verbs (Cassidy & Kelly, 1991; Durieux & Gillis, 2001; Kelly & Bock, 1988; Kelly, 1992; 1996; Monaghan, Chater & Christiansen, 2005; Soreno & Jongman, 1990). The phonological form of words in English might be useful to distinguish open from closed class words on the one hand, and nouns and verbs on the other.

In this dissertation, I will try to replicate some of these previous analyses as far as English nouns are concerned. More specifically, I will work out: (a) the degree of successful categorization (i.e. the percentage of nouns which are correctly categorized as nouns); (b) the degree of non-categorization (i.e. the percentage of nouns which lack the cues under consideration and which would, therefore, fall out of the noun category);

(c) the degree of miscategorization or overcategorization (i.e. the percentage of lexical items other than nouns which, bearing any of the cues under consideration, would be incorrectly classified as nouns on the basis of their phonological form). As with syntax, the prediction for the test with phonology is that a given set of sound cues alone will be enough to account for most of the nouns to which English learning children are exposed. Taken individually, phonological cues might not reach high completeness scores (although accuracy scores are expected to be high), but when taking them together, completeness scores might improve while noun categorization will still be accurate after all cues have been added.

4.2.3. Semantic cues

A third source of information that has proved to be helpful for the grammatical categorization of words is semantic information (Pinker, 1987). While it is still not clear whether the first analysis that children perform on the input is on notional grounds (and therefore, semantic cues are considered first) or distributional grounds (and thus, syntactic cues are considered first), it is widely accepted that the semantic notion of *object* and *action* might assist language learners in the identification of nouns and verbs respectively. One of the problems for *the Semantic Bootstrapping* proposal lies on the difficulty of identifying the meaning of unknown words and, consequently, their semantic category. Besides, the links between semantic categories and grammatical categories are not one-to-one but many-to-many (Mintz, 2003). Thus, for example, not all items within the semantic category of *actions* are verbs. An adjective such as *noisy*, or a noun such as *call*, can also be semantically classified as actions. In fact, as Nelson et al. (1993) have pointed out, the actual proportion of English nouns which conform to the semantic category of objects is only a small subset of the whole noun inventory.

Then, an additional analysis will be carried out in order to examine the usefulness and reliability of semantic cues for the categorization of nouns in English child-directed speech. While it is true that not all English nouns refer to objects exclusively, it is also true that the kind of interactions that very young children are involved in are often restricted to the *here* and *now* and to familiar objects within each child's reach (Baldwin, 1991; 1993; Clark, 2009). The prediction is that most nouns to which children are exposed conform to the pattern of being either object referents or substance referents or else they are proper nouns with a unique single referent that will be relatively easy to identify as well. Consequently, the amount of nouns which will be semantically too vague or general or too ambiguous as not having a direct referent will not be as important as to narrow down the usefulness of semantic information for grammatical categorization. A close evaluation of words which semantically refer to actions will be carried out in order to examine the possible overlap between the grammatical categories of nouns and verbs which share a common semantic content. Although it exists, such overlap is expected to be minimal and not in conflict with an overall sucessful categorization on the basis of semantic information.

4.2.4. Prediction about research question 1

In line with the results found in similar studies, and based on the evidence discussed in the previous sections, the following prediction can be established for Research Question 1:

Prediction 1:

All three sources of information for grammatical categorization will be useful, consistent and reliable by themselves, although not to the same extent. While most nouns are expected to be sucessfully categorized on the basis of distributional information alone, or on the basis of semantic information only, phonological cues will be weaker and will show lower completeness scores.

4.3. Research question 2: cues in interaction

In which way do distributional, phonological and semantic cues interact? Do they categorize the same elements and only a few (i.e. are they redundant)? Does the sum of all three categorize all English nouns (i.e. are they complementary)? Or do they ever enter into conflict?

While several studies have examined the effectiveness of a certain kind of cue for grammatical categorization, most of them have explored individual cues in isolation and only a few have considered the interaction of several different kinds of cues. In fact, only Monaghan, Chater & Christiansen (2005) or Monaghan, Christiansen & Chater (2007) have considered the interaction between phonological and distributional cues. They have found out that syntax and phonology are not redundant, since both sources do not exclusively categorize certain nouns and the same elements in each case, but they are not contradictory either, since what distributional information would not categorize as a noun is often categorized as such in terms of the word's own phonological form. In fact, Monaghan, Christiansen & Chater (2007) have put forward the *Phonological-Distributional Coherence Hypothesis*, according to which phonological and distributional cues operate in tandem as far as the grammatical categorization of elements is concerned (see chapter 3).

However, those studies only consider the most frequent 5,000 words in the whole CHILDES database, regardless their category or the age of the children to whom such words were addressed. A closer examination to the whole inventory of nouns to which very young English children are exposed is proposed here.

Besides, I will explore the interaction of all three possible sources of information for grammatical categorization. I will test whether all combined cues are reliable and consistent with one another or whether they ever enter into conflict (e.g. the likelihood of finding nouns in the input which are framed within a syntactic nominal context but are prosodically marked as verbs). On the basis of such analyses, I will work out the overall degree of accuracy with which nouns in English are likely to be correctly categorized on the basis of all possible sources of information available in the input. Then, for the analysis of cues in interaction and Research Question 2, the following prediction can be established:

> *Prediction 2:*
>
> The interaction of syntactic, phonological and semantic cues will be reliable and useful for children, since such cues will interact with one another and provide the successful categorization of most English nouns. Furthermore, the three different kinds of sources will work differently aas far as categorization is concerned and will complement one another. Thus, the elements for which phonological cues will be irrelevant can be correctly categorized on the basis of distributional and/or semantic information and viceversa.

Chapter 5: Methodology

5.1. Data source

The data used for the present analysis comes from the Manchester corpus (Theakston et al., 2001), available from the CHILDES database (MacWhinney, 2000). 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.

Within the individual corpus of every child, I selected and considered those transcriptions in which the child was approximately 2;6 years old and younger, and I discarded all those transcriptions in which the children were 2;7 years old or older. This was done in order to keep a similar sample as that from other studies with parallel goals (e.g. Mintz, 2003; Mintz et al., 2002), and in contrast to other more recent studies where all the speech by adults from all the English corpora in the database were extracted, regardless the age of the child to whom the utterances were addressed (i.e. Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007). 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 et al., 1997; 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, it is then necessary 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.

Among the four children, there were two boys (i.e. Aran and Carl) and two girls (i.e. Anne and Becky). All of them were from either Manchester or Nottingham and they were all being brought up by middle-class families. All children under consideration were first borns, monolingual English speakers (and they also spoke the same dialectal variety, namely British English) and were cared for primarily by their mothers.

All children were audiotaped in their homes for an hour on two separate sessions every three weeks. They engaged in normal play activities with their mothers. All speech was transcribed with the exception of speech not explicitly directed to the child (i.e. conversations between adults, telephone calls etc.). However, if the child produced an utterance in response to such speech, the relevant utterances were transcribed as well (Theakston et al., 2001).

For the present study, only adult language was taken into account, and utterances spoken by the child were not analyzed. 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. These adult utterances included, but were not limited to, the production of the child's mother and the investigator in charge of the recording of the sessions. In cases were friends or fathers also interacted with their children, those utterances were also considered.

Table 5.1 shows some of the characteristics of the children under analysis. While the sex of the child was not a variable taken into account in this study, special care was taken of the fact that, for all the transcriptions, none of the children was older than 2;7.

It is also important to note that children also come from the same dialectal and social environments, which provides linguistic homogeneity to the whole corpus subsample[6].

Table 5.1: A summary of the general characteristics of the children in the corpus subsample.

	Anne	**Aran**	**Becky**	**Carl**
Age	1;10.07 to 2;6.29	1;11.12 to 2;6.17	2;0.07 to 2;6.29	1;8.22 to 2;6.19
Sex	female	male	female	male
Origin	Nottingham	Manchester	Nottingham	Manchester
Social class	middle class	middle class	middle class	middle class
Participants	mother, investigator	mother, investigator	mother, investigator	mother, investigator, father, family friend
Corpus size: types	2,761	3,432	2,222	2,266
Corpus size: tokens[7]	103,457	118,469	65,411	76,859

5.2. 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

[6] Several studies have outlined a relationship between variation in L1 acquisition and the particular characteristics of the input that children hear from those around them. See Lieven (2006) for a review.

[7] The actual corpus size of every child was bigger, as children were recorded for longer than 2;6 years. These numbers correspond to the size of every corpus subsample chosen for the present study (i.e. those transcriptions that fall between the age range under consideration).

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, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007), or on the accuracy with which nouns and verbs can be grouped into different grammatical categories (e.g. Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007; 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 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 excluded from the analysis. Prepositions, interjections, onomatopoeic expressions and greetings or set phrases such as *goodbye* or *thank you* were also taken out.

Words were then classified into two different categories to be analyzed separately. One category included all nouns, and the other category, which was labelled "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 used as nouns and the number of tokens that were used as verbs in every transcript.

Individual words were then analyzed to see the degree to which they matched a selected set of cues which were said to identify the grammatical category of English nouns. If the group of selected nouns obtained a high score in the analysis, that would indicate that the selected set of cues accurately classified English nouns in their grammatical category. On the contrary, if low scores were obtained, that would indicate that the set of cues under consideration was bad at classifying nouns accurately.

The same procedure was carried out with the list of verbs, adjectives and adverbs obtained, but with the opposite expectation. Thus, a low score in the analysis of verbs, adjectives and adverbs would indicate that the set of cues considered were deficient for the classification of these words (as they should be, since the cues were selected according to the characterization of English nouns only). On the contrary, a high score with verbs, adjectives and adverbs would indicate that the same set of cues that classified nouns in English was also useful to classify other words other than nouns. Such contradictory and inconsistent information would be useless to the language learning infant who does not have any *a priori* innate linguistic knowledge of grammatical categories. All in all, in order for the cues to be valid for accurate categorization, and therefore, useful in the language learning process, one should expect a high score for the group of nouns and a low score for the group of verbs, adjectives and adverbs in the analysis.

5.3. Frequency groups

In order to explore the interaction between word frequency and accurate grammatical classification, the group of 9,641 word types obtained from the four corpora (corresponding to a total of 143,272 tokens) was further subdivided into three different groups, with different word frequency ranges each. The main criterion with which the frequency groupings were made was that each of the three groups that resulted from the split should be more or less equivalent in size. This would guarantee that correct classification of tokens was even across similar subgroups, and not just over a very small subsample of tokens. Special care was also taken that the noun *versus* other proportions arisen in each of the three groups was similar, not only from group to group, but also from the whole corpus. Thus, depending on their token frequency values, words were included in one of the following groups:

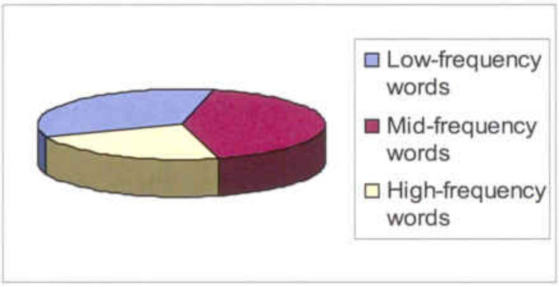
- *Low-frequency words*: this group included all words from the four corpora which had a token frequency of 1. A total of 1,883 nouns and 1,399 other open class words matched this description.

- *Mid-frequency words*: this group included all words from the four corpora whose token frequency values ranged from 2 to 9. A total of 2,337 nouns and 1,729 other open class words matched this description.

- *High-frequency words*: this group included all words from the four corpora with a token frequeny equivalent to 10 or higher. A total of 1,168 nouns and 1,105 other open class words matched this description.

Figure 5.1 shows the size of each frequency group as far as the whole corpus is concerned. As can be seen, the three groups keep more or less an even size, which makes them equivalent for direct comparisons and analyses.

Figure 5.1.: Distribution of the total number of cases in each of the frequency groups.



5.4. Cue derivation

For the present analysis, three different types of cues were taken into account: syntactic (or distributional), phonological and semantic cues. Each cue was labelled as a *Syn*, *Phon* or *Sem* cue, in order to indicate the type of information they contained (syntactic, phonological or semantic, respectively).

5.4.1. Distributional cues

As far as distributional cues are concerned, a total of 15 different cues were considered, each corresponding to a different distributional 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*)[8]. Following previous studies (Mintz, 2003; Mintz et al. 2002; Monaghan, Chater & Christiansen, 2005), the set of English articles, demonstrative determiners, possessive determiners,

[8] Anglo Saxon genitive *'s* was discarded from the analysis as it graphically overlapped with the way the verb *to be* was transcribed in all children's files. Thus, an element such as *boy's* obtained from the list generated by FREQ could either correspond to *the boy's teddy* (where it would be part of a nominal syntactic context) or *the boy's here, the boy's playing* (and thus be part of a verb phrase which followed a nominal phrase).

quantifiers, prepositions and *wh-* interrogative elements were considered as intervening items in the "*x* + *noun*" bigrams or the "*x* + *noun* + *-(e)s*" frames. Unlike Mintz (2003), this study only considered determiners as the first element within a bigram or frame, and plural noun morphology as the second categorizing element, and not just any independent word which would frequently correlate with the occurrence of 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. That would give distributional contexts greater consistency. Although weaker successful categorization scores are expected due to the lower number of frames and bigrams, the set of phonological and semantic cues would counterbalance the weaker categorization obtained on distributional grounds only.

For the present analysis, only extremely local distributional contexts were considered, unlike other previous studies (cf. Mintz et al., 2002; Redington, Chater & Finch, 1998). 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*. On the one hand, such an approach would make distributional information look qualitatively similar to phonological information to the language learning child. On the other hand, it will also 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 15 different distributional contexts was generated and every target word was analyzed to see whether its context matched any of the 15 established. Words scored 1 if they appeared in a syntactic context that matched the description of a given cue and they scored 0 if their context did not match. The set of distributional contexts considered include the following:

- *Syn0*: this grouped together words which were not preceded by any element which prototypically introduces nouns in English (e.g. articles, possessive determiners, etc.). Thus, this category included nouns which occurred with no distributional context (or at least not a context that would ease categorization) as well as verbs, adjectives and adverbs. Words in the corpora whose distributional context matched this description scored 1 here and 0 in all the other contexts. The *Syn0* group was further divided into two subgroups:

    - *Syn0a*: this grouped together all words which were not part of any set bigram and also did not have any nominal inflectional marker at the end, and thus did not occur within any nominal distributional context (e.g. bare singular nouns, or singular nouns preceded by a verb or an adjective; proper nouns also fell into this category).
    - *Syn0b*: this grouped together words which occurred in the same distributional contexts as those words in *Syn0a*, but which had a

nominal plural morpheme *–s* at or *–es* at the end (e.g. *books, houses*).[9]

- *Syn1*: under this label, singular words preceded by an indefinite article *a/an* were classified. Words within such a bigram scored 1 here and 0 elsewhere (e.g. *a dog, an apple*). Since it is local syntactic contexts that are being considered, an adjective such as *beautiful* in *a beautiful girl* would also score 1 here.

- *Syn2*: this label was meant to group together words which were introduced by a definite article *the*. Words found in such syntactic context scored 1 in the *Syn2* category and 0 elsewhere. The *Syn2* category was further subdivided into two groups:

    - *Syn2a*: this grouped together singular nouns which were preceded by a definite article (e.g. *the man*), as well as any other word found in the same context (e.g. *the clever man*). Irregular plural nouns which were introduced by a definite article (e.g. *the children*) also scored 1 here, since they would be ideally categorized on a bigram basis, and not as intervening elements within a frame.

[9] Note that, while being labelled the same way, the *Syn0a* and the *Syn0b* cues describe two different degrees of categorization: *Syn0a* would group words with no nominal distributional context (and it would therefore describe the degree of non-categorization of words), whereas *Syn0b* would describe elements not preceded by a determiner but followed by nominal inflectional morphology (and it would therefore describe the degree of categorization of words on a bigram basis). The labelling of both types of cues as *Syn0* is motivated by the fact that both contexts share the property of describing phrases not introduced by any kind of determiner.

  - *Syn2b*: this grouped together plural nouns which were preceded by a definite article and which contained the plural morphemes *–s* or *–es* (e.g. *the buses, the toys*).

- *Syn3*: under this label, words preceded by a demonstrative determiner were classified. Words within such a distributional context scored 1 here and 0 elsewhere. The *Syn3* category was further divided into two different subgroups:

  - *Syn3a*: this was the subgroup that included all words which were preceded by the demonstratives *this, that*, *these* or *those* and had no nominal morphology at the end, either because they were singular nouns, because they were irregular plural nouns, or because they were not nouns (e.g. *this doll, those men, that brown dog, these tiny birds*). They all had in common the fact that they were preceded by the same kind of element and were not followed by any morphological marker, which would make them classifiable on a bigram basis.
  - *Syn3b*: this grouped together words which were preceded by the demonstratives *these* or *those* and which contained the plural morphemes *–s* or *–es* (e.g. *these girls, those houses*). They all had in common the fact that they were preceded by similar elements and followed by similar morphological markers, which would make them classifiable on a frame basis.

- *Syn4*: this category included all words in the corpora which were preceded by any form of possessive determiner (i.e. *my*, *your*, *his*, *our*, *their*, etc.). Words within such distributional contexts scored 1 in the *Syn4* category and 0 in all other categories. Similar to most *Syn* categories, the *Syn4* category was further split into two morphologically different subgroups:

    - *Syn4a*: this subgroup included all words which were preceded by a possessive determiner and had no nominal inflectional morphology at the end (e.g. *my book, their mice*).
    - *Syn4b*: this was the subgroup for words with the plural morphemes *–s* or *–es* at the end, and which were equally introduced by a possessive adjective (e.g. *her books, your kisses*).

- *Syn5*: this category was meant to group items which were introduced by any English quantifier. Items found within such contexts scored 1 in this category and 0 everywhere else. Again, the *Syn5* group was further subdivided into two:

    - *Syn5a*: for singular countable nouns, mass nouns or irregular plural countable nouns preceded by a quantifier, or any other word which might accidentally fall within the same syntactic context (e.g. *much milk, many people, some good friends*).

  - *Syn5b*: for plural countable nouns which were introduced by a quantifier and had the plural morphemes *–s* or *–es* at the end (e.g. *many stamps, some sausages*).

- *Syn6*: this category was made to group together all words which were introduced by a preposition. Words in such distributional contexts scored 1 here and 0 in all other categories. Similar to other *Syn* categories, the *Syn6* category included two different subgroups:

  - *Syn6a*: this was the group for words with no nominal morphological marker and which were preceded by a preposition (e.g. *on time*).
  - *Syn6b*: this was the group for words with plural morphology and which were preceded by a preposition (e.g. *at weekends, with glasses*).

- *Syn7*: this category grouped together all words which were introduced by a *wh-* interrogative word that could syntactically introduce nouns in English (e.g. *which*, *what*, *whose*, etc.)[10]. As the other *Syn* categories, the *Syn7* group included two morphologically different subgroups:

  - *Syn7a*: for words with no nominal morphology and which were preceded by a *wh-* word (e.g. *which story*).

[10] *Wh-* words such as *who*, *when* or *where*, which are typically followed by an auxiliary verb and not a noun, were excluded and not considered here.

- *Syn7b*: for words which had a nominal morphological marker at the end and were preceded by a *wh-* word (e.g. *whose books*).

Table 5.2 shows a summary of the different elements that were taken as introducing elements of the bigrams and frames under analysis. Appendix A (tables 1-4) contain the tables with the particular words that were taken as framing elements in the corpus of every different child.

Some of the framing elements were common across the four corpora (e.g. the elements in the *Syn2* group). Other elements, on the other hand, were specific of a particular corpus, as they were elements which were only characteristic in the linguistic environment of a given child (e.g. some individual instantiations of quantifiers for the *Syn5* group, see appendix A for details). Token frequencies varied from corpus to corpus as well.

The same child-specific introducing elements were used for the analysis of nouns as well as for that of verbs, adjectives and adverbs of every corpus. 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 item, 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 analyzed in order to determine the number of nouns in each corpus that was successfully categorized by any of the distributional cues.

Table 5.2: Distributional contexts for the different *Syn* groupings.

Syn0	**a**	{∅} + x
	b	{∅} + x + *-(e)s*
Syn1	**a**	{*a, an*} + x
Syn2	**a**	{*the*} + x
	b	{*the*} + x + *-(e)s*
Syn3	**a**	{*this, that, these, those*} + x
	b	{*this, that, these, those*} + x + *-(e)s*
Syn4	**a**	{POSSESSIVE} + x
	b	{POSSESSIVE} + x + *-(e)s*
Syn5	**a**	{QUANTIFIER} + x
	b	{QUANTIFIER} + x + *-(e)s*
Syn6	**a**	{PREPOSITION} + x
	b	{PREPOSITION} + x + *-(e)s*
Syn7	**a**	{WH- ELEMENT} + x
	b	{WH- ELEMENT} + x + *-(e)s*

Similarly, other clusters of words containing verbs, adjectives or adverbs were analyzed 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 described above, 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*** *brown* ***dog***).

5.4.2. Phonological cues

The set of phonological cues that are said to indicate grammatical category membership include both segmental and suprasegmental features. The literature characterizes a number of different phonological cues to distinguish open from close class words on the one hand, and nouns and verbs on the other hand (cf. Durieux & Gillis, 2001; Kelly, 1992, 1996; Monaghan, Chater & Christiansen, 2005; see chapter 3 for a review). The phonological cues that make nouns differ from verbs describe word length (i.e. both in terms of syllable number and phoneme number, as well as word duration in general), consonant quality (i.e. nasality and vocal cord vibration), vowel quality (i.e. backness and frontness as well as vowel height) and suprasegmental features (i.e. primary stress position).

However, empirical studies that have tested the validity of every phonological cue have found out that many of those cues are rather weak at categorizing words when taken individually (Durieux & Gillis, 2001; Monaghan, Chater & Christiansen, 2005). Thus, for categorization purposes, a whole constellation of many different cues is said to be more powerful than individual cues.

One of the most recent studies that addresses this issue considers up to sixteen different phonological aspects that make members of different grammatical categories differ (see Monaghan, Chater & Christiansen, 2005). However, the purpose of the present study was to test the availability and usefulness of phonological cues to very young language learners. By having a multiplicity of phonological cues, together with all other types of cues considered, the risk of having children create several different categories for the same grammatical category of noun increased (i.e. one is likely to encounter the same problem as the one found with distributional analyses that consider a wide range of distributional frames (see section 5.3.1. for a discussion)). Furthermore, recognition and use of sets of multiple cues require more sophisticated abilities on the part of language learners. Thus, in order to minimize the risk of getting multiple categories inferred from a potentially too wide range of possible values, a limited selection of the strongest phonological cues was considered for this study.

Among all the possible cues, a total of four different cues were considered. They were labelled as *Phon1*, *Phon2*, *Phon3* and *Phon4*, and each one made reference to one of the features that are said to characterize English nouns vs. verbs (i.e. word length, consonant quality, vowel quality and stress position, respectively). The selected cues were defined as follows:

- *Phon1*: this makes reference to word length. Generally speaking, nouns are said to be longer and contain more syllables than verbs. Thus, words with two syllables or more scored 1 in this category, and monosyllabic words scored 0.

Therefore, the only cue to measure the superiority of nouns as far as length is concerned was syllable number and all the other length cues (i.e. number of phonemes and word duration) were discarded. The transcriptions from the corpora under consideration showed orthographic representations of mother-child interactions and there was no access to the phonological or phonetic transcriptions of those speech data. Therefore, the exact duration of individual phonemes or the possible reduction, vocalization or omission of certain segments could not be accurately retrieved on the basis of the orthographical transcriptions alone. That made number of phonemes and word duration not reliable measures for the present analysis and, therefore, they were discarded.

Besides, unlike some of the previous studies that had a similar goal (Kelly, 1992), this analysis considered word forms exactly as they were spoken to the child, that is, fully inflected forms and not only bases or bare forms. In this context, testing the *Phon1* cue turned particularly relevant in the sense that, unlike other word classes, verbs are more likely to get syllabic inflections (e.g. the *–ing* suffix). Since two-syllable words scored 1 in the *Phon1* category, the measure will be useful to see if verbal syllabic inflections neutralize the potential superiority of nouns as far as number of syllables is concerned (i.e. even if most English verbs are monosyllabic and most English nouns are disyllabic or even longer, the token frequency of verb forms such as *taking* or *putting* might neutralize the speakers' perception that nominal elements have more syllables than any other elements).

- *Phon2*: this refers to consonant quality. In particular, it is connected to the voice features of final consonants. If a word finishes with a consonant, it is more likely to be a voiced consonant if the word is a noun rather than a verb. Thus, words ending in a voiced consonant scored 1 in this category. Words which finished in a vowel or a voiceless consonant scored 0.

This was the only cue that was considered regarding consonant quality. The other consonantal cue (i.e. consonant nasality) was discarded because of the kind of input analyzed here. The literature claims that if a word finishes in a nasal consonant, it is more likely to be a noun than a verb (Kelly, 1992). However, empirical studies that have tested the claim have not considered fully inflected forms but bare forms and bases (Kelly, 1992; 1996). Thus, they did not consider verb forms with inflections which, unlike the inflections of all other word classes, are more likely to contain final nasal consonants (i.e. all *–ing* verb forms and also many past participle verb forms with very high token frequency like *been*, *done*, *gone*, *given* or *taken*). Thus, since some verbal inflections and their statistical frequency would have neutralized the final nasality effect of nouns, the cue was discarded.

- *Phon3*: this was connected to the height features of vowels. In English, vowels in nouns tend to be low, whereas vowels in verbs tend to be high. In order to test whether words fitted this description, the stressed vowel in every word from the corpora was analyzed. Thus, words from the corpus whose stressed vowel was low (i.e. /æ/, /ʌ/, /ɒ/ or /ɑ:/) scored 1 in this category. Words whose stressed vowel was mid (i.e. /e/, /ɜ:/ or /ɔ:/) or high (i.e. /u:/, /ʊ/, /ɪ/ or /i:/) scored 0. Words whose stressed syllable contained any of the English diphthongs equally scored 0, even if one of the vowels in the diphthong was also low.

The literature suggests that this difference is only apparent with high frequency words (Durieux & Gillis, 2001; Kelly, 1996; 1992). However, I tested every open class word in the corpora against this criterion in order to see the actual distribution of English low vowels regarding grammatical classes. These previous studies also provide another vocalic feature that distinguishes nouns from verbs (i.e. backness and frontness) which was discarded here. As with high features, the back vs. front difference is only apparent among high frequency words. Besides, since I wanted to keep as fewer phonological cues as possible (i.e. one cue of each type), I carried out a pilot analysis with one of the corpus which would help me decide which vowel quality cue to discard. From Anne's corpus, I examined only nouns and verbs and tested them against both vowel quality cues. Table 5.3 shows the results obtained.

Table 5.3: Analysis of all vowel quality phonological cues from the noun and verb tokens in the corpus of Anne.

	Nouns		**Verbs**	
	Total tokens	**Total %**	**Total tokens**	**Total %**
Back vowels	3,104	22.91	3,588	24.87
Low vowels	4,871	35.96	2,686	18.62

As can be seen, when considering all words regardless of their token frequency, the difference between nouns and verbs is more evident under the vowel height criterion than under the vowel backness criterion. Furthermore, in order to test the statistical weight of these two phonological cues, a Mann-Whitney U-test was performed with these cues as predictor variables in order to see whether their distribution made nouns and verbs differ significantly.

Results from the Mann-Whitney U-test show that the differences between nouns and verbs as far as vowel height is concerned were highly significant, with nouns being more likey to contain low vowels than verbs ($Z = -4.685, p < 0.001$). On the contrary, as far as back vowels are concerned, there were no significant differences found among their distribution in the noun and verb categories ($Z = -1.042, p > 0.05$). Thus, on the basis of this evidence, from the two possible vowel quality cues, the one which referred to vowel height was considered (i.e. *Phon3*) and the one which referred to vowel backness and frontness was discarded.

- *Phon4*: this reflected stress patterns within the word. Disyllabic words with an iambic stress pattern (i.e. stress on the last syllable) are more likely to be verbs, whereas words with a trochaic stress pattern (i.e. stress on the first syllable) are more likely to be nouns. Under this schema, disyllabic words from the corpora that had a trochaic stress pattern scored 1. Monosyllabic words or disyllabic words with an iambic stress pattern scored 0.

As with distributional cues, the set of English nouns obtained from the children corpora were tested against the selected phonological cues in order to see the accuracy with which such cues categorized English nouns. Similarly, the set of English verbs, adjectives and adverbs obtained from the corpora were analyzed under the same criteria, in order to test the likelihood with which the selected phonological cues might overcategorize other elements which are not nouns. So, unlike previous studies (cf. Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007; Kelly, 1996), not only verbs but also adjectives and adverbs were checked against nouns.

Besides, this study also differs from previous ones in as much as I did not test verbs with the set of phonological cues that categorize verbs, but I tested them with the same set of phonological cues that was used for nouns. The object of the study was not to see whether a given set of cues would successfully categorize verbs, but the aim was to see if the same set of phonological cues that categorized nouns was too general as to include other elements which were not nouns in the same category. Ideally, at the phonological test, a high score is expected from the analysis of nouns, as that would indicate accurate categorization of these elements. On the contrary, a low score is expected from the set of verbs, adjectives and adverbs, since that would indicate low accuracy in the categorization of such elements and thus, a low risk of overcategorization, which would make the whole set of phonological cues reliable for the categorization of nouns.

In the particular case of phonological cues, a further advantage of this study is that phonological cues were carefully selected so that each made reference to a different phonological property of the word and every cue was not incompatible with the rest. That would make it possible for tokens to get a gradable score (from 0 to 4) in the phonological test. Thus, if a token scored 4 in the phonological test, this means it fulfilled all four phonological criteria under consideration, which would describe that token as highly marked with nominal features as fas as phonology is concerned. On the contrary, if a word scored 0 or 1 in the phonological test, this would indicate that such word fulfilled either none or just one of the four phonological criteria that were selected. This would make that particular word weakly marked as a noun in terms of phonology. Therefore, even if individual cues are not powerful enough by themselves and a constellation of phonological cues is necessary to distinguish word classes, combinations themselves have a gradable nature that can distinghish between strongly marked tokens and weakly marked tokens. With the cues under consideration, a limited set of twelve possible combinations arose:

- *NoPhon*: words which scored 0 in all 4 *Phon* categories.
- *OnlyPhon1*: words which scored 1 in *Phon1* and 0 in the rest of *Phon* categories.
- *OnlyPhon2*: words which scored 1 in *Phon2* and 0 in the rest of *Phon* categories.
- *OnlyPhon3*: words which scored 1 in *Phon3* and 0 in the rest of *Phon* categories.
- *PhonComb12*: words which scored 1 in *Phon1* and *Phon2* only.
- *PhonComb13*: words which scored 1 in *Phon1* and *Phon3* only.
- *PhonComb14*: words which scored 1 in *Phon1* and *Phon4* only.
- *PhonComb23*: words which scored 1 in *Phon2* and *Phon3* only.
- *PhonComb123*: words which scored 1 in all four *Phon* categories except *Phon4*.
- *PhonComb124*: words which scored 1 in all four *Phon* categories except *Phon3*.
- *PhonComb134*: words which scored 1 in all four *Phon* categories except *Phon2*.
- *PhonComb1234*: words which scored 1 in all four *Phon* categories.[11]

Thus, words which fell into the *PhonComb1234* group or any of the *PhonComb* groups with a constellation of three out of four phonological features were considered to be strongly marked as a noun as far as phonology is concerned. On the contrary, words which fell within the *NoPhon* group or any of the three possible *OnlyPhonX* groups were considered to be weakly marked as nouns regarding phonology. The

[11] Since *Phon1* subsumes *Phon4*, it is impossible to have *OnlyPhon4* words, or *PhonComb24* words, or *PhonComb34* words or *Phoncomb234* words. In any case, these would be *Phon14* words, *Phon124* words, *Phon134* words and *Phon1234* words, respectively.

expected distribution is that more nouns will fall under the *PhonComb1234* group, while more words from other grammatical categories will fall within the *NoPhon* group.

5.4.3. Semantic cues

While access to meaning was irrelevant for infants facing initial word segmentation tasks, semantic information is crucial and of great value for older children when placing previously segmented lexical units into their corresponding grammatical category. As sketched in chapter 3, theories of semantic bootstrapping postulate that, when learning their first language, children make use of conceptual knowledge and semantic information to create grammatical categories (Pinker, 1984). Thus, they seem to equate the semantic category of "type of object/person" to the grammatical category of noun, as well as the semantic category of "action" to the grammatical category of verb.

However, as some authors have remarked, the grammatical category of nouns includes more than simply names of objects, animals and people (Nelson et al., 1993). Crucially, some nominal elements actually share more semantic characteristics with verbs than with other nouns, since they refer to actions rather than object referents (e.g. *a trip, a job*). The point turns to be even more difficult in the case of English, where certain words can actually be classified as both nouns and verbs (e.g. *call*, *kiss*, *walk*). In those cases, the nominal and the verbal form will not only share semantic features, but also phonological ones. The only thing that distinguishes nouns and verbs in those cases is their respective distributional context.

Acknowledging that the core traditional definition of nouns as labels for people, animals and things only includes a subset of all English nouns raises the question of what the actual percentage of nouns which belong to this subset is (i.e. how big the subset is, considering all English nouns). It also raises the issue of whether this smaller subset can be taken to account for all of the nouns that young children are exposed to and will later acquire.

Thus, the test consisted now in classifying all nouns from the children corpora into different groups, according to the semantic features they bore. The main objective is to analyze the consistency and reliability with which semantic information to grammatical categorization is represented in the input addressed to language learning children. Particular attention was paid to the semantic overlap between nominal elements which describe actions and prototypical verbal elements which equally describe actions. Such contradictory information might be especially misleading for any child who relies on semantic information to form grammatical categories, since they would wrongly classify action nouns as verbs.

Thus, as with distributional and phonological cues, a set of semantic cues which can be said to identify nouns was selected. Again, not only nouns, but also verbs, adjectives and adverbs from the corpora were tested against the set of selected cues. That way made it possible to analyze the degree of overlap between semantic and grammatical categories and, therefore, work out the risk of misclassifying elements into a wrong grammatical category on the basis of semantic information. However, in this case, the overlap across elements from different categories is expected to be minimal, or less important than in the case of distributional or phonological cues, given that nouns share no semantic properties with adjectives or adverbs, and they share very few of them with verbs. Following the work by Nelson et al. (1993), the set of semantic categories that was selected for the analysis include the following:

- *Sem1*: this was the group for proper nouns, which semantically refer to single individuals and never expand to whole-class reference. A number of classical studies (e.g. Katz, Baker & Macnamara, 1974; Gelman & Taylor, 1984) have already provided evidence for the fact that children understand proper nouns and common nouns differently from very early stages of language development. Proper nouns of people, animals, toys, stories, songs, places, holidays etc. scored 1 in this category and 0 everywhere else. Family terms like *mummy*, *daddy*, *granny*, etc. also scored 1 when they were used as vocatives or proper nouns and they had a unique conceptual referent. The KWAL utility of the CLAN program was used to distinguish the particular use each word had in every context (e.g. '*Daddy is coming now*' vs. '*This is Tim and that's his daddy*').

- *Sem2*: this was the category that best fitted the traditional notional definition of nouns as labels for names of people, animals and objects. In particular, the *Sem2* group was meant to include all common count nouns which could be considered members of the basic level category in Rosch's (1978) terms[12] (e.g. *dog*, *apple*, *chair*). Words from the corpora that matched this definition scored 1 here and they scored 0 everywhere else.

- *Sem3*: this was the category that included all mass nouns in the corpora (e.g. *milk*, *paper*). They differ from the nouns that fitted into the *Sem2* category in the sense that *Sem3* words do not denote discrete entities but

[12] See footnote 4.

whole substances. This difference is also reflected in syntax, since mass nouns do not combine with the same type of determiner as count nouns. Furthermore, children seem to be aware of such differences by their second year of life (Soja, 1992). Words that matched the *Sem3* description scored 1 here and they scored 0 everywhere else.

- *Sem4*: this was the category which Nelson et al. (1993) labelled as XBLOCS (i.e. not basic level object categories). These were words which fitted neither in the *Sem3* category (since they did not designate substances), nor in the *Sem2* category (either because they were not basic names for people, animals or objects, or else because the names of people, animals or objects that they designated were too abstract or too general for them to be considered basic level concepts). While Nelson et al. (1993) distinguish several different types of XBLOC nouns, for the purpose of the present study only two different groups of the *Sem4* category were made:

    - *Sem4a*: this was the group for words which described actions (e.g. *kiss*, *help*, *trip*). Nouns which matched this description scored 1 here and 0 in the other *Sem* categories. Crucially, verbs which matched this description also scored 1 here.
    - *Sem4b*: this group included all other XBLOC nouns. Among them were words which denoted locations and places, both indoors and outdoors (e.g. *kitchen*, *park*, *school*); words which described generic categories and belonged to the superordinate level in

Rosch's (1978) terms (e.g. *animal, person, object*); words which referred to abstract events or social gatherings (e.g. *lunch, party*); words which described person roles (e.g. *doctor, brother*)[13]; words which denoted natural phenomena (e.g. *sky, cloud, rain*); words which referred to temporal entities (e.g. *morning, day*); words which designated quantities (e.g. *drop, spoonful*); and words which described materials (e.g. *wood, wool*). Then, words which semantically referred to any of these groups scored 1 in this category and 0 in all the other *Sem* categories.

What motivates the division of the *Sem4* group into two subgroups is that, unlike Nelson et al. (1993), the purpose of my analysis is not to provide an accurate description and classification of XBLOC nouns in general. Instead, the main objective is to test the amount of English nouns that lack a direct semantic component available to language learners (i.e. which nouns are XBLOCS and which ones are not). Besides, it was also important to analyze the degree of semantic overlapping between nouns and other words such as verbs, for which the *Sem4a* category was created.

The corpora from all the four children were analyzed separately. After the analysis, all scores obtained for every noun, verb, adjective and adverb in every category were added up. Results and total scores are discussed in the following sections.

5.5. Statistical tests

Following previous studies with similar goals (e.g. Monaghan, Chater & Christiansen, 2005), two different types of statistical tests were used. On the one hand, a

[13] These differed from *Sem2* words like *man* or *boy* in the sense that *Sem2* words have a direct visual and conceptual mapping to the person they refer, whereas the words for person roles are conventionally or socially defined.

test of significance was carried out with all different types of cues, in order to examine the extent to which those cues can be used to distinguish English nouns from other kinds of open class words. When significant differences between nouns and other elements are found with a given set of cues, then it follows that this set of cues is useful to classify English nouns in a reliable way, and that the differences found between the subset of nouns and the rest of open class words considered in this study are not due to mere chance.

However, as Morgan, Shi & Allopenna (1996) point out, significant differences found between the means of nouns and other words in terms of individual cues does not yield cue validity for those invididual cues alone. The distributions of two groups of words with different means may still have a considerable overlap. This implies that individual cues cannot still be considered good predictors of category membership and the results from this study might not be generalizable to the whole set of words to which English learning children are exposed.

Therefore, a test of diagnosticity was also carried out with several groups of cues together. This would give reliable information as to the extent to which constellations of several cues together can accurately classify English nouns and distinguish them from all other open class words.

The results obtained from these two tests are given in chapter 6. Results obtained from individual measures (i.e. tests of significance) are reported separately, and they are followed by the results obtained from the analysis of the contribution of each combination of cues to word categorization (i.e. tests of diagnosticity).

Chapter 6:
Results

6.1. General data

As mentioned in chapter 5, 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. Since the age at which children were first recorded varies in each case, the total amount of time considered in this study will also be different for every child.

Thus, data from Anne include a total time span of 9 months (i.e. from 1;10 until she was 2;6), data from Aran include a total of 8 months (i.e. from 1;11 to 2;6), data from Becky include a total of 6 months (i.e. from 2;0,07 to 2;6) and data from Carl include a total of 11 months (i.e. from 1;8 to 2;6). On average, this study considers the language to which those chidren have been exposed during 8.5 months, ranging from the second half of their first year to the first half of their second year of life.

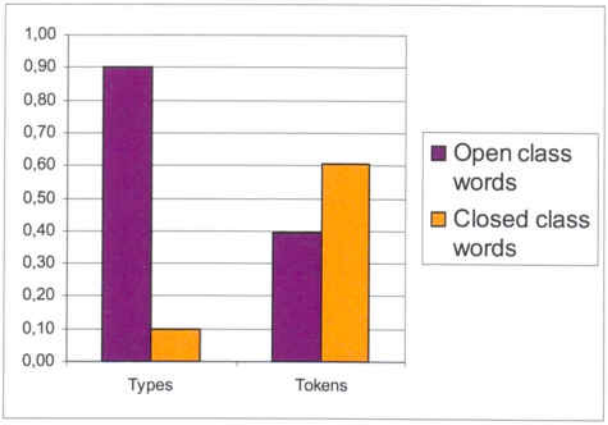
From the list of words in each corpus obtained with the FREQ utility of the CLAN program, a subsample of words was selected, from wich all function words were discarded, and open class words were classified into two different groups: nouns, on the one hand, and all other open class words (i.e. verbs, adjectives and adverbs) on the other (see section 5.2). Table 6.1 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. For the proportions obtained from every individual child, see appendix B.

Table 6.1: 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	9,641	143,272	0.067	0.90	0.39
Total nouns	5,388	51,577	0.104	0.56	0.36
Total other	4,233	88,047	0.048	0.44	0.61

The results obtained from the split between open class words and function words reflect what many previous studies have already mentioned regarding the open- vs. closed-class word distinction (e.g. Morgan, Shi & Allopenna, 1996; Shi, Morgan & Allopenna, 1998; Monaghan, Chater & Christiansen, 2005): while function words exhibit very little lexical variability (i.e. there are very few word types), each function word occurs with very high frequency (i.e. there are many token instantiations for every single type). Figure 6.1 shows the proportion of open class word types and tokens that were left in the sample of the corpora under consideration, once all function words had been discarded.

Figure 6.1.: Proportion of open and closed class types and tokens.

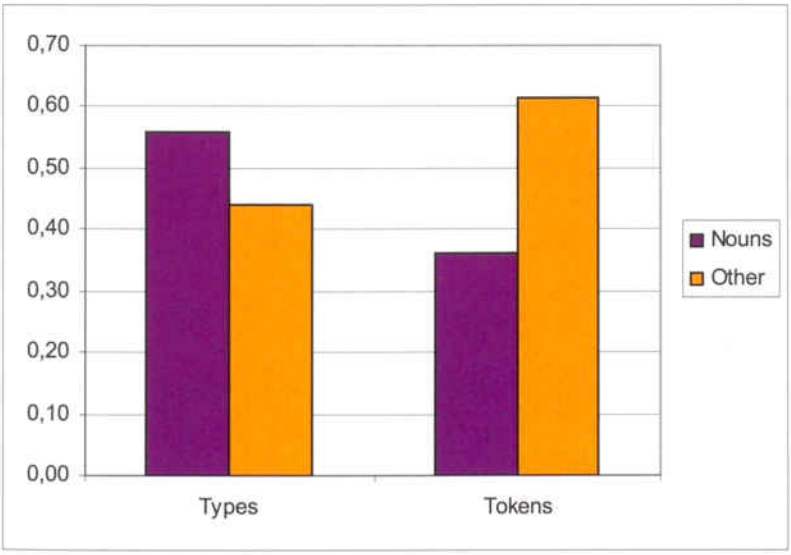


As can be seen in table 6.1, once the selected subcorpora was created for the present analysis by removing all closed class words, most types were kept from the original corpus size. However, the token proportion left over after the selection indicates that, while very few word types were removed, those types were extremely frequent.

As indicated in chapter 5, once all the closed class words were withdrawn, a further split was performed on the remaining open class word group. Thus, nouns were distinguished from the rest of open class words (i.e. verbs, adjectives and adverbs), which were labelled together as "other".

The type and token proportions of the newly emerged "noun" group and "other" group are also uneven (see figure 6.2). Thus, nouns exhibit a higher lexical variability than other open class words, since there are far more nominal types than verbal, adjectival or adverbial types. However, while being more lexically limited, other open class words are far more frequent than nouns, since there are far more tokens in the "other" group than in the "noun" group.

Figure 6.2.: Proportion of nouns and other open class types and tokens.



Once open class words were split into the "noun" group and the "other" group, each of the elements within each group was tested against the set of cues described in chapter 5, in order to see whether they satisfied or not the conditions associated to English nouns. The results obtained from these analyses are reported in the following sections. I will first report results obtained from each type of cue individually (i.e. distributional, phonological and semantic cues alone), and I will move then to the analyses of cues in combination.

6.2. Distributional cues

6.2.1. Descriptive data

For the analysis of distributional cues, the set of nouns and the set of other open class words obtained from the corpus preparation were tested against the set of distributional contexts described in chapter 5 (see section 5.4.1). The total number of nouns that satisfied each of the distributional contexts established is shown in table 6.2(a), while table 6.2(b) shows the results obtained from the equivalent analysis with the rest of open class words. Tables 6.2(a) and 6.2(b) show the total number of types and tokens that were found in each of the distributional contexts under consideration from the corpora of all four children together. Individual tables with data corresponding to the corpus of each individual child are found in Appendix C.

As shown in the tables, the total number of nominal types and tokens that were found in all distributional contexts considered (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 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 6.2.(a): Total of nouns found in distributional contexts.

	NOUNS				
	Total Types	Total Tokens	Type/Token Ratio	Type Proportion	Token Proportion
{Ø} + x (Syn0a)	2,438	17,405	0.140	0.45	0.34
{Ø} + 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
{this, that, these, those} + x (Syn3a)	728	2,292	0.318	0.14	0.04
{these, those} + 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 6.2.(b): 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
{this, that, these, those} + x (Syn3a)	433	1,615	0.268	0.10	0.02
{these, those} + 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

6.2.2. Tests of significance

6.2.2.1. 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 6.3 below.

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 indefinite articles (*Syn1*), definite articles (*Syn2*), demonstrative determiners (*Syn3*), possessive determiners (*Syn4*), quantifiers (*Syn5*), prepositions (*Syn6*) as well as *wh-* interrogative elements such as *which* or *whose* (*Syn7*). Nouns are also more likely than other open class words to be followed by morpheme *–(e)s* (*Syn0b*), and they are also more often found in morphosyntactic frames with one determiner on the left and morpheme *–(e)s* on the right (*Syn2b-Syn7b*). 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.

However, it should be noted that the fifteen distributional cues do not categorize the same proportion of nominal types equally, and all cues are different in their strength and saliency. Thus, for example, the difference between nouns and other open class words in terms of presence or absence of definite articles (*Syn2*) is considerably big. On the contrary, presence or absence of *wh-* interrogative elements (*Syn7*) gives a smaller difference between nouns and the rest of words.

Table 6.3.: 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

6.2.2.2. 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 6.4.

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 between a determiner and a plural morpheme (*Syn2b-Syn7b*) or followed by plural morphemes only (*Syn0b*).

Also in line with the results found with types, nominal tokens are also more likely to be preceded by indefinite articles (*Syn1*), definite articles (*Syn2*), demonstrative determiners (*Syn3*), possessive determiners (*Syn4*) and quantifiers (*Syn5*). However, while there was a significant difference between the "noun" group and the "other" group in terms of presence of a preposition (*Syn6a*) and a wh- interrogative element (*Syn7a*), the results indicate that it is other open class word tokens and not noun tokens which are more likely to be preceded by such elements.

Furthermore, as was also shown with types, even if distributional cues are highly significant and mean differences reveal an advantage for nouns over other open class words, there are differences among the various distributional contexts in terms of strength. Thus, distributional contexts like precedence of articles (both definite and indefinite) as well as precedence of possessive determiners appear to account for a larger proportion of nominal tokens than the rest of distributional cues.

Table 6.4.: 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	0.000
{∅} + x + -(e)s (Syn0b)	0.57	0.00	-26.452	0.000
{a, an} + x (Syn1)	1.19	0.47	-23.528	0.000
{the} + x (Syn2a)	1.73	0.26	-31.936	0.000
{the} + x + -(e)s (Syn2b)	0.28	0.00	-20.066	0.000
{this, that, these, those} + x (Syn3a)	0.43	0.38	-4.846	0.000
{these, those} + x + -(e)s (Syn3b)	0.07	0.00	-11.554	0.000
{POSSESSIVE} + x (Syn4a)	0.66	0.08	-20.467	0.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.18	0.00	-14.081	0.000
{QUANTIFIER} + x (Syn5a)	0.42	0.40	-2.575	0.010
{QUANTIFIER} + x + -(e)s (Syn5b)	0.22	0.00	-19.341	0.000
{PREPOSITION} + x (Syn6a)	0.37	0.39	-4.713	0.000
{PREPOSITION} + x + -(e)s (Syn6b)	0.06	0.00	-12.503	0.000
{WH- ELEMENT} + x (Syn7a)	0.14	0.15	-5.826	0.000
WH- ELEMENT} + x + -(e)s (Syn7b)	0.02	0.00	-6.531	0.000

6.2.2.3. Frequency groups

As seen in chapter 5, three different frequency groups were created out of the whole sample of all tokens in order to test the effects that word frequency had on the different classification systems that were established. Table 6.5 shows a summary of the total number of nouns and other open class words that fell within each of the frequency groups established. As shown in the table, the three frequency groups keep a similar noun *versus* other type proportion, with more nominal types than other open word types. These proportions are also similar to the ones obtained from the whole corpus (see figure 6.2). Therefore, the frequency groups that were established keep the natural characteristics of the kind of language to which children are exposed in general, and they are not biased towards nouns only or other open class words only.

Table 6.5: Total number of nominal and non-nominal types in each of the frequency groups.

	Low-frequency words		**Mid-frequency words**		**High-frequency words**		**Total corpus**	
	Total	**Prop.**	**Total**	**Prop.**	**Total**	**Prop.**	**Total**	**Prop.**
Noun	1.883	0.57	2,337	0.57	1,168	0.51	5,388	0.56
Other	1,399	0.42	1,729	0.42	1,105	0.48	4,233	0.44
Total	3,282	--	4,066	--	2,273	--	9,621	--

In order to test whether the significance values obtained from the analysis with all tokens together were affected by the split of tokens in three frequency groups, a Mann-Whitney U-test was performed with each of the resulting groups (i.e. low-frequency tokens, mid-frequency tokens and high-frequency tokens). Tables 6.6(a-c) show the results obtained from the Mann-Whitney U-tests performed with each of the token frequency groups.

Table 6.6(a): Mann-Whitney U-test for the 15 distributional cues with low-frequency tokens.

Distributional cues	Nouns	Other	Z	Significance
{∅} + x (Syn0a)	3.23	18.67	-31.618	0.000
{∅} + x + -(e)s (Syn0b)	0.57	0.00	-12.205	0.000
{a, an} + x (Syn1)	1.19	0.47	-9.156	0.000
{the} + x (Syn2a)	1.73	0.26	-12.224	0.000
{the} + x + -(e)s (Syn2b)	0.28	0.00	-8.707	0.000
{this, that, these, those} + x (Syn3a)	0.43	0.38	-1.072	0.283
{these, those} + x + -(e)s (Syn3b)	0.07	0.00	-3.114	0.002
{POSSESSIVE} + x (Syn4a)	0.66	0.08	-7.513	0.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.18	0.00	-5.200	0.000
{QUANTIFIER} + x (Syn5a)	0.42	0.40	-1.879	0.060
{QUANTIFIER} + x + -(e)s (Syn5b)	0.22	0.00	-7.703	0.000
{PREPOSITION} + x (Syn6a)	0.37	0.39	-3.231	0.001
{PREPOSITION} + x + -(e)s (Syn6b)	0.06	0.00	-4.497	0.000
{WH- ELEMENT} + x (Syn7a)	0.14	0.15	-3.302	0.001
WH- ELEMENT} + x + -(e)s (Syn7b)	0.02	0.00	-2.113	0.035

Table 6.6(b): Mann-Whitney U-test for the 15 distributional cues with mid-frequency tokens.

Distributional cues	Nouns	Other	Z	Significance
{∅} + x (Syn0a)	1.04	3.45	-38.559	0.000
{∅} + x + -(e)s (Syn0b)	0.41	0.00	-18.848	0.000
{a, an} + x (Syn1)	0.53	0.12	-15.477	0.000
{the} + x (Syn2a)	0.69	0.07	-21.869	0.000
{the} + x + -(e)s (Syn2b)	0.20	0.00	-13.667	0.000
{this, that, these, those} + x (Syn3a)	0.11	0.08	-2.443	0.015
{these, those} + x + -(e)s (Syn3b)	0.04	0.00	-6.935	0.000
{POSSESSIVE} + x (Syn4a)	0.22	0.04	-12.298	0.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.09	0.00	-9.015	0.000
{QUANTIFIER} + x (Syn5a)	0.18	0.11	-3.303	0.001
{QUANTIFIER} + x + -(e)s (Syn5b)	0.16	0.00	-13.048	0.000
{PREPOSITION} + x (Syn6a)	0.15	0.08	-4.652	0.000
{PREPOSITION} + x + -(e)s (Syn6b)	0.05	0.00	-7.867	0.000
{WH- ELEMENT} + x (Syn7a)	0.03	0.01	-3.841	0.000
WH- ELEMENT} + x + -(e)s (Syn7b)	0.01	0.00	-3.223	0.001

Table 6.6(c): Mann-Whitney U-test for the 15 distributional cues with high-frequency tokens.

Distributional cues	Nouns	Other	Z	Significance
{∅} + **x** (Syn0a)	12.34	65.05	-31.413	0.000
{∅} **+ x + -(e)s** (Syn0b)	1.65	0.00	-14.262	0.000
{a, an} + x (Syn1)	4.22	1.57	-18.248	0.000
{the} + x (Syn2a)	6.36	0.87	-23.932	0.000
{the} + x + -(e)s (Syn2b)	0.80	0.00	-12.314	0.000
{this, that, these, those} + x (Syn3a)	1.70	1.31	-7.013	0.000
{these, those} + x + -(e)s (Syn3b)	0.24	0.00	-9.411	0.000
{POSSESSIVE} + x (Syn4a)	2.52	0.22	-17.631	0.000
{POSSESSIVE} + x + -(e)s (Syn4b)	0.63	0.00	-10.097	0.000
{QUANTIFIER} + x (Syn5a)	1.51	1.32	-3.103	0.002
{QUANTIFIER} + x + -(e)s (Syn5b)	0.65	0.00	-12.661	0.000
{PREPOSITION} + x (Syn6a)	1.32	1.35	-3.371	0.001
{PREPOSITION} + x + -(e)s (Syn6b)	0.15	0.00	-9.192	0.000
{WH- ELEMENT} + x (Syn7a)	0.58	0.54	-5.256	0.000
WH- ELEMENT} + x + -(e)s (Syn7b)	0.07	0.00	-5.713	0.000

As can be seen, the results from the significance tests with each of the frequency groups are very similar to those obtained from the analysis with all tokens together. For both mid-frequency tokens and high-frequency tokens, all distributional cues taken as predictor variables make nouns and other open class words differ significantly. The results obtained with low-frequency tokens show highly significant differences for most distributional contexts as well, except for elements preceded by a demonstrative (*Syn3a*) and a quantifier (*Syn5a*).

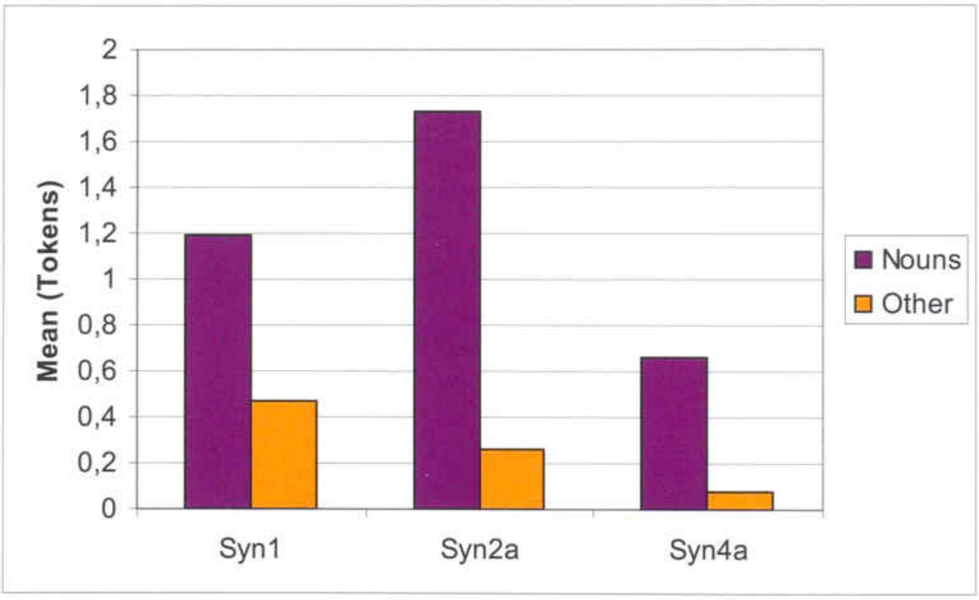
The patterns found with each of the frequency groups are also similar to those found with all tokens together: nouns appear to be more likely to be preceded by any of the syntactic elements taken as predictor variables except for Syn0a (i.e. absence of distributional context), which is a context where other open class words are more likely to be found. The only exceptions to this pattern are prepositions (Syn6a) and wh-interrogative elements (Syn7a), which appear to correlate with other open class words rather than nouns for both the low-frequency and the high-frequency tokens. However, such correlations were also found in the analysis with all tokens together.

6.2.3. 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? As Shi, Morgan & Allopenna (1998) pointed out, grammatical categories may differ from one another in terms of their means for a given set of cues. However, even if those differences are found significant, the overlap of categories for each of the cues may be considerable.

Figure 6.3 indicates the distribution of the mean differences of nouns and other open class words as far as tokens are concerned for three of the cues that reached the highest significance in the Mann-Whitney U-tests (i.e. preceding indefinite article (*Syn1*), preceding definite article (*Syn2a*) and preceding possessive determiner (*Syn4a*), which reached a significance of $p < 0.001$ for types as well as for tokens in general and within each of the frequency groups). Even for these three cues alone, there is a considerable overlap between nouns and other open class words.

Figure 6.3.: Distribution of nouns and other open class words for three of the most significant distributional contexts.



The data obtained indicate that there are still 2,009 non-nominal tokens that are preceded by an indefinite article (*Syn1*) or 1,112 preceded by a definite article (*Syn2a*). Overall, there are a total of 9,015 tokens which belong to the other open class word group (i.e. they are mainly adjectives) and which are found in any of the fourteen contexts which were estimated to be typical of nouns only. Given a language learner who has no prior knowledge of grammatical category membership, these environmental data would yield a number of incorrect classifications.

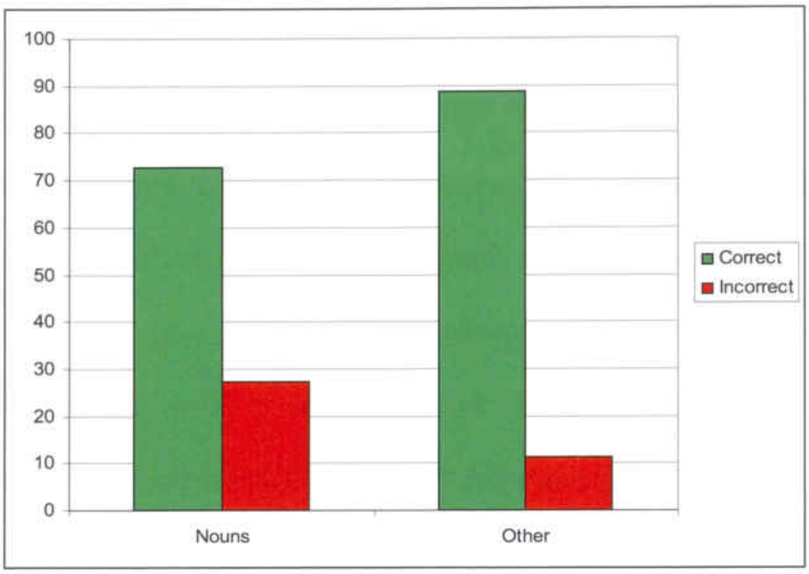
Therefore, the combined contribution of all the distributional cues towards correct classification of words was further tested using a multivariate linear discriminant analysis. Discriminant analysis provides a classification of items into categories based on a set of predictor variables. The chosen 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 following sections contain the results from such analyses for each of the groups under consideration.

6.2.3.1. 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). Results from this analysis are shown in figure 6.4.

Figure 6.4.: Classification of noun types and other types for all distributional contexts.



Correct classification of all types from the four corpora was further assessed using the leave-one-out cross-validation method. This is a method that works out the classification on all types except for one, and then assesses whether the resulting classification system correctly applies to the type that was left out. The results obtained from the leave-one-out discriminant analysis were identical to those obtained with the standard discriminant method (i.e. Wilks $\lambda = 0.611$, $\chi^2 = 4734.621$, $p < 0.001$) resulting in 72.8% of nouns and 88.8% of other open class words correctly classified, which indicates that the set of distributional contexts chosen as predictor variables is a valid criterion for classification purposes.

A further validation of these results was also 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 as to allow for such validation to be done without subsamples affecting significance results. With this method, 50% of the total number of types is 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.

The stepwise method is another way of developing a discriminant function. It starts with an empty equation and builds it by incorporating each of the predictor variables in one step at a time. Predictor variables are included in the equation if they significantly add to the predictive function. Thus, a stepwise discriminant analysis does not necessarily include all the predictor variables. This maximizes the explanatory strength of the variables as well as of the classificatory system as a whole. Furthermore, the order in which predictor variables are included in each step responds to the statistical weight of each of the variables, that is, the first-entered cues contribute most to correct classification (Meyers et al., 2006). Thus, we also obtain a variable ranking, with the strongest variable first and the weakest variable last.

When the stepwise discriminant analysis was performed with the types from the child corpora, only 11 of the 15 distributional contexts were entered as predictor variables. The order in which each of the distributional contexts was introduced in the discriminant function was the following:

1. Absence of nominal distributional context (*Syn0a*)
2. Precedence of definite article (*Syn2a*)
3. Precedence of possessive determiner (*Syn4a*)
4. Plural morphology at the end (*Syn0b*)
5. Precedence of indefinite article (*Syn1*)
6. Precedence of definite article and plural morphology at the end (*Syn2b*)
7. Precedence of preposition (*Syn6a*)
8. Precedence of quantifier and plural morphology at the end (*Syn5b*)
9. Precedence of *wh-* interrogative elements (*Syn7a*)
10. Precedence of demonstrative determiner (*Syn3a*)
11. Precedence of possessive determiner and plural morphology at the end (*Syn4b*)

Four other distributional cues were considered too weak to contribute significantly to the discriminant function, and were thus excluded from the analysis.

The four cues which were discarded from the stepwise discriminant analysis were the following ones: precedence of demonstrative determiner and plural morphology at the end (*Syn3b*), precedence of quantifier (*Syn5a*), precedence of preposition and plural morphology at the end (*Syn6b*), and precedence of *wh-* interrogative element and plural morphology at the end (*Syn7b*).

When the 11 strongest cues were entered in a stepwise analysis, the results were identical to those obtained in the above mentioned discriminant analysis when all 15 cues were entered together in a single step. Thus, correct classification of nominal types reached 72.8% of the cases, while correct classification of other open class words reached 88.8% of the cases, with an overall total successful classification of 79.8 types. This was highly significant: Wilks $\lambda = 0.611$, $\chi^2 = 4729.839$, $p < 0.001$.

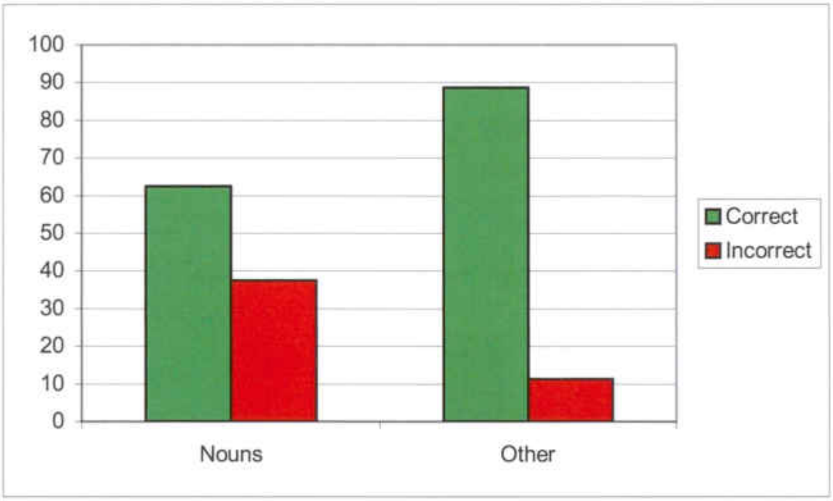
An additional objective of this study was to compare different categorization contexts and to test whether frames (i.e. two morphosyntactic categorizing elements with one intervening element in the middle) had equivalent categorization strength to bigrams (i.e. one morphosyntactic categorizing element with one intervening element to the left or to the right). From the results obtained from the stepwise discriminant analysis we can see that, as far as types are concerned, bigrams had a bigger classificatory strength than frames, given the fact that very few biframe contexts were found among the 11 strongest cues in the stepwise analysis.

Thus, from the six different distributional frames that were considered in the initial analysis, only three of them contributed significantly to the discriminant function in the stepwise analysis, and even then, none of them occupied the top positions in any of the first steps (i.e. biframe '*the* + x + *-(e)s*' (*Syn2b*) was included in the sixth step, biframe 'QUANTIFIER + x + *(e)s*' (*Syn5b*) was included in the eighth step, and biframe 'POSSESSIVE + x + *-(e)s*' (*Syn4b*) was included in the last step). Furthermore, from the four distributional cues that were discarded from the stepwise analysis, three of them corresponded to briframes (i.e. biframe 'DEMONSTRATIVE + x + *-(e)s*' (*Syn3b*), biframe 'PREPOSITION + x + *-(e)s*' (*Syn6b*) and biframe 'WH- ELEMENT + x + *-(e)s*' (*Syn7b*)) and only one of the discarded cues corresponded to a bigram (i.e. bigram 'QUANTIFIER + x' (*Syn5a*)).

In order to further test the hypothesis that bigrams had a greater contribution to grammatical categorization that frames did, two additional discriminant analyses were performed, one with all the distributional cues that described bigrams only, and one with all the distributional cues that described frames separately. For the discriminant analysis using bigrams only, eight different cues were introduced simultaneously: plural morphology at the end (i.e. *Syn0b*), precedence of indefinite article (i.e. *Syn1*), precedence of definite article (i.e. *Syn2a*), precedence of demonstrative determiner (i.e. *Syn3a*), precedence of possessive determiner (i.e. *Syn4a*), precedence of quantifier (i.e. *Syn5a*), precedence of preposition (i.e. *Syn6a*) and precedence of *wh-* interrogative element (i.e. *Syn7a*).

The results found from this analysis revealed a correct classification for nominal types that reached 62.5%, as well as a total of 88.7% types that were correctly classified from the other open class word group. This reveals a considerable drop in completeness scores, although successful scores in accuracy are kept (i.e. most other open class words are correctly classified as "other" and not included in the noun category). Overall, correct classification reached a total of 74.0% types in both groups together, Wilks λ = 0.741, $\chi^2 = 2876.629$, $p < 0.001$. The results from this analysis are shown in figure 6.5.

Figure 6.5.: Classification of noun types and other types for the distributional contexts that corresponded to bigrams.



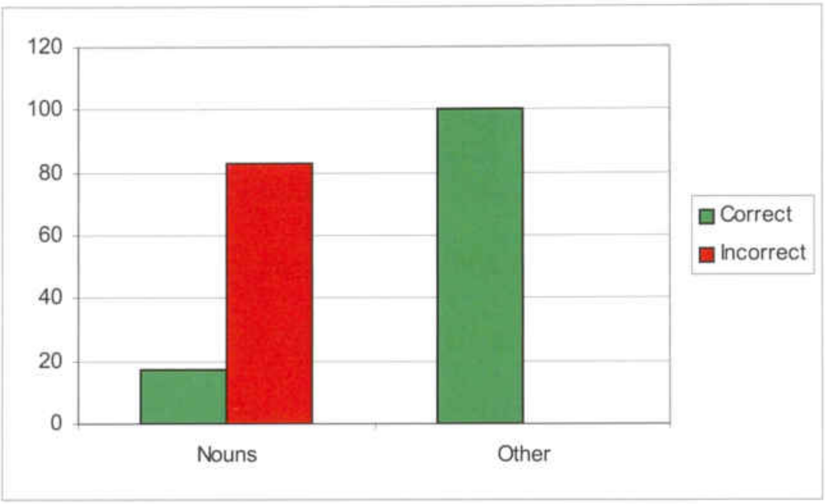
For the discriminant analysis using frames only, six different cues were introduced simultaneously as predictor variables: precedence of definite article and plural morphology at the end (i.e. *Syn2b*), precedence of demonstrative determiner and plural morphology at the end (i.e. *Syn3b*), precedence of possessive determiner and plural morphology at the end (i.e. *Syn4b*), precedence of quantifier and plural morphology at the end (i.e. *Syn5b*), precedence of preposition and plural morphology at the end (i.e. *Syn6b*) and precedence of wh- element and plural morphology at the end (i.e. *Syn7b*)[14]. The results found from this analysis are shown in figure 6.6.

As seen from figure 6.6, the discriminant analysis with frames revealed a correct classification for nominal types of only 17.0%, but 100% of accurate classification for types which belong to the other open class word group. This is not surprising, as it only

[14] Note that the *Syn0a* cue was not introduced in any of these last analyses, neither with bigrams only nor with frames. *Syn0a* subsumed all those elements that were not found in any of the nominal distributional contexts established in the study, that is, it described absence of any categorizing element either to the left or to the right of the intervening element. Thus, its description did not match the bigram or the frame requirements and was not considered for these two last analyses.

indicates that there were no adjectives, verbs or adverbs in English child directed speech that contained a plural nominal *–(e)s* morpheme[15].

Figure 6.6.: Classification of noun types and other types for the distributional contexts that corresponded to frames.



However, despite the high score in accuracy, the system showed very low completeness scores, as most of the elements within the noun category actually fell out of the classification, as biframe morphosyntactic contexts in English can only account for a limited number of nominal elements. So, despite the overall correct classification of 53.5% of all types in both groups, the analysis with frames only was considered as non-successful as far as the categorization of nouns is concerned.

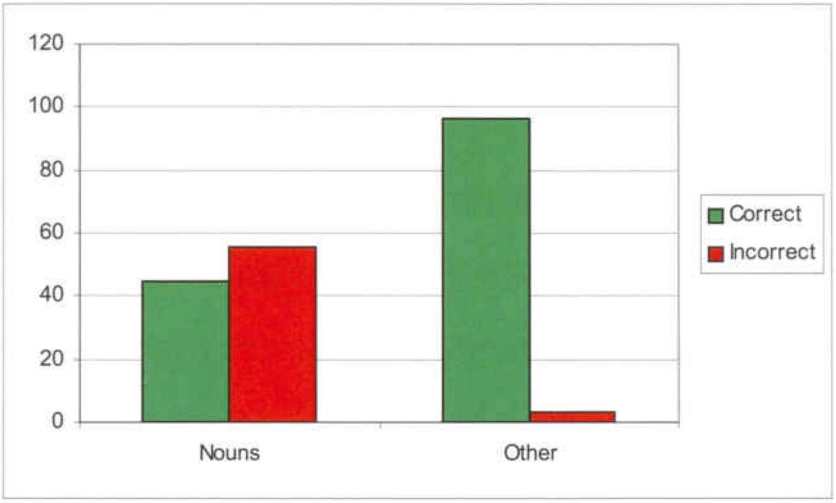
6.2.3.2. Tokens

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

[15] It is true that the English verbal morpheme to indicate third person singular in the present tense has the same form as the nominal morpheme for plural. Thus, there do exist verb forms which end in *–(e)s*. However, they were always preceded by either a noun or a strong subject pronoun, but they never occurred preceded by any of the determiners considered in the analysis.

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 $\lambda = 0.921$, $\chi^2 = 790.324$, $p < 0.001$). Results from this analysis are shown in figure 6.7 below.

Figure 6.7.: Classification of noun tokens and other tokens for all distributional contexts.



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).

With the leave-one-out cross-validation method, the discriminant analysis for tokens when all distributional cues were entered simultaneously revealed almost identical results to those of the standard discriminant analysis mentioned above: 44.5% of nouns and 96.5% of other open class words were correctly classified, Wilks λ = 0.921, $\chi^2 = 790.324$, $p < 0.001$. 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 $\lambda = 0.919$, $\chi^2 = 401.987$, $p < 0.001$.

A further stepwise discriminant analysis with distributional cues was also performed with tokens, as well as with types. When the stepwise discriminant analysis was performed with all tokens from the child corpora, 12 of the 15 distributional contexts were entered as predictor variables (i.e. one more variable than in the corresponding analysis with types), and three variables were discarded. Each cue was entered in one step at a time, with the statistically strongest cues being entered first, and the weakest cues last. The order in which each of the distributional cues was introduced in the discriminant function was the following:

1. Precedence of definite article (*Syn2a*)
2. Absence of nominal distributional context (*Syn0a*)
3. Precedence of quantifier and plural morphology at the end (*Syn5b*)
4. Precedence of possessive determiner (*Syn4a*)
5. Plural morphology at the end (*Syn0b*)
6. Precedence of definite article and plural morphology at the end (*Syn2b*)

7. Precedence of possessive determiner and plural morphology at the end (*Syn4b*)
8. Precedence of indefinite article (*Syn1*)
9. Precedence of preposition (*Syn6a*)
10. Precedence of demonstrative determiner (*Syn3a*)
11. Precedence of preposition and plural morphology at the end (*Syn6b*)
12. Precedence of demonstrative determiner and plural morphology at the end (*Syn3b*)

Accuracy scores for nouns in the stepwise analysis were very similar to those found with the standard discriminant analysis (i.e. 96.6% of other open class words were correctly classified as "other"). However, completeness scores did not improve, as only 43% of nouns were correctly classified as nouns, and 57% of nouns missed their correct classification. Overall, the stepwise discriminant analysis gave a total of 66.6% of tokens which were correctly classified, Wilks $\lambda = 0.921$, $\chi^2 = 787.315$, $p < 0.001$.

Interestingly, most of the cues that were entered in the stepwise discriminant analysis with types were also entered in the corresponding analysis with tokens, and they were entered in more or less the same step. This indicates that each of the distributional contexts had a similar statistical weight for types and tokens. The only striking difference is that biframe distributional contexts were slightly more powerful in the token analysis than in the type analysis.

To start with, biframe *Syn5b* (i.e. 'QUANTIFIER + x + *-(e)s*') was taken at the third step in the token analysis, while it only occupied the eighth step in the analysis with types. Similarly, biframe *Syn4b* (i.e. 'POSSESSIVE + x + *-(e)s*') occupied the seventh step in the token analysis, while it was taken in the last step in the analysis with types.

Furthermore, frames *Syn6b* (i.e. 'PREPOSITION + x + *-(e)s*') and *Syn3b* (i.e. 'DEMONSTRATIVE + x + *-(e)s*') were included in the stepwise analysis with tokens (although they were in the last steps), but they were discarded in the stepwise analysis with types. Thus, from the three distributional cues that were discarded in the stepwise discriminant analysis with tokens, only one of them corresponded to a biframe (i.e. biframe *Syn7b*, '*wh-* ELEMENT + x + *-(e)s*'), and the other two cues corresponded to bigram contexts (i.e. *Syn5a*, 'QUANTIFIER + x' and *Syn7a*, '*wh-* ELEMENT + x').

In order to see whether this translated into a discriminant analysis with frames only with better results than those obtained with types, two further discriminant analyses were also ran with all tokens: one in which only bigram cues were entered as predictor variables, and another one where only biframe cues were entered. However, the results found from the analyses with tokens were not better than those found with the equivalent analyses ran with types (see figures 6.8 and 6.9 below, as well as figures 6.5 and 6.6. for types).

For the analysis with bigrams only, the same distributional cues that were used in the analysis with types were also used in the analysis with tokens. The results show a slight improvement in accuracy, since 88.7% of types were correctly classified, but the token percentage of correct classification rose to 96.3% in the token analysis. However, completeness scores were quite low, since only 32.4% of the nominal tokens were assigned their right category in the discriminant analysis, Wilks $\lambda = 0.956$, $\chi^2 = 433.990$, $p < 0.001$.

Figure 6.8.: Classification of noun tokens and other tokens for the distributional contexts that corresponded to bigrams.

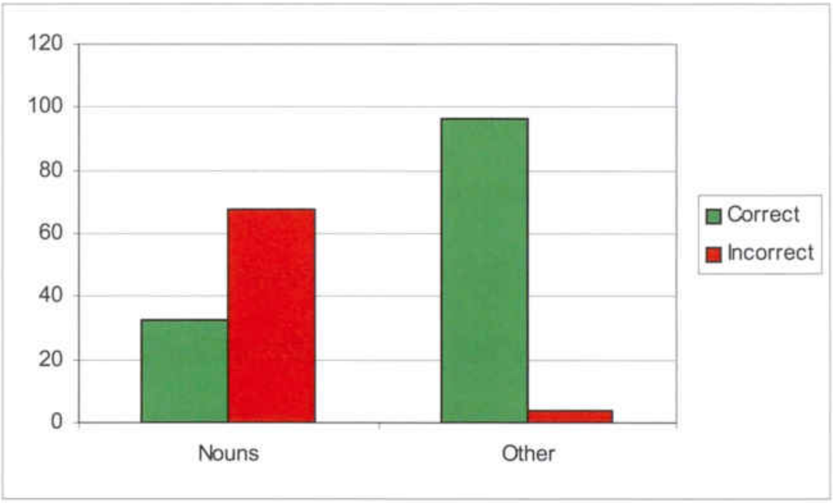
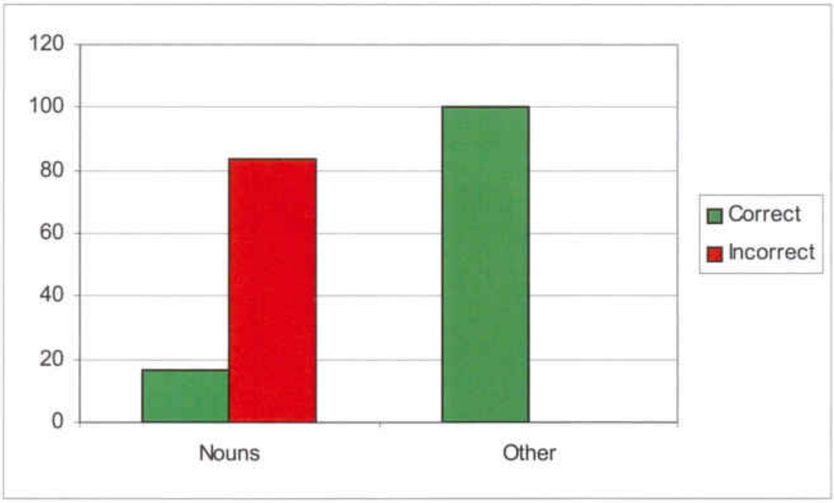


Figure 6.9.: Classification of noun tokens and other tokens for the distributional contexts that corresponded to frames.

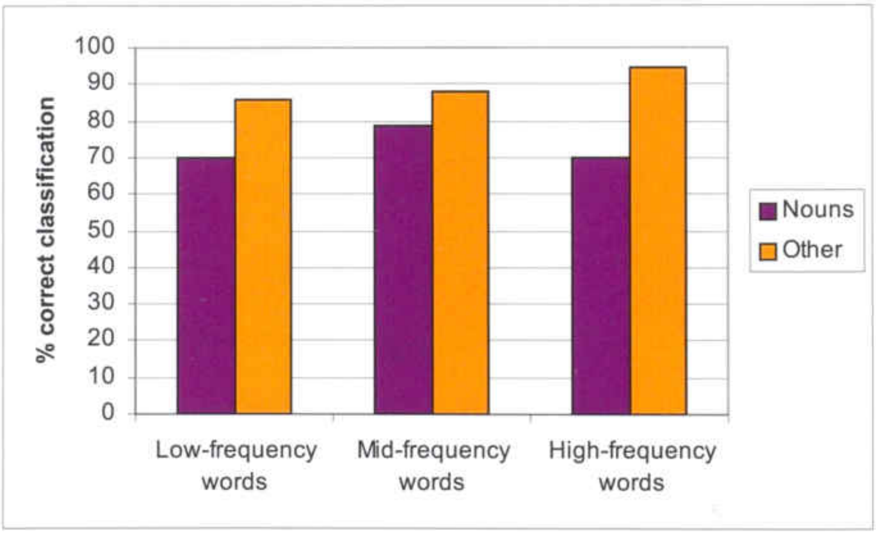


As far as frames are concerned, completeness scores did not improve either, and the results were almost identical to the analysis ran with types. Thus, although classification was very accurate (i.e. 100% of other open class word types and tokens were correctly classified), there were very low scores for completeness, as very few nominal elements were assigned to their right category (i.e. 17% of nominal types and 16.7% of nominal tokens). Therefore, although the analysis with frames only produced a total of 53.4% of correctly classifed tokens, and the discriminant analysis results were highly significant (i.e. Wilks $\lambda = 0.977$, $\chi^2 = 228.242$, $p < 0.001$), the biframe only classificatory system was overall taken as non-successful.

6.2.3.3. Frequency groups

In order to test whether token frequency had an effect on correct classification, a discriminant analysis was also carried out with all the fifteen distributional cues together for each of the three frequency groups separately. Figure 6.10 shows the percentages of correct classification only for both nouns and other open class words in each of the frequency groups.

Figure 6.10: Percentage of correct classification of noun tokens and other tokens for all distributional contexts across different frequency groups.



In the analysis with low-frequency words, a total of 70.1% of nouns and 85.6% of other open class words were correctly classified in their corresponding grammatical categories. This was highly significant, Wilks $\lambda = 0.668$, $\chi^2 = 1320.053$, $p < 0.001$. The same analysis with mid-frequency words gave a total of 78.5% of nouns and 88% of other open class words which were correctly classified, Wilks $\lambda = 0.604$, $\chi^2 = 2046.444$, $p < 0.001$. Last, the same analysis with high-frequency words resulted in a total of 65.8% of nouns and 94.7% of other open class words correctly classified.

Thus, surprisingly, while accuracy and completeness scores were relatively high for low- and mid-frequency words, and accuracy scores were still high among high-frequency words, there was a drop in completeness as fas as high-frequency nominal tokens are concerned. This might have been due to the fact that the high-frequency noun group includes a particular subset of elements that might have altered completeness results: the group of proper nouns. Because of the particular characteristics of child-directed speech, proper nouns exhibit quite a high token frequency, especially regarding the name of each child in each corpus. As seen in chapter 3, proper nouns are not necessarily difficult to categorize, because of their heavy semantic load as well as their features of referentiality. However, precisely because they are referential in nature, they exhibit their own particular syntactic characteristics, and thus, the syntactic environments in which they occur makes them very different from the rest of common nouns.

Thus, it might follow that the low completeness scores obtained in the discriminant analysis with high-frequency words might have been affected by the high frequency of proper nouns and their special syntactic characteristics. This would have biased somehow the overall correct classification among high-frequency words. In order to eliminate the outlier effect that proper nouns might have had, an additional discriminant analysis was run with a new subset of high-frequency words from which all proper nouns were removed. While the whole original high-frequency group contained a total of 2,273 cases, the new high-frequency group which was obtained after removing all proper nouns included a total of 2,165 cases.

When the new discriminant analysis was carried out with these 2,165 high-frequency words and all the fifteen distributional cues as independent variables, a total of 70.2% of nouns and 94.8% of other open class words were correctly classified, Wilks $\lambda = 0.713$, $\chi^2 = 727.759$, $p < 0.001$. 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. Within mid-frequency words, a total of 82.5% of tokens were correctly classified and overall correct classification among low-frequency words reached a total of 76.7% of tokens. Therefore, as frequency increases, a slight improvement in accuracy is observed as far as overall correct classification is concerned.

For the leave-one-out cross-validation discriminant analyses, the results found in each of the frequency groups were almost identical to the results obtained from the standard discriminant analyses. Thus, low-frequency words exhibited an overall correct classification of 76.7% of correctly classified tokens, 70.1% of correctly classified nouns, and 85.6% of correctly classified other open class words, Wilks $\lambda = 0.668$, $\chi^2 = 1320.053$, $p < 0.001$. Mid-frequency words exhibited a total of 82.5% of correctly classified words, 78.4% were correctly classified nouns and 87.9% were correctly classified other open class words, Wilks $\lambda = 0.604$, $\chi^2 = 2046.444$, $p < 0.001$. For the cross-validation of high-frequency words, 82.7% of tokens were correctly classified, with correct nouns reaching 70.2% and other open class words 94.8%, Wilks $\lambda = 0.713$, $\chi^2 = 727.759, p < 0.001$.

A further cross-validation with 50% of randomly selected cases was also performed for each of the frequency groups. The results found from this second cross-validation were again very similar to those found in the original analysis, indicating that the set of predictor variables is robust and valid as a classification method. Overall correct classification among low-frequency words reached 77.6% (Wilks $\lambda = 0.681$, $\chi^2 = 626.982$, $p < 0.001$), 83.2% for mid-frequency words (Wilks $\lambda = 0.607$, $\chi^2 = 990.099$, $p < 0.001$) and 82.8% for high-frequency words (Wilks $\lambda = 0.702$, $\chi^2 = 387.876$, $p < 0.001$).

Additional stepwise discriminant analyses with distributional cues were also performed for each of the frequency groups, the same as with all types and all tokens.Table 6.7 shows a summary of the number of steps which discriminant analyses took for each frequency group as well as the particular distributional cue which was introduced as predictor variable at every step.

Table 6.7: Statistical weight of distributional cues in stepwise discriminant analyses with all frequency groups.

Step	**Low-frequency words**	**Mid-frequency words**	**High-frequency words**
1.	{∅} + x (***Syn0a***)	{∅} + x (***Syn0a***)	{*the*} + x (***Syn2a***)
2.	{DEMONSTR.} + x (***Syn3a***)	{*the*} + x (***Syn2a***)	{∅} + x (***Syn0a***)
3.	{QUANTIFIER} + x (***Syn5a***)	{∅} + x + -*(e)s* (***Syn0b***)	{QUANT}+ x + -*(e)s* (***Syn5b***)
4.	{PREPOSITION} + x (***Syn6a***)	{*a, an*} + x (***Syn1***)	{POSSESSIVE} + x (***Syn4a***)
5.	{*a, an*} + x (***Syn1***)	{POSSESSIVE} + x (***Syn4a***)	{POSS.} + x + -*(e)s* (***Syn4b***)
6.	{*the*} + x (***Syn2a***)	{*the*} + x + -*(e)s* (***Syn2b***)	{∅} + x + -*(e)s* (***Syn0b***)
7.	{POSSESSIVE} + x (***Syn4a***)	{PREPOSITION} + x (***Syn6a***)	{*a, an*} + x (***Syn1***)
8.	--	{QUANT}+ x + -*(e)s* (***Syn5b***)	{*the*} + x + -*(e)s* (***Syn2b***)
9.	--	{QUANTIFIER} + x (***Syn5a***)	{QUANTIFIER} + x (***Syn5a***)
10.	--	{POSS.} + x + -*(e)s* (***Syn4b***)	{DEMONSTR.} + x (***Syn3a***)
11.	--	{WH- ELEMENT} + x (***Syn7a***)	--
12.	--	{PREP.} + x + -*(e)s* (***Syn6b***)	--

As shown in the table, although the number of steps varies across different frequency ranges, distributional cues contribute in a similar way in the three discriminant functions. Thus, several cues were similarly weighted in each of the analyses (i.e. cues '{∅} + x' (*Syn0a*), '{*a, an*} + x' (*Syn1*), '{*the*} + x' (*Syn2a*), '{POSSESSIVE} + x' (*Syn4a*) and '{QUANTIFIER} + x' (*Syn5a*) were more or less equally considered in the three analyses). There was a slight difference between the low-frequency word analysis on the one hand, and the mid- and high-frequency word analyses on the other, since fewer steps were required in the low-frequency word analysis. Besides, all the distributional cues that significantly contributed to the discriminant function in the low-frequency analysis corresponded to bigram contexts, while mid- and high-frequency words required both bigram and biframe contexts.

Furthermore, those distributional cues which proved to be weak and were discarded from the discriminant analyses were the same across the three frequency groups. Thus, none of the analyses included frames *Syn3b* (i.e. {DEMONSTRATIVE} + x + *-(e)s*) or *Syn7b* (i.e. {WH- ELEMENT} + x + *-(e)s*). Besides, contexts *Syn6b* (i.e. {PREPOSITION} + x + *-(e)s*) and *Syn7a* (i.e. {WH- ELEMENT} + x) were also commonly discarded. They were only taken in the mid-frequency word analysis, but they occupied the eleventh and twelfth step respectively (i.e. the last steps).

As for accuracy and completeness, the stepwise analyses revealed very similar results for correct classification to the initial standard discriminant analyses. Thus, overall correct classification for low-frequency tokens was 76.7% (70.1% nouns and 85.6% other open class words, Wilks $\lambda = 0.668$, $\chi^2 = 1321.358$, $p < 0.001$). Correct classification for mid-frequency words was a total of 82.6% of tokens (78.5% nouns and 88.2% other open class words, Wilks $\lambda = 0.605$, $\chi^2 = 2040.658$, $p < 0.001$). Finally, correct classification among high-frequency words reached a total of 82.9% tokens

(70.6% nouns and 94.8% other class words, Wilks $\lambda = 0.714$, $\chi^2 = 725.855$, $p < 0.001$). This confirms, on the one hand, the robustness of the set of predictor variables as a valid system for classification and, on the other hand, the tendency towards the increase of successful categorization as word frequency increases.

6.2.4. Summary of results with distributional cues

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. Significance tests using these cues over a sample of English words have revealed that the differences between English nouns and other English open class words regarding their distributional contexts is highly significant and, therefore, those differences are not the product of mere chance. Furthermore, when diagnosticity was evaluated by means of discriminant analyses, a very high percentage of correctly classified words was obtained as well. For the token analysis, accuracy scores were very high (i.e. 96.5% of correctly classified other open class words) although there was a considerable drop in completeness scores (i.e. only 44.6% of the nouns were correctly classified as nouns). The type analysis yielded better results, with correct classification reaching 72.8% for nouns, and 88.8% for other open class words.

The most important findings as far as distributional cues are concerned lie in the interaction between correct classification and word frequency. With frequency groups, the completeness scores obtained with tokens improved significantly, giving results above chance in all cases. More interestingly, there was a close interaction between word frequency and accurate classification, with high-frequency words being classified more accurately than low-frequency words. Thus, there was a total of 76.7% of low-frequency words which were correctly classified, 82.5% of mid-frequency words and 82.8% of high-frequency words.

6.3. Phonological cues

6.3.1. Descriptive data

For the analysis of phonological cues, the set of nouns and the set of other open class words obtained from the four child corpora were tested against all the possible combinations of phonological cues described in chapter 5. A total of four different phonological cues were used, and all the possible combinations of all these cues yielded a total of twelve different phonological contexts, which were used as predictor variables in the present analysis. These different contexts ranged from total absence of phonological cues, to presence of one single cue only, combinations of two different cues, combinations of three different cues, as well as convergence of all four cues in one word (see section 5.4.2).

The total number of nouns that satisfied each of the phonological variables established is shown in table 6.8(a), and table 6.8(b) shows the results obtained from the same analysis with the rest of open class words. As with distributional cues, the four child corpora were merged and totals were worked out from each corpus, so tables 6.8(a) and 6.8(b) show the total number of types and tokens that satisfied each of the phonological variables under consideration from the corpora of all four children together. For individual tables with data corresponding to the corpus of each individual child, see appendix D.

Table 6.8(a): Total of nouns that satisfied the features of each phonological variable.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	759	8,994	0.084	0.14	0.17
Two syllables or more (***Phon1***)	200	883	0.227	0.04	0.02
Final voiced consonant (***Phon2***)	883	9,071	0.097	0.16	0.18
Low stressed vowel (***Phon3***)	348	3,713	0.094	0.06	0.07
Two syllables or more + Final voiced consonant (***Phoncomb12***)	255	985	0.259	0.05	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	119	468	0.254	0.02	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	497	5,225	0.095	0.09	0.10
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	281	5,438	0.052	0.05	0.11
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	167	1,079	0.155	0.03	0.02
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	986	6,019	0.164	0.18	0.12
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	324	4,446	0.073	0.06	0.09
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	569	5,256	0.108	0.11	0.10

Table 6.8(b): Total of other open class words that satisfied the features of each phonological variable.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	819	32,817	0.025	0.19	0.37
Two syllables or more (***Phon1***)	157	1,467	0.107	0.04	0.02
Final voiced consonant (***Phon2***)	730	19,075	0.038	0.17	0.22
Low stressed vowel (***Phon3***)	332	8,819	0.038	0.08	0.10
Two syllables or more + Final voiced consonant (***Phoncomb12***)	220	1,596	0.138	0.05	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	55	450	0.122	0.01	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	299	3,030	0.099	0.07	0.03
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	120	4,236	0.028	0.03	0.05
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	60	397	0.151	0.01	0.00
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	930	13,061	0.071	0.22	0.15
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	132	833	0.158	0.03	0.01
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	379	2,266	0.167	0.09	0.03

As shown in the tables, the four phonological features chosen for this study which are said to be typical of nouns are also present in a number of other open class word types and tokens as well. However, the more phonological features that converge in one word, the more likely this word is to be a noun. Thus, absence of phonological features (i.e. Phon0) appears to be typical of other open class words, as well as simple phonological variables in which only one phonological feature converges (i.e. Phon1, Phon2, Phon3). However, as clusters of phonological features become more complex, (i.e. combinations of two, three or four features) the likelihood for such words to be nouns also increases. The following sections examine in detail the statistical significance of the differences outlined above as well as the weight of each of the phonological variables and their categorization strength in discriminant analyses.

6.3.2. Tests of significance

6.3.2.1. Types

In order to test the significance of phonological cues among types, parallel to the significance analyses with distributional cues, 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 6.9.

As the table shows, most of the phonological variables that were used in the analysis were highly significant ($p \leq 0.001$ for seven of the twelve phonological variables). Those highly significant variables mostly corresponded to "absence of phonological cues" (i.e. *Phon0*), most clusters made up of two cues (i.e. *Phoncomb13*, *Phoncomb14* and *Phoncomb23*), as well as all the clusters which were made up of three cues (i.e. *Phoncomb123*, *Phoncomb124* and *Phoncomb134*). The variable which contained the cluster of all four phonological features (i.e. *Phoncomb1234*) was still significant, with $p < 0.01$ and so was *Phon3*, which grouped all words that contained a stressed low vowel.

Table 6.9: Mann-Whitney U-test for the 12 phonological cues with all types.

Phonological variables	Nouns	Other	Z	Significance
No phonological cues (***Phon0***)	0.14	0.19	-6.947	0.000
Two syllables or more (***Phon1***)	0.04	0.04	-0.130	0.897
Final voiced consonant (***Phon2***)	0.16	0.17	-1.111	0.267
Low stressed vowel (***Phon3***)	0.06	0.08	-2.630	0.009
Two syllables or more + Final voiced consonant (***Phoncomb12***)	0.05	0.05	-1.096	0.273
Two syllables or more + Low stressed vowel (***Phoncomb13***)	0.02	0.01	-3.263	0.001
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	0.09	0.07	-3.731	0.000
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	0.05	0.03	-5.837	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	0.03	0.01	-5.443	0.000
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	0.18	0.22	-4.391	0.000
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	0.06	0.03	-6.501	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	0.11	0.09	-2.614	0.009

The only exception to significance among clusters which contained two phonological features was variable *Phoncomb12*, which grouped together words which had at least two syllables and ended in a voiced consonant. The features connected to word length and final consonant voicing did not reach any significance in the test either when they were in variables that contained only that single feature (i.e. variables *Phon1* and *Phon2*).

As for the distribution in terms of mean differences, the tendency outlined above holds for most of the variables which showed significant differences between nouns and other open class words. That is, absence of any of the four chosen phonological features or presence of just one single feature appears to be characteristic of other open class words, while nouns are more likely to be described by variables which contain clusters of two, three or four phonological feaures. Thus, for *Phon0* and *Phon3*, there is an advantage of other open class words over nouns. On the contrary, for the two-, three- and four-feature combinations, nouns have an advantage over other open class words.

The only exception to this pattern would be variable *Phoncomb124*, which appears to be more typical of other open class words than of nouns. This is not surprising though, as *Phoncomb124* describes words with two syllables, final voicing and trochaic stress pattern. Most of the elements within the other open class word group were verbs, and verbs are more likely than nouns to receive syllabic inflections. It should be noted that, unlike previous analyses which only include bases and stems (i.e. Kelly, 1992; 1996), my analyses include real naturalistic data with fully inflected forms. Thus, verbal morphemes such as *–ing*, which are rather frequent, might have altered the pattern and given these results. If true, this hypothesis should be confirmed with the analysis of significance for tokens, when frequency effects are taken into account.

6.3.2.2. Tokens

As with distributional cues, in order to test the significance of phonological variables among 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 6.10.

As shown in the table, the results found in the test with tokens are very similar to those obtained with types that were outlined in the previous section. As with types, most of the phonological variables that were used in the analysis were highly significant ($p \leq 0.001$ for seven of the twelve phonological variables again). Those highly significant variables corresponded to the same variables that were equally significant in the analysis with types, namely "absence of phonological cues" (i.e. *Phon0*), three of the clusters that made up of two cues (i.e. *Phoncomb13*, *Phoncomb14* and *Phoncomb23*), and the clusters which were made up of three cues (i.e. *Phoncomb123*, *Phoncomb124* and *Phoncomb134*). Those variables which reached a significance of $p <$ 0.01 in the analysis with types (i.e. *Phon3* and *Phoncomb1234*) also reached the same significance with tokens.

As for the distribution in terms of mean differences, the tendency outlined in the analysis of types is confirmed with the results obtained from the analysis of tokens. Thus, for variables including no phonological features (i.e. *Phon0*) and presence of one single phonological feature (i.e. *Phon1*, *Phon2* and *Phon3*), mean proportions were higher among other open class words than among nouns, indicating that other open class words are more likely than nouns to contain none of the phonological features under consideration, or just one of them.

Table 6.10: Mann-Whitney U-test for the 12 phonological cues with all tokens.

Phonological variables	Nouns	Other	Z	Significance
No phonological cues (***Phon0***)	1.67	7.76	-7.460	0.000
Two syllables or more (***Phon1***)	0.16	0.35	-0.102	0.918
Final voiced consonant (***Phon2***)	1.68	4.51	-1.296	0.195
Low stressed vowel (***Phon3***)	0.69	2.08	-2.707	0.007
Two syllables or more + Final voiced consonant (***Phoncomb12***)	0.18	0.38	-1.114	0.265
Two syllables or more + Low stressed vowel (***Phoncomb13***)	0.09	0.11	-3.256	0.001
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	0.97	0.72	-3.746	0.000
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	1.01	1.00	-5.770	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	0.20	0.09	-5.440	0.000
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	1.12	3.08	-4.680	0.000
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	0.83	0.20	-6.567	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	0.97	0.54	-2.761	0.006

On the contrary, for the three- and four-feature combinations, mean proportions were higher among nouns than among other open class words, indicating that nouns are more likely than other open class words to contain at least three of the phonological features under consideration, if not all of them. The only exception to this pattern is again variable *Phon124*, where mean proportions of other open class words are higher that those of nouns. Recall that the same distribution was also found in the analysis with types. As mentioned before, most of the elements within the "other" group were verbs, since adjectives and adverbs are far less frequent, both in terms of lexical variability (i.e. types) as well as in terms of token frequency. Since verbs are more likely than nouns to receive syllabic inflections, then the frequency of fully inflected verbal forms such as those containing the verbal morpheme *–ing* would have given these results.

6.3.2.3. Frequency groups

In line with the analyses using distributional values, a Mann-Whitney U-test was performed with each of the frequency groups (i.e. low-frequency words, mid-frequency words and high-frequency words) in order to see whether significance was affected. Tables 6.11(a-c) show the results obtained from the Mann-Whitney U-tests performed with each of the token frequency groups.

As seen from the tables, significance was clearly affected by word frequency, given the fact that only three out of twelve phonological variables reached significance above chance levels among low-frequency words (i.e. *Phomcomb124*, with $p < 0.01$; *Phoncomb23* and *Phoncomb123*, with $p < 0.001$). Significance levels improved with higher frequency. Thus the results with mid-frequency words are very similar to those obtained with types and all tokens together: there were significant differences for all predictor variables except for variables *Phon1*, *Phon2* and *Phoncomb12*, which were also found nonsignificant in the previous analyses, and *Phoncomb1234*. For high-frequency words, the variable indicating absence of phonological cues (i.e. *Phon0*) was highly significant, together with all variables that contained clusters of three and four phonological feataures (i.e. *Phoncomb123*, *Phoncomb124*, *Phoncomb134* and

Phoncomb1234). However, none of the variables with a single phonological feature were found significant, neither were any of the variables with combinations of two features, except for variables *Phoncomb14*, which reached significance of $p < 0.5$.

Table 6.11(a).: Mann-Whitney U-test for the 12 phonological cues with low-frequency tokens.

Phonological variables	Nouns	Other	Z	Significance
No phonological cues (***Phon0***)	0.12	0.14	-1.375	0.169
Two syllables or more (***Phon1***)	0.05	0.05	-0.224	0.823
Final voiced consonant (***Phon2***)	0.13	0.15	-1.110	0.267
Low stressed vowel (***Phon3***)	0.06	0.06	-0.301	0.764
Two syllables or more + Final voiced consonant (***Phoncomb12***)	0.07	0.08	-1.100	0.271
Two syllables or more + Low stressed vowel (***Phoncomb13***)	0.03	0.02	-1.646	0.100
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	0.09	0.08	-0.761	0.447
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	0.05	0.02	-4.168	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	0.04	0.02	-3.612	0.000
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	0.21	0.24	-2.416	0.016
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	0.05	0.04	-1.746	0.081
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	0.11	0.11	-0.400	0.689

Table 6.11(b).: Mann-Whitney U-test for the 12 phonological cues with mid-frequency tokens.

Phonological variables	Nouns	Other	Z	Significance
No phonological cues (***Phon0***)	0.54	0.77	-4.228	0.000
Two syllables or more (***Phon1***)	0.14	0.12	-1.258	0.208
Final voiced consonant (***Phon2***)	0.68	0.61	-0.457	0.648
Low stressed vowel (***Phon3***)	0.26	0.38	-3.547	0.000
Two syllables or more + Final voiced consonant (***Phoncomb12***)	0.16	0.20	-0.874	0.382
Two syllables or more + Low stressed vowel (***Phoncomb13***)	0.08	0.04	-3.108	0.002
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	0.35	0.28	-2.541	0.011
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	0.25	0.11	-4.607	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	0.10	0.06	-2.620	0.009
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	0.71	0.91	-3.721	0.000
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	0.24	0.11	-4.463	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	0.41	0.36	-0.968	0.333

Table 6.11(c).: Mann-Whitney U-test for the 12 phonological cues with high-frequency tokens.

Phonological variables	Nouns	Other	Z	Significance
No phonological cues (***Phon0***)	6.43	28.33	-6.237	0.000
Two syllables or more (***Phon1***)	0.40	1.07	-1.900	0.057
Final voiced consonant (***Phon2***)	6.19	16.11	-1.323	0.186
Low stressed vowel (***Phon3***)	2.56	7.31	-0.664	0.507
Two syllables or more + Final voiced consonant (***Phoncomb12***)	0.42	1.05	-0.368	0.713
Two syllables or more + Low stressed vowel (***Phoncomb13***)	0.19	0.32	-0.136	0.892
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	3.63	2.21	-3.173	0.002
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	4.08	3.64	-1.068	0.286
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	0.66	0.25	-3.185	0.001
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	3.41	10.07	-2.460	0.014
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	3.25	0.53	-5.314	0.000
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	3.50	1.36	-4.722	0.000

The pattern found among the three of the frequency groups was also identical to the one found in the analyses with types and tokens: while other open class words are more likely to be subsumed by those phonological variables that describe words with either just one phonological feature or none of them (i.e. either *Phon1*, *Phon2*, *Phon3*, or *Phon0*), nouns are more likely to be subsumed by those phonological variables that describe words with at least two of the phonological features under consideration. The only exception to this pattern is again variable *Phoncomb124*, which describes bisyllabic words with a trochaic stress pattern and with a final voiced consonant. For this variable, other open class word mean proportions are higher than noun mean proportions, probably because of the effect of verbal inflectional morphology.

6.3.3. Tests of diagnosticity

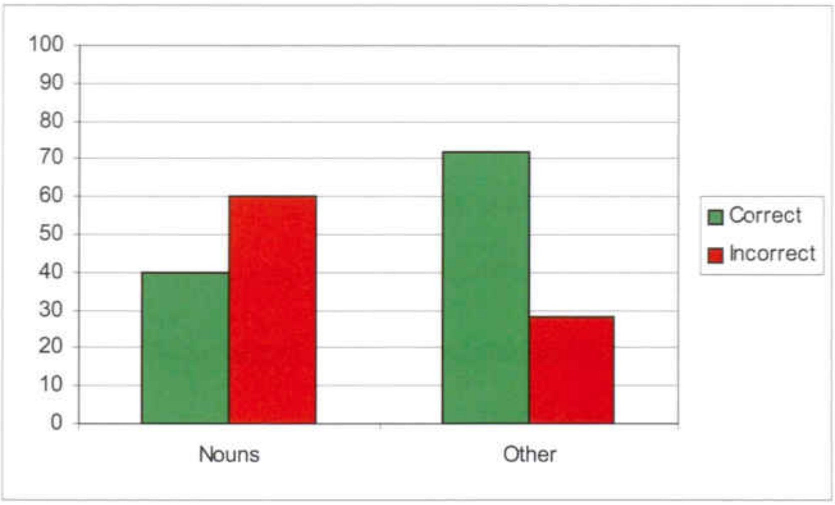
Previous studies which employ a similar methodology to the present study (i.e. Monaghan, Chater & Christiansen, 2005; Shi, Morgan & Allopenna, 1998) show that, although a given variable might not be significant in terms of its mean differences, it might nevertheless contribute to the discriminant function in a multivariate linear discriminant analysis. This only suggests that, in spite of not being significant, this variable combines with the other variables towards accurate classification of a given set of words. The following sections explore the results found with the various discriminant analyses which were performed with phonological variables.

6.3.3.1. Types

First of all, correct classification of all types from the four corpora was assessed for the twelve phonological cues together. When all cues were entered simultaneously, 40.1% of nouns and 71.6% of other open class words were correctly classified. Overall, 53.9% of types were correctly classified. This was highly significant (Wilks $\lambda = 0.980$, $\chi^2 = 197.638$, $p < 0.001$). The cross-validation using the leave-one-out method was also performed, and exactly the same results were found.

Thus, as far as types are concerned, the results suggest that phonological cues are successful at categorizing English nouns in terms of accuracy (i.e. as shown by the high percentage of correctly classified "other" words), although phonological cues are less successful in terms of completeness, given the low percentage of correctly classified nouns, indicating that almost 60% of the nouns in the corpus fell out of the nominal classification and were wrongly classified in the "other" group. Results from this analysis are shown in figure 6.11.

Figure 6.11.: Classification of noun types and other types for all phonological variables together.



The validity of these results was further tested using a cross-validation with 50% of the cases which were randomly selected, as with all previous analyses. With this second cross-validation, very similar results were found as well. There was a slight drop in completeness scores (36.8% of nouns were correctly classified), but there was a slight improvement in accuracy scores as well (74.7% of other open class words were correctly classified). Overall correct classification reached a total of 53.5% of the words, Wilks $\lambda = 0.980$, $\chi^2 = 95.431$, $p < 0.001$.

A further discriminant analysis was performed with all the types from the corpus and using phonological variables, this time using the stepwise method. As in all previous analyses, this would provide information about the statistical weight that each of the variables has in the discriminant function, as well as the way and the order in which they contribute to the classificatory system.

From the twelve possible phonological variables, the stepwise discriminant analysis took a total of eight variables and it discarded the other four. The order in which these eight variables were entered in the discriminant function was the following:

1. Absence of phonological features (*Phon0*)
2. Two or more syllables + final voiced consonant + trochaic stress pattern (*Phoncomb124*)
3. Two or more syllables + low stressed vowel + trochaic stress pattern (*Phoncomb134*)
4. Final voiced consonant + low stressed vowel (*Phoncomb23*)
5. Two or more syllables + final voiced consonant + low stressed vowel (*Phoncomb123*)
6. Two or more syllables + trochaic stress pattern (*Phoncomb14*)
7. Two or more syllables + low stressed vowel (*Phoncomb13*)
8. All four phonological features (*Phoncomb1234*)

As seen from the steps above, none of the cues which were found non-significant in the Mann-Whitney U-tests were included in the stepwise discriminant analysis (i.e. *Phon1*, *Phon2* and *Phoncomb12*). Furthermore, variable *Phon3* (i.e. the one that grouped together all words with stressed low vowel only), which reached a significance of $p < 0.01$ in the Mann-Whitney U-test, was also discarded in the stepwise discriminant analysis. Thus, the variables that proved to be statistically stronger for types and the ones that contributed the most to the discriminant function were those which contained clusters of at least two different phonological features (except for *Phoncomb12*) as well as the variable which grouped all words with none of the four phonological features (i.e. *Phon0*).

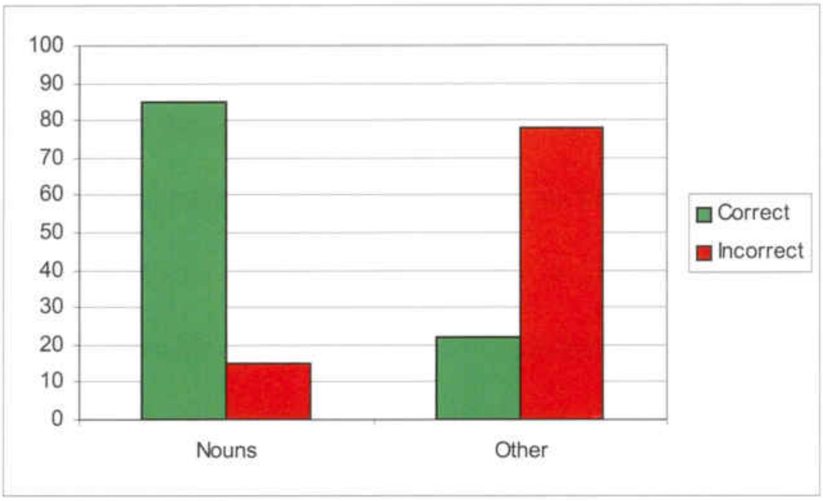
The results found in the stepwise analysis were very similar to the ones found with the standard discriminant analysis: correct classification among nouns only reached a total of 36.3%, but 75.2% of other open class words were correctly classified. Overall correct classification in the stepwise analysis with types was 53.4%, Wilks $\lambda = 0.980$, $\chi^2 = 193.997$, $p < 0.001$.

6.3.3.2. Tokens

As in the previous analyses with distributional cues, correct classification of all tokens from the four corpora was also assessed for the twelve phonological cues together. When all cues were entered simultaneously, 84.9% of nouns and 22.1% of other open class words were correctly classified. Overall, 57.3% of types were correctly classified, Wilks $\lambda = 0.982$, $\chi^2 = 178.169$, $p < 0.001$. Results from this analysis are illustrated in figure 6.12.

Thus, overall correct classification among tokens using phonological cues is very similar to the percentage of correct classification among types using the same cues, with a slight improvement in the token analysis (i.e. 53.9% for types and 57.3% for tokens). However, phonological cues in general are less successful than distributional cues, given the lower overall correct classification scores obtained both in the type and token analyses.

Figure 6.12.: Classification of noun tokens and other tokens for all phonological variables together.



Furthermore, the type and token analyses using phonological cues also differ in that the opposite pattern is found from one analysis to the other. Thus, while there was good accuracy but low completeness among types (i.e. 40.1% of correctly classified noun types *versus* 71.6% of correctly classified other open class words), the reversed pattern was obtained in the token analysis, with very high completeness scores (i.e. 84.9% of the nouns were correctly classified) and very low accuracy scores (i.e. only 22.1% of other open class tokens were correctly classified in their appropiate category, and most of them were wrongly misclassified as nouns).

The same validations that were performed in previous analyses were also performed with the token analysis using all the phonological variables. For the discriminant analysis with the leave-one-out cross-validation method, results were identical to those with the standard discriminant analysis using the same variables. For the second cross-validation using 50% of randomly selected cases, very similar results were found as well, with 82.7% of correctly classified nouns and and 22.4% of correctly classified other open class words. Overall, the cross-validation using 50% of randomly selected cases gave a total of 56.1% of correctly classified words, Wilks $\lambda = 0.979$, $\chi^2 = 102.854$, $p < 0.001$.

A stepwise discriminant analysis was also performed with all tokens using phonological variables. From the twelve possible phonological variables, the stepwise discriminant analysis with tokens took eight variables as well, in line with the analysis with types. Therefore, four of them were discarded as they were found not to contribute significantly to the discriminant function. The order in which the remaining eight variables were introduced is the following:

1. Absence of phonological features (*Phon0*)
2. Final voiced consonant (*Phon2*)
3. Two syllables or more + final voiced consonant + trochaic stress pattern (*Phoncomb124*)
4. Low stressed vowel (*Phon3*)
5. Two syllables or more + low stressed vowel + trochaic stress pattern (*Phoncomb134*)
6. Two or more syllables + final voiced consonant (*Phoncomb12*)
7. Two or more syllables (*Phon1*)
8. All four phonological features (*Phoncomb1234*)

Interestingly, in line with previous studies (i.e. Monaghan, Chater & Christiansen, 2005; Shi, Morgan & Allopenna, 1998) variables which were not found significant (i.e. in this case *Phon1*, *Phon2* and *Phoncomb12*) contributed to the discriminant function, which indicates that, although they are not significant, they combine well with other variables in order to classify words in their grammatical category. Thus, the variables which were discarded from the stepwise discriminant analysis with tokens were, for the most part, all those variables which subsumed clusters of two phonological features (i.e. *Phoncomb13*, *Phoncomb14* and *Phoncomb23*), the exception being variable *Phoncomb12*, which was introduced in the sixth step. Variable *Phoncomb123* was also discarded from the analysis.

The classification results from the stepwise analysis with tokens were very similar to those obtained from the standard discriminant analysis. This time, correct classification reached 85.7% for nouns and 21.3% for other open class words. Overall correct classification was 57.4% for all words, Wilks $\lambda = 0.982$, $\chi^2 = 173.120$, $p < 0.001$. The following section presents the results from the corresponding analyses from each of the frequency groups.

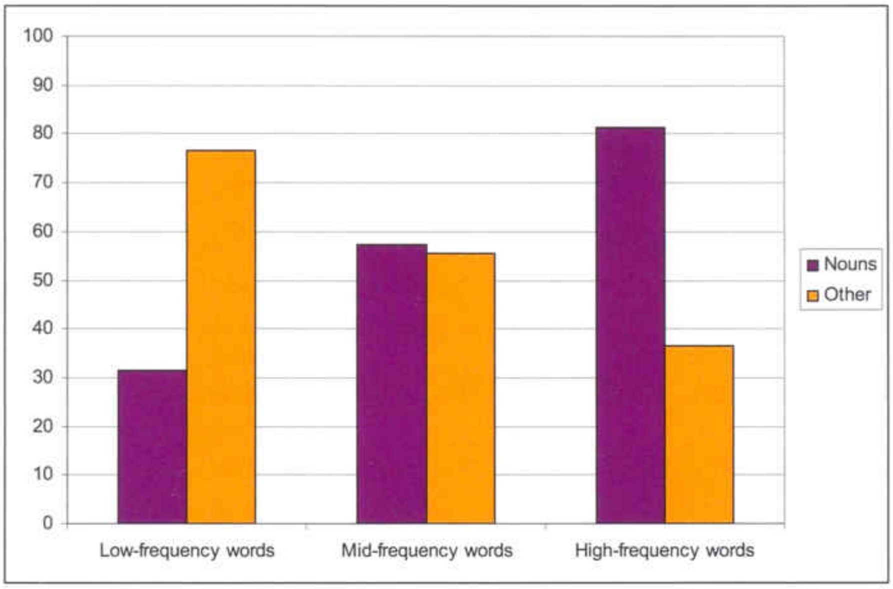
6.3.3.3. Frequency groups

As in the analyses with distributional cues, in order to test whether word frequency had an effect on correct grammatical categorization given a set of predictor variables, a discriminant analysis was also carried out with the twelve phonological cues together for each of the three frequency groups separately. Figure 6.13 shows the percentages of correct classification only for both nouns and other open class words in each of the frequency groups.

In the analysis with low-frequency words, a total of 31.4% of nouns and 76.7% of other open class words were correctly classified in their corresponding grammatical categories. This was highly significant, Wilks $\lambda = 0.986$, $\chi^2 = 45.084$, $p < 0.001$. The same analysis with mid-frequency words gave a total of 57.2% of nouns and 55.6% of other open class words which were correctly classified, Wilks $\lambda = 0.977$, $\chi^2 = 93.195$, $p < 0.001$. Last, the same analysis with high-frequency words resulted in a total of 81.4%

of nouns and 36.4% of other open class words correctly classified, Wilks $\lambda = 0.936$, $\chi^2 = 148.643$, $p < 0.001$.

Figure 6.13: Percentage of correct classification of noun tokens and other tokens for all phonological variables across different frequency groups.



Overall correct classification reached a total of 50.7% of low-frequency words, 56.5% of mid-frequency words and 59.5% of high-frequency words. Thus, there is an overall tendency for categorization to improve as word frequency increases. However such overall tendencies are not even across frequency groups and words classes, since mid-frequency words are the only group which show a more or less even distribution of correct classification in nouns and other open class words.

On the contrary, low-frequency words show a distributions which is similar to the one found in the analysis with types, that is, very low completeness scores (i.e. very few nouns which were correctly classified as nouns) but very high accuracy scores (i.e. many other open class words which were correctly classified as such). On the other hand, high-frequency words exhibit a similar pattern to the one found in the analysis with all tokens, which is the opposite to what types or low-frequency-words show. Thus, among high-frequency words, there are very high scores in completeness (i.e. most nouns are correctly classified as nouns), but very low scores in accuracy (i.e. few other open class words are correctly classified as such and many of them are misclassified as nouns).

The two cross-validations that were performed in each of the previous analyses were also carried out here. In the discriminant analyses using the leave-one-out method, identical results to the standard analyses were found both among low- and mid-frequency words. Almost identical results were found among high-frequency words as well, but with a slight decrease in completeness (i.e. 81.4% of nouns were correctly classified according to the standard analysis, and 81.2% of nouns were correctly classified in a cross-validation discriminant analysis with the leave-one-out method).

As to the second cross-validation using 50% of randomly selected cases, very similar results were found as well. Within the low-frequency word group, this cross-validation resulted in 30.5% of correctly classified nouns and 76.2% of correctly classified other open class words. There was an overall correct classification of 50.5% of words, Wilks $\lambda = 0.986$, $\chi^2 = 23.478$, $p < 0.05$. Within the mid-frequency word group, the same analysis resulted in 56.6% of correctly classified nouns and 53.2% of correctly classified other open class words. Overall correct classification was 55.2% of words, Wilks $\lambda = 0.972$, $\chi^2 = 57.154$, $p < 0.001$. Last, within high-frequency words, this analysis resulted in 80.9% of correctly classified nouns and 34.3% of correctly classified other open class words. Overall correct classification in this frequency group was 58.3% of the words, Wilks $\lambda = 0.931$, $\chi^2 = 81.568$, $p < 0.001$.

As in all previous cases, additional stepwise discriminant analyses with phonological cues were also performed for each of the frequency groups. Table 6.12

shows a summary of the number of steps which discriminant analyses took for each frequency group as well as the particular phonological cue which was introduced as predictor variable at every step.

Table 6.12: Statistical weight of phonological cues in stepwise discriminant analyses with all frequency groups.

Step	Low-frequency words	Mid-frequency words	High-frequency words
1.	Final voiced consonant + Stressed low vowel (***Phoncomb23***)	Two syllables or more + Stressed low vowel + Trochaic stress pattern (***Phoncomb134***)	Absence of phonological cues (***Phon0***)
2.	Two syllables or more + Final voiced consonant + Stressed low vowel (***Phoncomb123***)	Final voiced consonant + Stressed low vowel (***Phoncomb23***)	Final voiced consonant (***Phon2***)
3.	Two syllables or more + Stressed low vowel + Trochaic stress pattern (***Phoncomb134***)	Absence of phonological feeatures (***Phon0***)	Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)
4.	Two syllables or more + Stressed low vowel + (***Phoncomb13***)	Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	Low stressed vowel (***Phon3***)
5.	--	Low stressed vowel (***Phon3***)	Two syllables or more + Stressed low vowel + Trochaic stress pattern (***Phoncomb134***)
6.	--	Two syllables or more + Final voiced consonant + Stressed low vowel (***Phoncomb123***)	Two syllables or more (***Phon1***)
7.	--	--	All four phonological features (***Phoncomb1234***)
8.	--	--	Two syllables or more + Final voiced consonant + (***Phoncomb12***)

As shown in the table, although the number of steps varies across different frequency ranges, phonological cues contribute in a similar way in the three discriminant functions. Thus, the kind of phonological variables which mostly contributed to each of the discriminant functions in the three stepwise analyses were those variables that contain phonological clusters of two and three features, and also variable *Phon0*, which described absence of phonological features.

The exception to this pattern would be the analysis run with high-frequency words where, besides clusters of two and three features, all variables with only one phonological feature were also included, although they were found non-significant in the corresponding Mann-Whitney U-test. The stepwise discriminant analysis with high-frequency words also took variable *Phoncomb1234* (i.e. the one that contains all four features), unlike the other two analyses, all be it, the variable was only taken at the seventh step.

The low-frequency word analysis was also different from the rest in that it took remarkably less steps than the other two or than the two previous stepwise analyses with types and tokens. However, it should be noted that most of the phonological variables were found to be non-significant in the Mann-Whitney U-test as well.

As to the results obtained, the analysis with low-frequency words revealed a non-successful categorization system, with only an overall 48.1% of words which were correctly classified (16.5% of correctly classified nouns and 90.6% of correctly classified other open class words, Wilks $\lambda = 0.988$, $\chi^2 = 39.802$, $p < 0.001$). Correct classification improved in the analysis of mid-frequency words, with an overall 56.9% of correctly classified words (61.6% of correctly classified nouns and 50.7% of correctly classified other open class words, Wilks $\lambda = 0.980$, $\chi^2 = 83.761$, $p < 0.001$). Finally, among high-frequency words, there was a total of 59.5% of correctly classified words: 83.1% of correctly classified nouns and 34.6% of correctly classified other open class words, Wilks $\lambda = 0.939$, $\chi^2 = 143.400$, $p < 0.001$. Thus, except for the analysis with low-frequency words, the rest of stepwise discriminant analyses gave very similar results to the standard discriminant analyses.

6.3.4. Summary of results with phonological cues

In general terms, the strength of phonological variables was directly proportional to the amount of information that each of the variables carried. Thus, those variables that contained clusters of three or four phonological features (i.e. *Phoncomb123*, *Phoncomb124*, *Phoncomb134* and *Phoncomb1234*) were stronger than those which contained just a single phonological feature (i.e. *Phon1*, *Phon2* and *Phon3*). Evidence for this pattern was found in the tests of significance as well as in the stepwise discriminant analyses, where variables of a single feature were often discarded.

As to the success with which phonological variables classified words in their appropriate grammatical category, there was a tendency towards very high accuracy scores and very low completeness scores when frequency was not taken into account (i.e. in the type analysis or in the low-frequency word analysis). However, the pattern was progressively reversed as frequency increased. Thus, with mid-frequency words, there was a levelling effect, with almost as many nouns as other open class words which were correctly classified, each group reaching scores which were slightly above chance. The pattern is ultimately reversed as word frequency reaches its top, with very high completeness scores but very low accuracy scores both in the analysis with all tokens and the analysis with high-frequency words.

6.4. Semantic cues

6.4.1. Descriptive data

For the analysis of semantic cues, the set of nouns and the set of other open class words obtained from the corpus preparation were tested against the four different semantic cues described in chapter 5 (see section 5.4.3). The total number of nouns that met each of the semantic descriptions under consideration is shown in table 6.13(a), while table 6.13(b) shows the results obtained from the equivalent analysis with the rest of open class words. As with the rest of data presented in this dissertation, tables 6.13(a) and 6.13(b) show the total number of types and tokens from the corpora of all four children together. Individual tables with data corresponding to the corpus of each individual child are found in Appendix E.

Table 6.13.(a): Total number of nouns in each semantic category.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	699	9,516	0.073	0.13	0.18
Basic level count nouns (*Sem2*)	2,675	25,369	0.105	0.50	0.49
Basic level mass nouns (*Sem3*)	288	2,655	0.108	0.05	0.05
Action words (*Sem4a*)	228	1,419	0.161	0.04	0.03
Non-basic level words (*Sem4b*)	1,507	12,618	0.119	0.28	0.24

Table 6.13.(b): Total number of other open class words in each semantic category.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	0	0	--	0.00	0.00
Basic level count nouns (*Sem2*)	0	0	--	0.00	0.00
Basic level mass nouns (*Sem3*)	0	0	--	0.00	0.00
Action words (*Sem4a*)	2,520	46,530	0.054	0.60	0.53
Non-basic level words (*Sem4b*)	59	1,318	0.045	0.01	0.01

In the case of semantic cues, nouns and other open class words only overlapped as far as cue *Sem4* is concerned (i.e. the one that described XBLOC nouns, that is, those nouns which fell out of the basic-level object category). Crucially, the XBLOC included all nouns which had a semantic content denoting "actions", which makes them overlap with most verbs. Thus, the values in variable *Sem4a* in tables 6.13 indicate the number of nominal elements that denote actions in the case of nouns (e.g. *a kiss*, *a call*, *some help*) and the number of adjectives and verbs that had a similar semantic content in the case of other open class words. As seen from the tables, the number of other open class words that denote actions exceed by far the number of nouns that have a similar semantic content. Thus, given these data, it looks like very few nouns would be misclassified as verbs on the basis of the semantic information they carry. The following section will examine the statistical validity of such a claim.

Variable *Sem4b* included all other XBLOC nouns which were not actions (i.e. generic nouns, locations, person roles, etc. See section 5.4.3. for a full list). Crucially, the XBLOC group also included such semantic notions as "material", so nouns that matched this semantic description were included in this group. Furthermore, some of the other open class words shared a similar semantic content, since all adjectives that indicated material or colour were included in this group as well. The following section presents the results obtained from the significance tests using semantic variables.

6.4.2. Tests of significance

6.4.2.1. Types

In line with the tests of significance that were performed using distributional and phonological cues, 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 children together. The results of the tests are found in table 6.14.

Table 6.14.: Mann-Whitney U-test for the semantic cues with all types.

Semantic cues	Nouns	Other	Z	Significance
Proper nouns (*Sem1*)	0.13	0.00	-24.334	0.000
Basic level count nouns (*Sem2*)	0.50	0.00	-53.978	0.000
Basic level mass nouns (*Sem3*)	0.05	0.00	-15.272	0.000
Action words (*Sem4a*)	0.04	0.59	-59.566	0.000
Non-basic level words (*Sem4b*)	0.28	0.01	-34.954	0.000

As seen before, in the case of semantic variables, nouns and other open class words only overlapped as far as variable *Sem4(a-b)* is concerned, that is, in the fact that some nouns might contain semantic features which are typically associated to nouns (i.e. denoting actions) or that some adjectives might contain semantic features which are typically associated to nouns (i.e. denoting materials or colours). However, even if descriptive data showed that there were indeed a number of nouns which denoted actions and a number of adjectives which were semantically similar to nouns, the differences between nouns and other open class words, as far as semantic information is concerned, is still highly significant, as shown in the Mann-Whitney U-test. The mean distributions also indicate that other open class words are more likely to refer to actions than nouns, while nouns are more likely to refer to materials and colours than any other open class word.

Furthermore, with semantic variables, a further significance test was performed using data from the noun group only. This was carried out in order to test the degree to which nouns bearing semantic features that would help children in categorization (i.e. *Sem1*, *Sem2* and *Sem3* nouns) differed significantly from those which would not (i.e. *Sem4a* and *Sem4b* nouns). For this purpose, a Wilcoxon signed-rank test was performed. The results from this test using types revealed a highly significant difference between *Sem1*, *Sem2* and *Sem3* nouns on the one hand *versus Sem4* nouns on the other, Z = -17.954, $p < 0.001$. There was an advantage of semantically categorizable nouns (i.e. mean 0.55) over semantically non-categorizable nouns (i.e. mean 0.32).

6.4.2.2. Tokens

The same significance test carried out for types using semantic cues was also performed on the 51,577 noun tokens and the 88,047 other open class word tokens from the corpora of the four children. The results of the Mann-Whitney U-tests are shown in table 6.15.

Table 6.15.: Mann-Whitney U-test for the semantic cues with all tokens.

Semantic cues	Nouns	Other	Z	Significance
Proper nouns (*Sem1*)	1.77	0.00	-24.312	0.000
Basic level count nouns (*Sem2*)	4.71	0.00	-53.066	0.000
Basic level mass nouns (*Sem3*)	0.49	0.00	-15.269	0.000
Action words (*Sem4a*)	0.26	11.02	-58.842	0.000
Non-basic level words (*Sem4b*)	2.35	0.28	-34.647	0.000

As with types, the significance test with tokens revealed highly significant differences between nouns and other open class words as far as semantic cues are concerned. The mean distributions of both groups also indicate that, while there do exist some nominal tokens with semantic information typically associated to verbs (i.e. *Sem4a* nouns) and there are other open class word tokens with semantic information associated to nouns (i.e. *Sem4b* other open class words), the means obtained do not alter the highly significant differences between the grammatical categories concerned.

Following the analyses used with types as well, a Wilcoxon signed-rank test was performed to see whether *Sem1*, *Sem2* and *Sem3* noun tokens differed significantly from *Sem4* noun tokens. The results from this test revealed a highly significant difference between *Sem1*, *Sem2* and *Sem3* nouns on the one hand *versus Sem4* nouns on the other, Z = -19.399, $p < 0.001$. There was an advantage of semantically categorizable nouns (i.e. mean 5.25) over semantically non-categorizable nouns (i.e. mean 2.61).

6.4.2.3. Frequency groups

As in previous analyses, I also tested whether the significance values obtained from the analysis with all tokens were affected by the split of tokens in three frequency groups. Thus, new Mann-Whitney U-tests were performed with each of the frequency groups (i.e. low-frequency tokens, mid-frequency tokens and high-frequency tokens). Tables 6.16(a-c) show the results obtained each test.

Table 6.16(a): Mann-Whitney U-test for the semantic cues with low-frequency words.

Semantic cues	Nouns	Other	Z	Significance
Proper nouns (*Sem1*)	0.14	0.00	-14.842	0.000
Basic level count nouns (*Sem2*)	0.42	0.00	-27.845	0.000
Basic level mass nouns (*Sem3*)	0.05	0.00	-8.524	0.000
Action words (*Sem4a*)	0.05	0.59	-34.279	0.000
Non-basic level words (*Sem4b*)	0.34	0.01	-23.143	0.000

Table 6.16(b): Mann-Whitney U-test for the semantic cues with mid-frequency words.

Semantic cues	Nouns	Other	Z	Significance
Proper nouns (*Sem1*)	0.51	0.00	-15.985	0.000
Basic level count nouns (*Sem2*)	2.12	0.00	-35.464	0.000
Basic level mass nouns (*Sem3*)	0.20	0.00	-9.521	0.000
Action words (*Sem4a*)	0.16	2.48	-39.724	0.000
Non-basic level words (*Sem4b*)	0.92	0.04	-20.547	0.000

Table 6.16(c): Mann-Whitney U-test for the semantic cues with high-frequency words.

Semantic cues	Nouns	Other	Z	Significance
Proper nouns (*Sem1*)	6.90	0.00	-10.351	0.000
Basic level count nouns (*Sem2*)	16.79	0.00	-28.826	0.000
Basic level mass nouns (*Sem3*)	1.79	0.00	-8.503	0.000
Action words (*Sem4a*)	0.82	37.61	-27.303	0.000
Non-basic level words (*Sem4b*)	8.42	0.99	-16.035	0.000

As shown in the tables, the results from the tests with each of the frequency groups were identical to the former analyses with types and tokens as far as significance is concerned. Mean distributions also show the same pattern as previous analyses, with verbs being more likely to be associated to the semantic feature of "action" and nouns being more likely to be associated to the XBLOC semantic category as a whole.

As in the analyses used with types and tokens, a Wilcoxon signed-rank test was also performed with the nouns in each of the frequency groups in order to see whether *Sem1*, *Sem2* and *Sem3* noun tokens differed significantly from *Sem4* noun tokens. The results from this test for low-frequency words revealed a highly significant difference between *Sem1*, *Sem2* and *Sem3* nouns on the one hand *versus Sem4* nouns on the other, Z = -4.061, $p < 0.001$, with an advantage of semantically categorizable nouns (i.e. mean 0.47) over semantically non-categorizable nouns (i.e. mean 0.38). With mid-frequency words, a highly significant difference was found as well between both nominal groups, Z = -5.223, $p < 0.001$, but this time there was an advantage of semantically non-categorizable nouns (i.e. mean 1.67) over categorizable nouns (i.e. mean 1.34). Finally, the results from this test for high-frequency words revealed a highly significant difference again between *Sem1*, *Sem2* and *Sem3* nouns *versus Sem4* nouns, Z = -10.197, $p < 0.001$, with an advantage of semantically categorizable nouns (i.e. mean 18.58) over semantically non-categorizable nouns (i.e. mean 9.24).

As seen in previous chapters, the goal of the semantic analysis was not only to work out the degree with which nouns would overlap with other open class words, especially verbs, as far as their meaning was concerned. Besides, the idea was to test the likelihood with which the semantic features which have traditionally been associated to nouns alone (which are captured here as *Sem1*, *Sem2* or *Sem3*) could account for the correct classification· of most, if not all nouns to which English-learning children are exposed. For this purpose, variables *Sem4a* and *Sem4b* were united into a single *Sem4* variable, which would contain all nouns which lack their prototypical semantic content (i.e. absence of *Sem1*, *Sem2* or *Sem3* features), plus all other open class words which equally carry semantic information which is likely to be associated to the XBLOC semantic category. The results from the tests of diagnosticity using these four resulting variables (i.e. *Sem1-4*) are reported in the next section.

6.4.3. Tests of diagnosticity

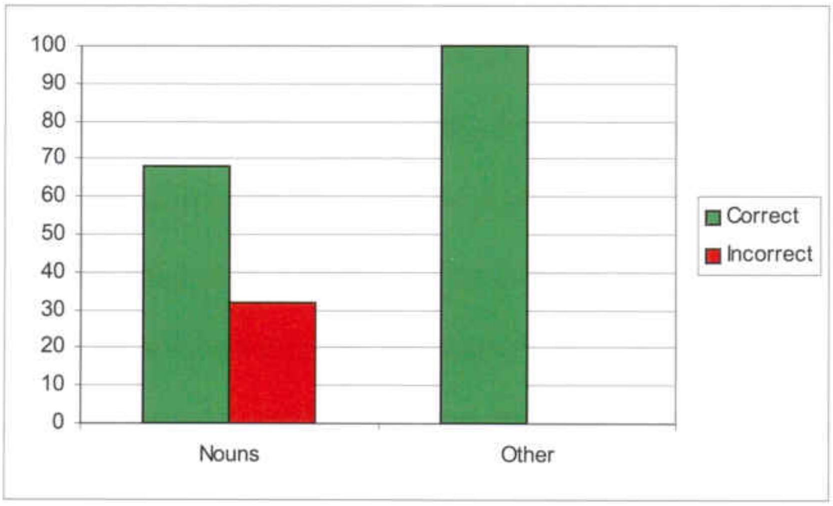
6.4.3.1. Types

As in previous analyses with distributional and phonological cues, correct classification of all types from the four corpora was assessed for the four semantic cues. When the cues were entered simultaneously, 67.9% of nouns were correctly classified. As far as other open class words are concerned, correct classification reached 100% of the cases. This is not surprising, as such high accuracy scores only indicate that there were no other open class words which interfered in the noun category by carrying any of the semantic features which are typically associated to nouns (i.e. there are no verbs, adjectives or adverbs which denote "proper noun" or "common count basic level object" or "substance", which are variables *Sem1*, *Sem2* and *Sem3*, respectively).

However, the lower completeness scores indicate that, statistically speaking, there do exist a number of nouns which would fall out of the nominal category on the basis of the semantic information that they carry. Thus, 67.9% of nouns have a semantic content which would help children identify them as nouns (i.e. they were included in variable *Sem1*, *Sem2* or *Sem3*), while the rest of nouns lack this semantic content (i.e. they were included in variable *Sem4*). Overall correct classification was 82.0%, Wilks $\lambda = 0.439$, $\chi^2 = 7916.359$, $p < 0.001$. However, the percentage of overall correct classification in the case of semantic cues is biased by the high score obtained in accuracy, which was obvious in this case. Results from this analysis are illustrated in figure 6.14.

As in previous analyses, the cross-validations using the leave-one-out method and the one using 50% of randomly selected cases were performed here as well. The results from the discriminant analysis using the leave-one-out method were identical to those obtained using the standard method. As for the cross-validation using 50% of randomly selected cases, results in accuracy were obviously the same as before (i.e. 100% of correctly classified other open class words), and there was a slight improvement in completeness scores, with 69.1% of correctly classified nouns, Wilks $\lambda = 0.450$, $\chi^2 = 3813.063$, $p < 0.001$.

Figure 6.14.: Classification of noun types and other types for all semantic variables together.



As in the previous analyses as well, a further stepwise discriminant analysis was carried out with all types using semantic variables. The stepwise discriminant analysis took all four semantic variables and did not discard any of them. When the four variables were introduced using the stepwise method, they were entered in the following order:

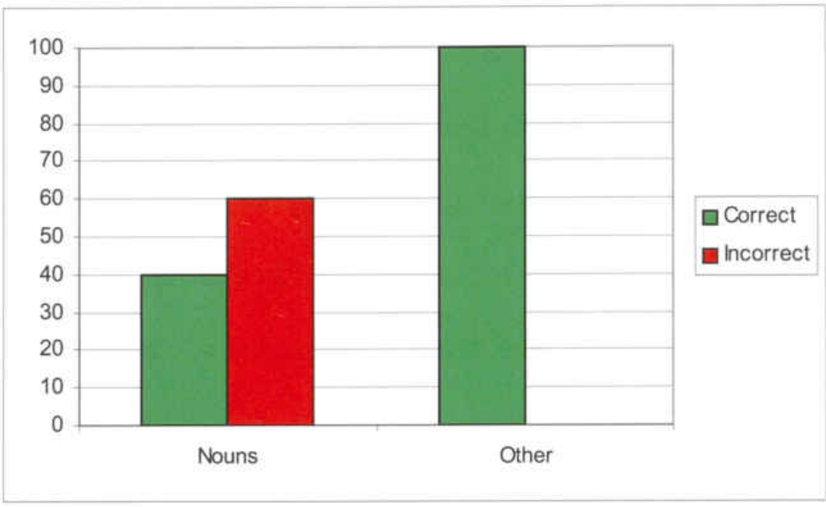
1. Common count nouns (*Sem2*)
2. Proper nouns (*Sem1*)
3. Common mass nouns (*Sem3*)
4. XBLOC words (*Sem4*)

The results obtained from this new stepwise discriminant analysis were also identical to those obtained from the standard discriminant analysis. Thus, the idea that basic level objects (i.e. *Sem2* nouns) do not account for the whole inventory of English types to which children are exposed is confirmed by the stepwise analysis as well. However, despite not getting perfect completeness scores, the variable that described the most prototypical denotational meaning of nouns is still statistically stronger than the rest, as it is the one which was first introduced in the stepwise analysis.

6.4.3.2. Tokens

The same analyses that were run with types using the four semantic cues were also run with all tokens from the four corpora. For the token analysis, when all four cues were entered simultaneously in the standard discriminant analysis, 39.9% of nouns and 100% of other open class words were correctly classified. Overall correct classification was 66.3% of tokens, Wilks $\lambda = 0.945$, $\chi^2 = 544.696$, $p < 0.001$. The results from the token analysis are illustrated in figure 6.15.

Figure 6.15.: Classification of noun tokens and other tokens for all semantic variables together.



Thus, the accuracy scores obtained from the type analysis are replicated with tokens, indicating that other open class words are never misclassified as nouns (i.e. there are no other open class word tokens with semantic features associated to nouns). However, there were very low completeness scores in the analysis (i.e. only 39.9% of nouns were correctly classified). This indicates that there is a number of noun tokens which lack the semantic features with which nouns are associated, and whose semantic information is either ambiguous or too broad (i.e. they were grouped in variable *Sem4*). This means that children would not be able to not work out the grammatical category to which these nouns belong on the basis of semantic information alone.

These results were cross-validated using the same methods as in all the previous analyses. In the cross-validation discriminant analysis with tokens using the leave-one-out method, the results obtained were identical to those from the standard discriminant analysis. In the further discriminant analysis using 50% of randomly selected cases, identical results were found for accuracy (i.e. 100% of other open class words were correctly classified) but completeness scores dropped, with only 33.8% of correctly classified nouns, Wilks $\lambda = 0.938$, $\chi^2 = 305.836$, $p < 0.001$.

When semantic cues were not introduced simultaneously, but using the stepwise method, all four variables were also found to contribute significantly to the discriminant function in the analysis, as was seen with types as well. In the stepwise discriminant analysis with tokens, the order in which each of the semantic variables were introduced was the following:

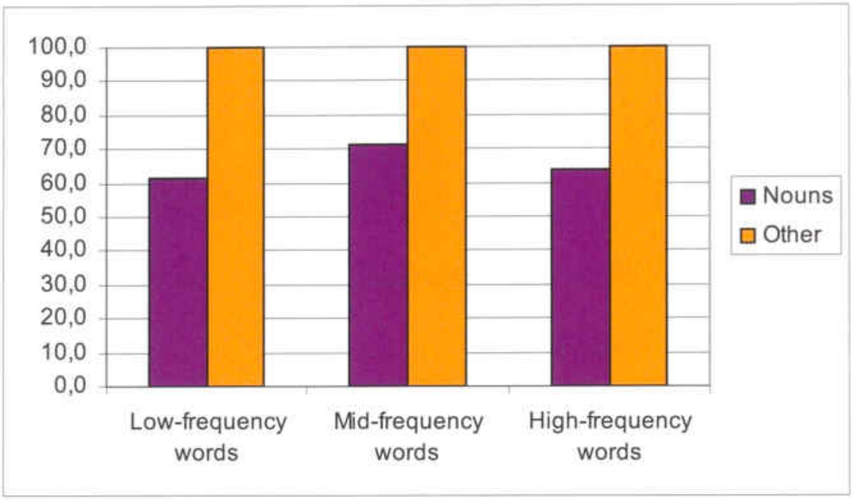
1. Common count nouns (*Sem2*)
2. XBLOC words (*Sem4*)
3. Common mass nouns (*Sem3*)
4. Proper nouns (*Sem1*)

Thus, although the order in which variables were introduced was different in the type and token analysis, they both have in common the fact that variable *Sem2* (i.e. the one that contains nouns with their prototypical semantic features) is still statistically the strongest, as it is the one which is introduced first in both stepwise discriminant analyses. The results obtained from this stepwise discriminant analysis with tokens were also identical to the results found from the equivalent analysis using the standard method and introducing all four variables simultaneously.

6.4.3.3. Frequency groups

As in all the previous analyses, in order to test whether word frequency had an effect on correct grammatical categorization using semantic variables, a discriminant analysis was also carried with the four semantic cues together for each of the three frequency groups. Figure 6.16 shows the percentages of correct classification only for both nouns and other open class words in each of the frequency groups.

Figure 6.16: Percentage of correct classification of noun tokens and other tokens for all semantic variables across different frequency groups.



In the analysis with low-frequency words, a total of 61.6% of nouns and 100% of other open class words were correctly classified. Overall correct classification was 77.9%, Wilks λ = 0.485, χ^2 = 2373.876, $p < 0.001$. The same analysis with mid-frequency words gave a total of 71.5% of nouns and 100% of other open class words which were correctly classified, with an overall 83.6% of tokens which were correctly classified, Wilks λ = 0.647, χ^2 = 1769.864, $p < 0.001$. Last, the same analysis with high-frequency words resulted in a total of 63.8% of nouns and 100% of other open class words correctly classified, with an overall of 81.4% of correctly classified tokens, Wilks λ = 0.824, χ^2 = 438.948, $p < 0.001$.

When using the discriminant analysis with the leave-one-out method for cross validation purposes, the results obtained in each of the analysis for each frequency group were identical to those obtained from the standard analyses. As for the second cross-validation using 50% of randomly selected cases, completeness scores improved among low-frequency words, with 62.2% of correctly classified nouns, Wilks $\lambda = 0.498$, $\chi^2 = 1145.469$, $p < 0.001$. A similar pattern emerged out of the cross-validation with mid-frequency words, where completeness scores improved giving a total of 72.9% of correctly classified nouns, Wilks $\lambda = 0.655$, $\chi^2 = 842.760$, $p < 0.001$. Last, the same cross-validation analysis with high-frequency words made completeness scores improve up to a 66.6% of correctly classified nouns, Wilks $\lambda = 0.795$, $\chi^2 = 260.563$, $p < 0.001$.

Following previous analyses, additional stepwise discriminant analyses with semantic cues were also performed for each of the frequency groups.Table 6.17 shows a summary of the number of steps which discriminant analyses took for each frequency group as well as the particular semantic cue which was introduced as predictor variable at every step.

Table 6.17: Statistical weight of semantic cues in stepwise discriminant analyses with all frequency groups.

Step	Low-frequency words	Mid-frequency words	High-frequency words
1.	Common count nouns (***Sem2***)	Common count nouns (***Sem2***)	Common count nouns (***Sem2***)
2.	Proper nouns (***Sem1***)	Proper nouns (***Sem1***)	Common mass nouns (***Sem3***)
3.	XBLOC words (***Sem4***)	Common mass nouns (***Sem3***)	XBLOC words (***Sem4***)
4.	Common mass nouns (***Sem3***)	--	Proper nouns (***Sem1***)

As shown in the table, the steps taken in each of the discriminant analyses are very similar across different frequency groups, and they are all very similar to the stepwise analyses with types and tokens as well. Thus, although the semantic notion that nouns refer to objects does not account for all English nouns to which children are exposed, the variable in question (i.e. *Sem2*) is still the one with more statistical weight in all stepwise analyses that have been performed, as it is always the one that is introduced first and the one which most contributes to each of the discriminant functions.

As to the outcome of each of the analyses, the results from the three stepwise analyses were identical, in all cases, to the previous analyses performed with each frequency group using the standard discriminant analysis method and introducing all variables together at the same time. Surprisingly, results were identical in all analyses, including the analyses performed with mid-frequency words, which took a step less than the rest of words in the stepwise analysis. This indicates that variable *Sem4* (i.e. the one which was discarded from the stepwise analysis with mid-frequency words) had absolutely no contribution to the discriminant function or to word categorization.

6.4.4. Summary of results with semantic cues

Significance tests with semantic variables revealed a highly significant difference between nouns and other open class words in all cases as far as each of the four semantic variables are concerned. Significance values were more or less the same for all variables, but from all of them, stepwise discriminant analyses revealed that it was variable *Sem2* (i.e. basic level object common nouns) the one which is statistically stronger than the rest.

As far as categorization force is concerned, discriminant analysis revealed very good accuracy scores for semantic cues in all cases, indicating that other open class words are very unlikely to be misclassified as nouns on the basis of their semantic content. However, completeness scores were far lower, even below chance in the analysis with all tokens together. As fas as frequency groups were concerned, completeness scores were relatively low among low-frequency words, they improved a bit among mid-frequency words, but dropped again among high-frequency words. This indicates that absence of prototypical semantic features in many of the nouns might lead to them being miscategorized outside the noun group on the basis of semantic information.

6.5. Cues in interaction

The previous sections have presented the results of correct classification of all words based on the discriminant analyses of distributional, phonological and semantic cues entered separately for types, tokens and each of the frequency groups. Interestingly, correct classification has been observed to interact with word frequency. Thus, for high-frequency words, distributional and phonological cues appear to be better than for low-frequency words. The reverse pattern seems to be true of semantic cues, for which overall correct classification is better for mid-frequency words, and slightly decreases among high-frequency words. The difference between semantic cues and the rest of cues are bigger when only completeness scores are considered. Semantic cues show such high overall correct classification scores because accuracy scores were very high, but completeness scores were far lower (see section 6.4.3). Thus, if different types of cues contribute differently towards accurate classification of English nouns, combining several different cues together might improve the overall categorization power of all cues in general. The following sections provide the results obtained from discriminant analyses after combining different types of cues.

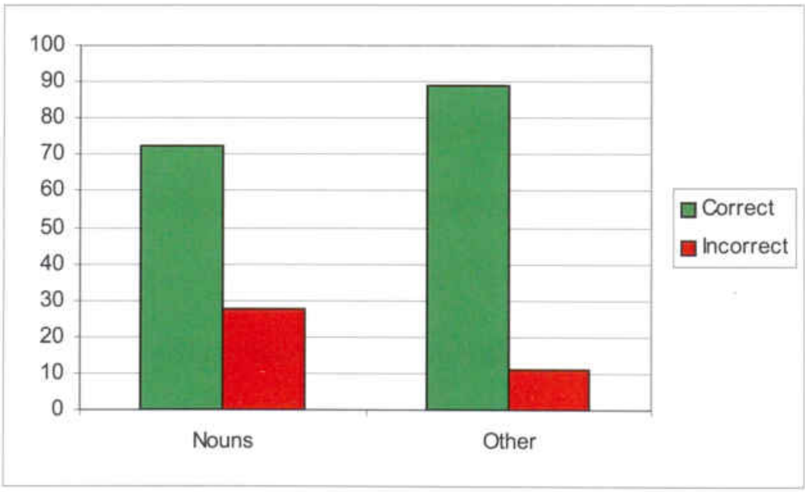
For these analyses of the interaction of all different kinds of cues, the variables that were used were exactly the same as those that were used in the analyses with individual cues. Therefore, for cues in interaction, only tests of diagnosticity will be reported here, given the fact that tests of significance would be the same as those which have already been reported for the separate analyses of distributional, phonological and semantic cues.

6.5.1. Distributional and phonologial cues

6.5.1.1. Types

A standard discriminant analysis with types was performed using all fifteen distributional cues (i.e. *Syn0a-Syn7b*) as well as the twelve phonological cues together (i.e. *Phon0-Phoncomb1234*). This yielded a total of 27 cues used as predictor variables. When all these cues were entered simultaneously, there was a total of 72.1% of correctly classified noun types and 88.9% of correctly classified other open class word types. Overall correct classification reached a total of 79.5% of types, Wilks $\lambda = 0.599$, $\chi^2 = 4926.809$, $p < 0.001$. Figure 6.17 illustrates the results obtained from this analysis.

Figure 6.17.: Classification of noun types and other types for distributional and phonological variables combined.



When cross-validation analyses were performed, a discriminant analysis using the leave-one-out method gave identical results to the ones obtained from the standard discriminant analysis. When a second cross-validation was performed using 50% of randomly selected cases, identical results were found for accuracy scores (i.e. 88.9% of correctly classified other open class words) but there was a slight drop in accuracy, with 71.5% of correctly classified nouns, Wilks $\lambda = 0.590$, $\chi^2 = 2518.206$, $p < 0.001$.

When all distributional and phonological cues were not entered simultaneously, but were entered one at each step, the stepwise discriminant analysis took a total of 19 out of the 27 possible, and discarded 8 of them. The selected 19 cues were entered in the following order:

1. "{∅} + x" (*Syn0a*)
2. "{*the*} + x" (*Syn2a*)
3. "{POSSESSIVE} + x" (*Syn4a*)
4. "{∅} + x + *-(e)s*" (*Syn0b*)
5. "{*a, an*} + x" (*Syn1*)
6. "{*the*} + x + *-(e)s*" (*Syn2b*)
7. "Two syllables or more + final voiced consonant + low stressed vowel" (*Phoncomb123*)
8. "Two syllables or more + low stressed vowel + trochaic stress pattern" (*Phoncomb134*)
9. "Two syllables or more + low stressed vowel" (*Phoncomb13*)
10. "{PREPOSITION} + x" (*Syn6a*)
11. "{QUANTIFIER} + x + *-(e)s*" (*Syn5b*)
12. "{WH- ELEMENT} + x" (*Syn7a*)
13. "Two syllables or more" (*Phon1*)
14. "Final voiced consonant + low stressed vowel" (*Phoncomb23*)

15. "Two syllables or more + trochaic stress pattern" (*Phoncomb14*)

16. "Presence of all four phonological cues" (*Phoncomb1234*)

17. "{*this*, *that*, *these*, *those*} + x" (*Syn3a*)

18. "Two syllables or more + final voiced consonant" (*Phoncomb12*)

19. "{POSSESSIVE} + x + *-(e)s*" (*Syn4b*)

From the 19 cues that most significantly contributed to the discriminant function for types, 11 cues were distributional, while 8 cues were phonological. Thus, when distributional and phonological cues are combined and interact with one another, distributional cues seem to be stronger and win over phonological cues, although many of the phonological cues equally contributed to the discriminant function as well. This claim is based on the fact that, on the one hand, there were proportionally more distributional cues than phonological cues in the stepwise analysis and, on the other hand, distributional cues seem to be stronger, as they are the ones that were introduced first in the stepwise analysis (i.e. the first phonological cue found to contribute to the discriminant function was only entered at the seventh step).

Among the selected distributional cues, most of them corresponded to bigram contexts. Among the selected phonological variables, most of them corresponded to clusters of two or three features. Thus, the discarded cues were mainly distributional biframe contexts and single-feature phonological variables. This is consistent with the data obtained from the analyses with each cue type separately.

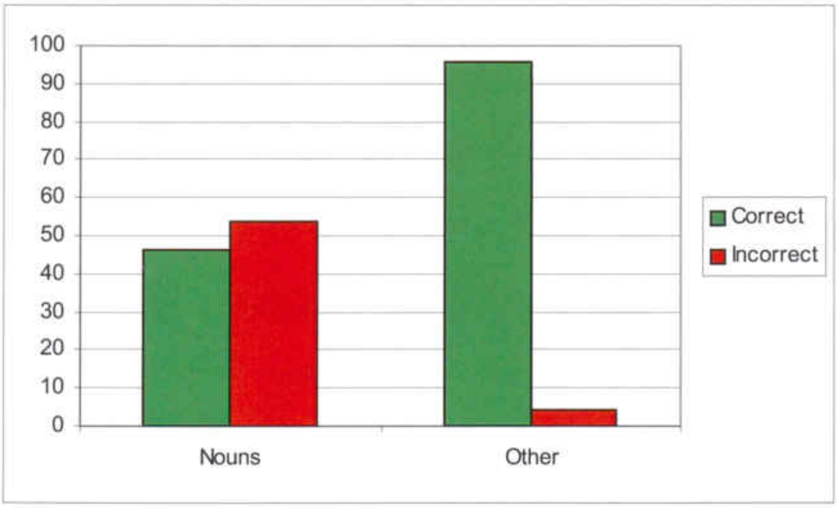
As to the results obtained from the stepwise discriminant analysis, they were almost identical to the ones found with the standard discriminant analysis when all the variables were entered simultaneously. Correct classification reached 72.1% for nouns and 88.9% for other open class words, Wilks $\lambda = 0.599$, $\chi^2 = 4923.146$, $p < 0.001$.

6.5.1.2. Tokens

The analysis that was carried out with types using distributional and phonological variables together was also performed with tokens using the same variable combination. When the 27 cues were entered together in a standard discriminant analysis, there was a total of 46.4% of correctly classified nouns and 95.6% of correctly classified other open class words. Overall correct classification was 68.0% of tokens, Wilks $\lambda = 0.913$, $\chi^2 = 873.343$, $p < 0.001$. Results from this analysis are illustrated in figure 6.18.

When these results were cross-validated using the leave-one-out method, results were almost identical to the ones from the standard analysis, with only minimally lower results both for accuracy and completeness. Thus, in the leave-one-out discriminant analysis, there were 46.2% of correctly classified nouns and 95.5% of correctly classified other open class words, with an overall correct classification of 67.9% of tokens, Wilks $\lambda = 0.913$, $\chi^2 = 873.343$, $p < 0.001$.

Figure 6.18.: Classification of noun tokens and other tokens for distributional and phonological variables combined.



In the cross-validation using 50% of cases selected at random, correct classification rates improved, especially for completeness scores, with 53.2% of correctly classified nouns. Accuracy scores were very similar, with 94.1% of correctly classified other open class words. Overall correct classification in this second cross-validation was 71.3% of tokens, Wilks $\lambda = 0.905$, $\chi^2 = 475.033$, $p < 0.001$.

The stepwise discriminant analysis with tokens took one step less than the equivalent analysis with types. Thus, out of the 27 possible distributional and phonological variables in combination, a total of 18 variables were introduced this time. The exact variables that were chosen and the order in which they were entered was the following:

1. "{*the*} + x" (*Syn2a*)
2. "{∅} + x" (*Syn0a*)
3. "{QUANTIFIER} + x + *-(e)s*" (*Syn5b*)
4. "{POSSESSIVE} + x" (*Syn4a*)
5. "{∅} + x + *-(e)s*" (*Syn0b*)
6. "{*the*} + x + *-(e)s*" (*Syn2b*)
7. "Final voiced consonant + low stressed vowel" (*Phoncomb23*)
8. "{POSSESSIVE} + x + *-(e)s*" (*Syn4b*)
9. "Two syllables or more + low stressed vowel + trochaic stress pattern" (*Phoncomb134*)
10. "Two syllables or more + final voiced consonant + low stressed vowel" (*Phoncomb123*)
11. "{*a, an*} + x" (*Syn1*)
12. "Final voiced consonant" (*Phon2*)

13. "{PREPOSITION} + x" (*Syn6a*)

14. "{*this, that, these, those*} + x" (*Syn3a*)

15. "Two syllables or more" (*Phon1*)

16. "Two syllables or more + final voiced consonant" (*Phoncomb12*)

17. "{PREPOSITION} + x + *-(e)s*" (*Syn6b*)

18. "{PREPOSITION} + x + *-(e)s*" (*Syn3b*)

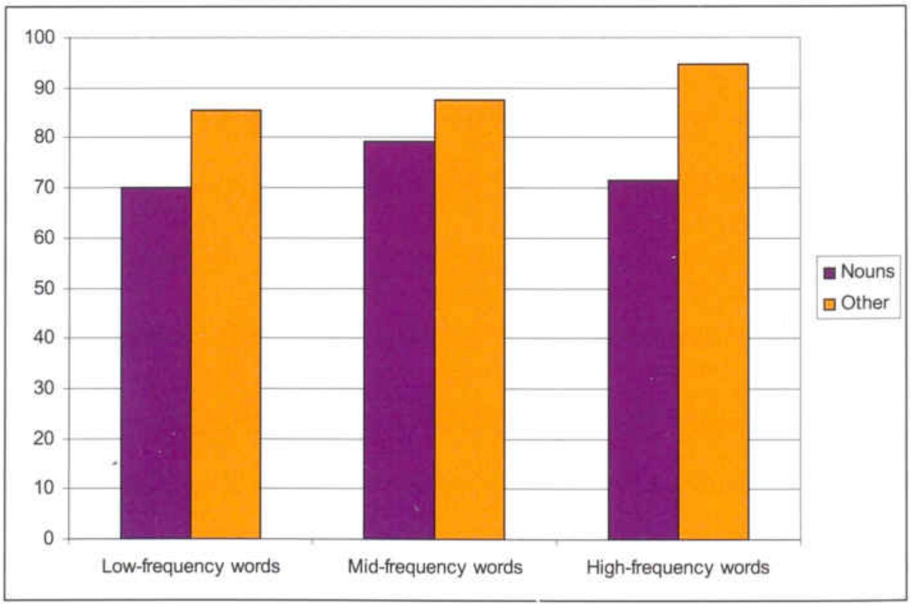
For the token analysis, from the 18 variables that were introduced, there were 12 distributional variables and 6 phonological variables. Therefore, the pattern obtained in the type analysis is replicated here in that, even if phonological cues interact with distributional cues, it is distributional cues that appear to be stronger than phonological cues. This can be seen by the fact that the proportion of distributional variables is higher than that of phonological variables in the stepwise analysis, and also because distributional variables occupy the first steps, while the strongest phonological variable is only introduced at the seventh step.

As to the results obtained from this stepwise discriminant analysis, classification improved slightly, compared to the results obtained from the standard discriminant analysis. Thus, the stepwise discriminant analysis gave a total of 47.3% of correctly classified nouns and 95.5% of correctly classified other open class words, with overall 68.5% of correctly classified tokens, Wilks $\lambda = 0.914$, $\chi^2 = 867.399$, $p < 0.001$. This indicates that the classificatory system actually worked better with the selected variables rather than with all 27 variables at once and that, therefore, the 9 variables that were discarded, not only did not contribute significantly to the discriminant function, but they also added confusion and fuzziness between the grammatical groups to be classified and impeded successful classification to a certain extent.

6.5.1.3. Frequency groups

The interaction between distributional and phonological cues was also examined with each of the frequency groups. Results from this analysis are illustrated in figure 6.19. Among low-frequency words, there were 70.1% of correctly classified nouns and 85.6% of correctly classified other open class words. There was an overall correct classification of 76.7% of words, Wilks $\lambda = 0.657$, $\chi^2 = 1373.980$, $p < 0.001$. Among mid-frequency words, there were 79.3% of correctly classified nouns and 87.4% of correctly classified other open class words, with an overall correct classification of 82.8%, Wilks $\lambda = 0.594$, $\chi^2 = 2110.940$, $p < 0.001$. Among high-frequency words, there was a total of 71.6% of correctly classified nouns and 94.7% of correctly classified other open class words, with an overall correct classification of 83.4%, Wilks $\lambda = 0.697$, $\chi^2 = 775.353$, $p < 0.001$.

Figure 6.19: Percentage of correct classification of noun tokens and other tokens for distributional and phonological variables across different frequency groups.



When these results were cross-validated using the leave-one-out method, results were very similar to the ones found with the standard method when all cases were entered. For low-frequency words, results were identical in both analyses. For mid- and high-frequency words, overall correct classification scores were slightly lower. For mid-frequency words, the cross-validation analysis gave a total of 79.1% of correctly classified nouns and 87.3% of correctly classified other open class words, with an overall correct classification of 82.6%. For high-frequency words, correct classification reached a total of 71.0% of nouns and 94.5% of other open class words, with an overall correct classification of 83.0%.

As to the results found in the cross-validation analyses using 50% of randomly selected cases, correct classification in general improved in the three frequency groups, although the degree to which they improved varied in each group. Thus, for low-frequency words, there was a slight improvement both in accuracy and completeness, with 71.2% of correctly classified nouns and 85.3% of correctly classified other open class words. Overall correct classification in this group was 77.4%, Wilks $\lambda = 0.668$, $\chi^2 = 655.819$, $p < 0.001$. Among mid-frequency words, both accuracy and completeness scores improved as well, with 79.9% of correctly classified nouns and 88.5% of correctly classified other open class words. Overall correct classification among mid-frequency words was 83.5%, Wilks $\lambda = 0.594$, $\chi^2 = 1029.965$, $p < 0.001$. Among high-frequency words, there was a very small improvement in completeness scores, but not in accuracy scores, although accuracy scores were already the highest from the three frequency groups. Thus, correctly classified high-frequency nouns reached a total of 71.9%, while correctly classified other open class words reached a total of 94.2%. Overall correct classification was almost identical to what was obtained from the standard discriminant analysis, with 83.5% of correctly classified words, Wilks λ = 0.682, $\chi^2 = 417.827, p < 0.001$.

A stepwise discriminant analysis was also performed using distributional and phonological cues in interaction with each of the frequency groups. Table 6.18 shows a summary of the number of steps that discriminant analyses took with each of the frequency groups, as well as the kind of variable that whas introduced at each step.

As far as the strength of variables is concerned, the stepwise analyses with frequency groups replicate the results obtained from the type and the token analyses. Therefore, it is distributional cues the ones that are introduced at the first steps and, therefore, the ones that are found to contribute most to the discriminant function and the ones which are statistically the strongest.

However, the interaction between distributional and phonological variables varies across different frequency groups. Thus, for the task of low-frequency word categorization, there are nearly as many distributional cues as phonological ones (i.e. from the 12 variables that were introduced, there were 7 distributional and 5 phonological). However, for the task of mid- and high-frequency words, distributional cues are clearly dominant, since they are not only introduced first, but they also exceed in number and proportion phonological cues. Thus, for mid-frequency words, 17 variables were used out of which 12 were distributional variables. Similarly, for high-frequency words, there were a total of 16 variables used in the stepwise analysis, out of which 11 were distributional.

Table 6.18: Statistical weight of distributional and phonological cues combined in stepwise discriminant analyses with all frequency groups.

Step	Low-frequency words	Mid-frequency words	High-frequency words
1.	{∅} + x (***Syn0a***)	{∅} + x (***Syn0a***)	{*the*} + x (***Syn2a***)
2.	{DEMONST.} + x (***Syn3a***)	{*the*} + x (***Syn2a***)	{∅} + x (***Syn0a***)
3.	{QUANTIF.} + x (***Syn5a***)	{∅} + x + *-(e)s* (***Syn0b***)	{QUANTIF.} + x + *-(e)s* (***Syn5b***)
4.	{PREP.} + x (***Syn6a***)	{*a, an*} + x (***Syn1***)	{POSS.} + x (***Syn4a***)
5.	{*a, an*} + x (***Syn1***)	{POSS.} + x (***Syn4a***)	{POSS.} + x + *-(e)s* (***Syn4b***)
6.	Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	{*the*} + x + *-(e)s* (***Syn2b***)	{∅} + x + *-(e)s* (***Syn0b***)
7.	Final voiced consonant + Low stressed vowel (***Phoncomb23***)	{PREP.} + x (***Syn6a***)	{*a, an*} + x (***Syn1***)
8.	Two syllables or more + Low stressed vowel (***Phoncomb13***)	{QUANTIF.} + x + *-(e)s* (***Syn5b***)	{*the*} + x + *-(e)s* (***Syn2b***)
9.	{*the*} + x (***Syn2a***)	Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)
10.	{POSS.} + x (***Syn4a***)	{QUANTIF.} + x (***Syn5a***)	Two syllables or more + Trochaic stress pattern (***Phoncomb14***)
11.	Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	{POSSESSIVE} + x + *-(e)s* (***Syn4b***)	Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)
12.	Two syllables or more (***Phon1***)	{WH- ELEMENT} + x (***Syn7a***)	{QUANTIF.} + x (***Syn5a***)
13.	--	{PREP.} + x + *-(e)s* (***Syn6b***)	{DEMONST.} + x (***Syn3a***)
14.	--	Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	Final voiced consonant (***Phon2***)
15.	--	Two syllables or more + Low stressed vowel (***Phoncomb13***)	{PREP.} + x (***Syn6a***)
16.	--	Two syllables or more (***Phon1***)	Two syllables or more (***Phon1***)
17.	--	Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	--

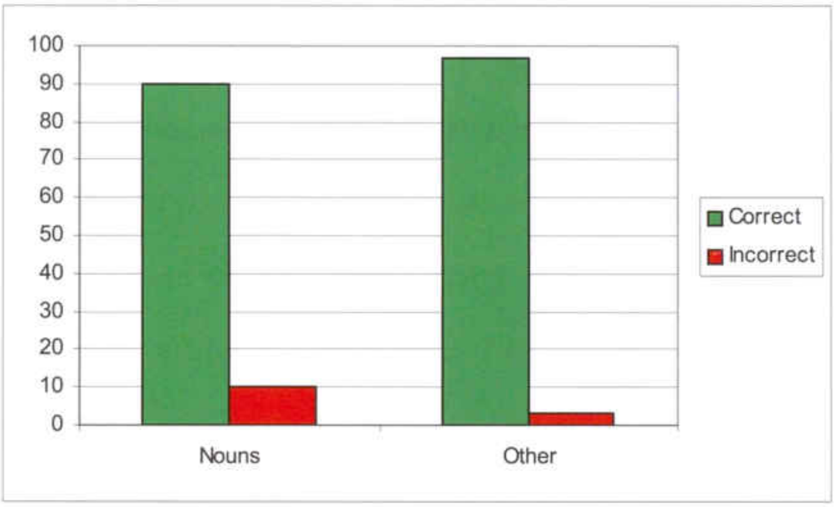
As far as the results from these stepwise analyses are concerned, the results obtained from the low-frequency word analysis are identical to the ones found with the corresponding standard discriminant analysis with the same frequency group, with 70.1% of correctly classified nouns and 85.6% of correctly classified other open class words, Wilks $\lambda = 0.658$, $\chi^2 = 1370.825$, $p < 0.001$. For mid-frequency words, results were very similar as well: there was a total of 78.9% of correctly classified nouns and 87.9% of correctly classified other open class words, Wilks $\lambda = 0.596$, $\chi^2 = 2098.040$, $p < 0.001$. The stepwise analysis with high-frequency words displayed categorization results which are very similar to the ones obtained from the standard discriminant analysis using all 27 variables. With just 16 variables in the stepwise analysis, successful categorization among high-frequency words reached 71.7% in the case of nouns, and 94.9% in the case of other open class words, Wilks $\lambda = 0.699$, $\chi^2 = 770.649$, $p < 0.001$. Therefore, the discarded variables in each of the cases were completely redundant and their presence did not contribute to improve word categorization.

6.5.2. Distributional and semantic cues

6.5.2.1. Types

In order to examine the effects of the interaction between distributional and semantic cues with types, the set of 15 distributional variables and the 4 semantic variables considered in this study were introduced together as predictor variables in a discriminant analysis with all types from the four child corpora. When the total 19 cues were introduced, there were 90.1% of correctly classified noun types and 96.8% of correctly classified other open class word types. Overall correct classification was 93.0% of all types, Wilks $\lambda = 0.312$, $\chi^2 = 11196.554$, $p < 0.001$. The results of this analysis are illustrated in figure 6.20.

Figure 6.20.: Classification of noun types and other types for distributional and semantic variables combined.



In the cross-validation discriminant analysis using the leave-one-out method, results were identical to those obtained in the standard discriminant analysis. A further cross-validation was performed, using 50% of randomly selected cases. For this second cross-validation, results were almost identical as well, with 90.2% of correctly classified nouns and 97.0% of correctly classified other open class words, Wilks $\lambda = 0.315$, $\chi^2 = 5516.626$, $p < 0.001$.

A further stepwise discriminant analysis was performed in order to evaluate the strength of each distributional and semantic cue in interaction with one another. When variables were introduced one at a time, a total of 13 variables were used, out of the 19 possible variables, so there were 6 variables which were discarded. The variables that were selected and the order in which they were introduced are the following:

1. "Common count nouns" (***Sem2***)
2. "Proper nouns" (***Sem1***)

3. "{∅} + x" (***Syn0a***)

4. "Common mass nouns" (***Sem3***)

5. "XBLOC words" (***Sem4***)

6. "{*the*} + x" (***Syn2a***)

7. "{*a, an*} + x" (***Syn1***)

8. "{POSSESSIVE} + x" (***Syn4a***)

9. "{∅} + x + *-(e)s*" (***Syn0b***)

10. "{*the*} + x + *-(e)s*" (***Syn2b***)

11. "{WH- ELEMENT} + x" (***Syn7a***)

12. "{QUANTIFIER} + x + *-(e)s*" (***Syn5b***)

13. "{QUANTIFIER} + x" (***Syn5a***)

From the set of selected variables, the four possible semantic variables were selected, and they were also the ones which were found to contribute most to the discriminant function, as they were introduced at the first steps. From the set of distributional variables that were selected, most of them corresponded to bigram contexts.

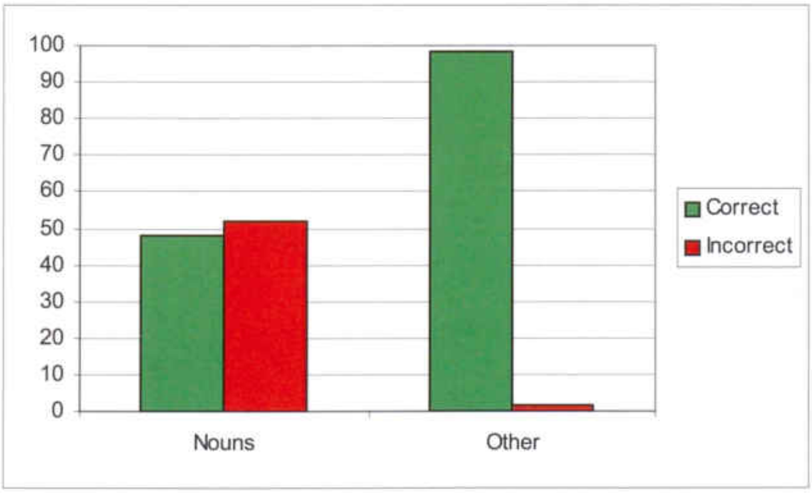
As to the results obtained from the stepwise analysis, results were nearly identical to those found in the corresponding standard analysis where the 19 variables were used. Thus, in the stepwise analysis with just 13 variables, there were 90.0% of correctly classified nouns, and 96.8% of correctly classified other open class words, with an overall correct classification of 93.0% of types, Wilks $\lambda = 0.312$, $\chi^2 = 11189.241$, $p < 0.001$.

6.5.2.2. Tokens

The same classificatory system made up of the combination of distributional and semantic variables was tested against the set of tokens from the four child corpora. Results from this analysis are illustrated in figure 6.21. When the 19 variables were introduced simultaneously as predictor variables in a standard discriminant analysis, there were 47.9% of correctly classified noun tokens and 98.3% of correctly classified other open class word tokens. Overall correct classification reached 70.1% of tokens, Wilks $\lambda = 0.899$, $\chi^2 = 1028.161$, $p < 0.001$.

When these results were cross-validated using the leave-one-out method, the results obtained were identical to those found in the standard discriminant analysis. The cross-validation with a random selection of 50% of the cases gave slightly better results than the standard discriminant analysis, with 48.7% of correctly classified nouns and 99.4% of correctly classified other open class words. Overall correct classification in this analysis was 71.0%, Wilks $\lambda = 0.889$, $\chi^2 = 559.408$, $p < 0.001$.

Figure 6.21.: Classification of noun tokens and other tokens for distributional and semantic variables combined.



A final discriminant analysis was performed with tokens using the stepwise method, in order to test the way in which each of the variables contributed to the discriminant function, if they did. From the 19 possible distributional and semantic variables, the stepwise analysis with all tokens selected a total of 16 variables. These 16 variables were introduced in the following order:

1. "Common count nouns" (***Sem2***)
2. "{∅} + x" (***Syn0a***)
3. "{QUANTIFIER} + x + *-(e)s*" (***Syn5b***)
4. "Proper nouns" (***Sem1***)
5. "Common mass nouns" (***Sem3***)
6. "XBLOC words" (***Sem4***)
7. "{∅} + x + *-(e)s*" (***Syn0b***)
8. "{*the*} + x + *-(e)s*" (***Syn2b***)
9. "{*the*} + x " (***Syn2a***)
10. "{POSSESSIVE} + x" (***Syn4a***)
11. "{POSSESSIVE} + x + *-(e)s*" (***Syn4b***)
12. "{PREPOSITION} + x" (***Syn6a***)
13. "{DEMONSTRATIVE} + x" (***Syn3a***)
14. "{*a, an*} + x" (***Syn1***)
15. "{PREPOSITION} + x + *-(e)s*" (***Syn6b***)
16. "{DEMONSTRATIVE} + x + *-(e)s*" (***Syn3b***)

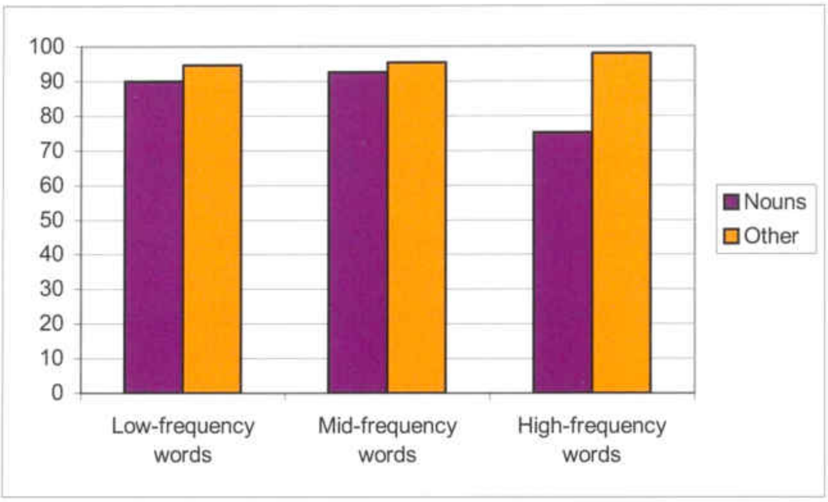
The kind of steps taken in the token stepwise discriminant analysis was then very similar to the ones taken in the type stepwise discriminant analysis. As with types, all the four semantic variables were found to significantly contribute to the discriminant function and were the ones which were introduced earlier, although the advantage of semantic cues over distributional cues was slightly bigger for types than for tokens. Within the selected distributional cues, the advantage of bigram contexts over frames is less evident here, since there were nearly as many bigrams as frames, and bigrams were not necessarily introduced in the analysis earlier than frames. Therefore, this pattern replicates the findings obtained from all previous analyses using distributional cues, either independently or in combination with other variables.

The results obtained from this stepwise discriminant analysis with tokens were very similar to the ones found in the standard discriminant analysis, with only a slight decrease in completeness scores. Thus, there was a total of 47.3% of correctly classified nouns and 98.3% of correctly classified other open class words. Overall correct classification was 69.7% of tokens, Wilks $\lambda = 0.899$, $\chi^2 = 1025.929$, $p < 0.001$.

6.5.2.3. Frequency groups

In order to examine the effect that word frequency had on word categorization when distributional and semantic information was considered, discriminant analyses were also performed with each of the frequency groups using both distributional and semantic cues in combination. The results from these analyses are illustrated in figure 6.22.

Figure 6.22: Percentage of correct classification of noun tokens and other tokens for distributional and semantic variables across different frequency groups.



With low-frequency words, when the 19 variables were introduced simultaneously using the standard method, there were 89.9% of correctly classified nouns and 94.6% of correctly classified other open class words, with an overall correct classification of 91.9% of low-frequency tokens, Wilks λ = 0.346, $\chi^2 = 3379.573$, $p < 0.001$. With mid-frequency words, the same analysis using the same predictor variables gave a total of 92.5% of correctly classified nouns and 95.4% of correctly classified other open class words. Overall correct classification was 93.8% of words, Wilks λ = 0.392, $\chi^2 = 3792.545$, $p < 0.001$. Last, the same analysis with high-frequency words gave a total of 75.2% of correctly classified nouns and 97.9% of correctly classified other open class words. Overall correct classification was 86.2% of words, Wilks λ = 0.693, $\chi^2 = 828.060$, $p < 0.001$.

In the cross-validation analyses using the leave-one-out method, results were very similar to the ones obtained from the discriminant analyses using the standard method, with only slightly lower correct classification scores in all cases. Thus, the leave-one-out discriminant analysis with low-frequency words resulted in 88.9% of correctly classified nouns and 94.6% of correctly classified other open class words, with an overall correct classification score of 91.3%. The leave-one-out discriminant analysis with mid-frequency words gave a total of 92.3% of correctly classified nouns and 95.3% of correctly classified other open class words, with overall correct classification reaching 93.6%. The leave-one-out discriminant analysis with high-frequency words gave a total of 74.8% of correctly classified nouns and 97.7% of correctly classified other open class words, with correct classification reaching 86.0%.

The second cross-validation using 50% of the cases which were randomly selected gave very similar results as well. This second cross-validation with low-frequency words gave nearly identical results, with 89.8% of correctly classified nouns and 94.5% of correctly classified other open class words, Wilks $\lambda = 0.359$, $\chi^2 = 1676.617$, $p < 0.001$. The same analysis with mid-frequency words gave a total of 92.0% of correctly classified nouns and 95.4% of correctly classified other open class words, Wilks $\lambda = 0.393$, $\chi^2 = 1854.910$, $p < 0.001$. The same cross-validation analysis with high-frequency words resulted in a total of 73.9% of correctly classified nouns, and 99.3% of correctly classified other open class words, Wilks $\lambda = 0.670$, $\chi^2 = 451.115$, $p < 0.001$.

Last, but not least, a stepwise discriminant analysis was performed with each of the frequency groups in order to analyze the kind of classificatory strength that each of the variables had with each word kind. The number of steps and variables that were employed with each of the frequency groups is summarized in table 6.19.

Table 6.19: Statistical weight of distributional and semantic cues combined in stepwise discriminant analyses with all frequency groups.

Step	Low-frequency words	Mid-frequency words	High-frequency words
1.	"{∅} + x" (***Syn0a***)	"{∅} + x" (***Syn0a***)	"Common count nouns" (***Sem2***)
2.	"Proper nouns" (***Sem1***)	"Proper nouns" (***Sem1***)	"{∅} + x" (***Syn0a***)
3.	"Common count nouns" (***Sem2***)	"Common count nouns" (***Sem2***)	"Proper nouns" (***Sem1***)
4.	"XBLOC words" (***Sem4***)	"Common mass nouns" (***Sem3***)	"{QUANTIF.} + x + *-(e)s*" (***Syn5b***)
5.	"Common mass nouns" (***Sem3***)	"XBLOC words" (***Sem4***)	"Common mass nouns" (***Sem3***)
6.	"{DEMONSTRATIVE} + x" (***Syn3a***)	"{∅} + x + *-(e)s*" (***Syn0b***)	"XBLOC words" (***Sem4***)
7.	"{PREPOSITION} + x" (***Syn6a***)	"{DEMONSTRATIVE} + x" (***Syn3a***)	"{POSSESSIVE} + x" (***Syn4a***)
8.	"{QUANTIFIER} + x" (***Syn5a***)	"{QUANTIFIER} + x" (***Syn5a***)	"{*the*} + x " (***Syn2a***)
9.	--	"{POSS.} + x + *-(e)s*" (***Syn4b***)	"{POSS.} + x + *-(e)s*" (***Syn4b***)
10.	--	"{DEMONS.} + x + *-(e)s*" (***Syn3b***)	"{*the*} + x + *-(e)s*" (***Syn2b***)
11.	--	"{PREPOSITION} + x" (***Syn6a***)	"{PREPOSITION} + x" (***Syn6a***)
12.	--	--	"{*a, an*} + x" (***Syn1***)
13.	--	--	"{∅} + x + *-(e)s*" (***Syn0b***)

As seen in the table, when distributional and semantic cues are combined, it is semantic cues that become more prominent and contribute more to the discriminant function, as they are the ones which are introduced first. Therefore, the stepwise analyses with frequency groups replicate the pattern obtained from the type and the token analyses. However, the categorization strength of semantic cues decreases as word frequency increases. This is a pattern that also correlates with successful categorization scores when semantic cues were used in isolation, with lower categorization scores among high-frequency words. As to the kind of distributional cues that were used with each of the frequency groups, the pattern found in all previous analyses with distributional cues is also replicated, since biframe distributional contexts are absolutely absent from the low-frequency word analysis, and they become more prominent as word frequency increases.

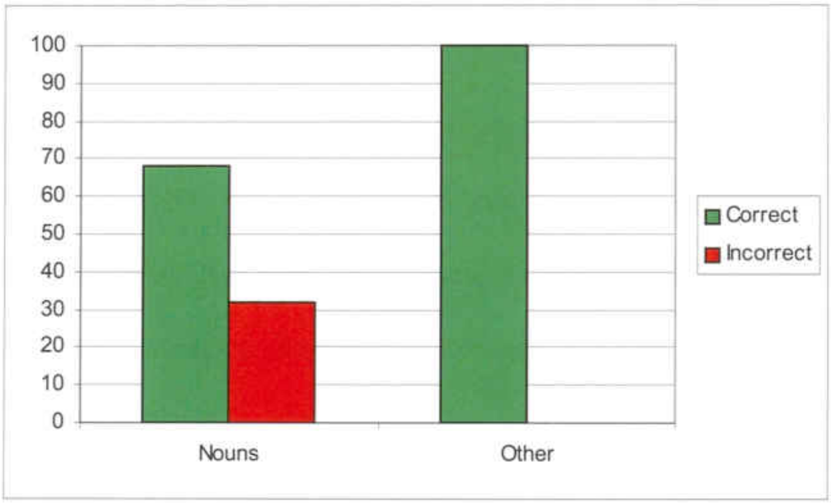
Regarding the results found from these stepwise discriminant analyses, the low-frequency word analysis provided identical results, with 89.9% of correctly classified nouns and 94.6% of correctly classified other open class words, Wilks $\lambda = 0.357$, $\chi^2 = 3374.441$, $p < 0.001$. The mid-frequency word analysis provided slightly lower completeness scores, with 90.8% of correctly classified nouns and 95.5% of correctly classified other open class words, Wilks $\lambda = 0.394$, $\chi^2 = 3783.684$, $p < 0.001$. Completeness scores were also a bit lower among high-frequency words, with 74.7% of correctly classified nouns and 97.9% of correctly classified other open class words, Wilks $\lambda = 0.695$, $\chi^2 = 824.019$, $p < 0.001$.

6.5.3. Phonological and semantic cues

6.5.3.1. Types

The interaction between both phonological and semantic cues towards word categorization was first investigated using all the types from the four child corpora. The results and the outcome of this interaction are illustrated in figure 6.23. When the 16 variables were introduced simultaneously, (i.e. the four semantic variables and the twelve phonological variables) the discriminant analysis with types provided a total of 67.9% of correctly classified nouns and 100% of correctly classified other open class words. Overall correct classification was 82.0%, Wilks $\lambda = 0.431$, $\chi^2 = 8095.422$, $p <$ 0.001. However, despite the high overall correct classification scores, they might be affected by the high accuracy scores obtained. Completeness scores, on the other hand, were far lower. In fact, the scores obtained out of the interaction between phonological and semantic cues with types are exactly the same as the ones that were obtained with semantic cues alone, showing that, with types, there is very little contribution of phonological variables towards word categorization.

Figure 6.23.: Classification of noun types and other types for phonological and semantic variables combined.



The cross-validation discriminant analysis using the leave-one-out method also gave identical results. In the second cross-validation using 50% of randomly selected cases, results for accuracy were identical as well (i.e. 100% of correctly classified other open class words), but completeness scores improved a bit, with 69.1% of correctly classified nouns, and an overall correct classification of 82.7% of types, Wilks λ = 0.443, $\chi^2 = 3891.439$, $p < 0.001$.

A final stepwise discriminant analysis was also performed with all types using the combination of phonological and semantic cues together, in order to examine the exact weight that each cue had in the classificatory system. Out of the 16 possible variables, the stepwise discriminant analysis selected 13 variables and discarded the other three. The exact variables that were chosen and the order in which they were entered was the following:

1. "Common count nouns" (***Sem2***)
2. "Proper nouns" (***Sem1***)
3. "Common mass nouns" (***Sem3***)
4. "XBLOC words" (***Sem4***)
5. "Two syllables or more + trochaic stress pattern" (***Phoncomb14***)
6. "Two syllables or more + low stressed vowel + trochaic stress pattern (***Phoncomb134***)
7. "Two syllables or more + final voiced consonant + low stressed vowel" (***Phoncomb123***)
8. "Two syllables or more" (***Phon1***)
9. "Two syllables or more + low stressed vowel" (***Phoncomb13***)
10. "Two syllables or more + final voiced consonant" (***Phoncomb12***)
11. "Final voiced consonant + low stressed vowel" (***Phoncomb23***)
12. "Two syllables or more + final voiced consonant + low stressed vowel + trochaic stress pattern" (***Phoncomb1234***)
13. "Final voiced consonant" (***Phon2***)

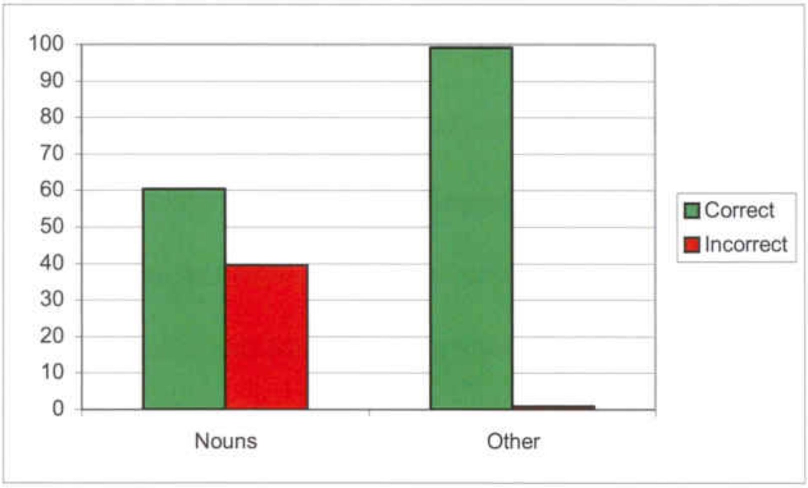
Thus, it is mainly semantic variables the ones that had more statistical weight and the ones that most contributed to the discriminant function (i.e. the four semantic variables occupy the four first steps of the analysis respectively). Variable *Sem2* appeared to be the strongest of the four, as in most of the previous analyses with semantic variables. From the selection of phonological variables, most of them corresponded to clusters of two or three features, and the four-feature variable was also included. This also replicates the findings of previous analyses with phonological variables.

The results obtained from this stepwise analysis were identical to those obtained from the standard analysis when all cues were entered simultaneously. Recall that these results were also identical to the results found with semantic variables alone (i.e. both the standard analysis and the stepwise analysis). Therefore, although the six phonological variables that were taken in this stepwise discriminant analysis were selected for being strong enough to contribute to the discriminant function in terms of statistics, the effect that such strength had on overall correct word classification was minimal.

6.5.3.2. Tokens

As with types, the interaction between phonological and semantic information was also examined with tokens. The standard discriminant analysis with tokens when the combination of all phonological and semantic variables were entered simultaneously gave a total of 60.4% of correctly classified nouns and 99.1% of correctly classified other open class words. There was an overall correct classification of 77.5% of tokens, Wilks $\lambda = 0.922$, $\chi^2 = 780.538$, $p < 0.001$. The results from this analysis are shown in figure 6.24.

Figure 6.24.: Classification of noun tokens and other tokens for phonological and semantic variables combined.



Recall that the same analysis with only phonological variables gave very high completeness scores but very low accuracy scores (see section 6.3.3.2), while the same analysis with only semantic variables gave the opposite pattern: very low completeness scores but very high accuracy scores (see section 6.4.3.2). Thus, word categorization is seen to benefit from the interaction between phonological and semantic variables, since the high accuracy scores provided by semantic information are kept and are not affected by phonological information (i.e. there were 99.1% of correctly classified other open class words). At the same time, the completeness scores provided by semantic information are significantly improved by the help of phonological information and, even if they do not reach the high percentage obtained with phonological variables alone, the score obtained out of the combination of phonological and semantic information is quite above chance levels.

The cross-validation analysis using the leave-one-out method provided exactly the same results. The cross-validation analysis using 50% of randomly selected cases gave very similar results as well, with identical results for completeness and slightly lower results for accuracy. Thus, the total of nouns that resulted correctly classified from this cross-validation was 60.4%, while 98.3% of other open class words were correctly classified. Overall correct classification was 77.1%, Wilks $\lambda = 0.915$, $\chi^2 = 423.016$, $p < 0.001$.

A stepwise discriminant analysis was also performed with all tokens using both phonological and semantic variables. When the 16 variables were used, the stepwise analysis with tokens took 13 variables and discarded the other three. The variables that were chosen and the order in which they were entered was the following:

1. "Common count nouns" (***Sem2***)
2. "Absence of selected phonological features" (***Phon0***)
3. "Common mass nouns" (***Sem3***)
4. "Final voiced consonant" (***Phon2***)
5. "Two syllables or more + final voiced consonant + trochaic stress pattern" (***Phoncomb124***)
6. "Stressed low vowel" (***Phon3***)
7. "Proper nouns" (***Sem1***)
8. "Final voiced consonant + stressed low vowel" (***Phoncomb23***)
9. "Two syllables or more + trochaic stress pattern" (***Phoncomb14***)
10. "Two syllables or more" (***Phon1***)
11. "Two syllables or more + final voiced consonant" (***Phoncomb12***)
12. "XBLOC words" (***Sem4***)
13. "Two syllables or more + final voiced consonant + stressed low vowel + trochaic stress pattern" (***Phoncomb1234***)

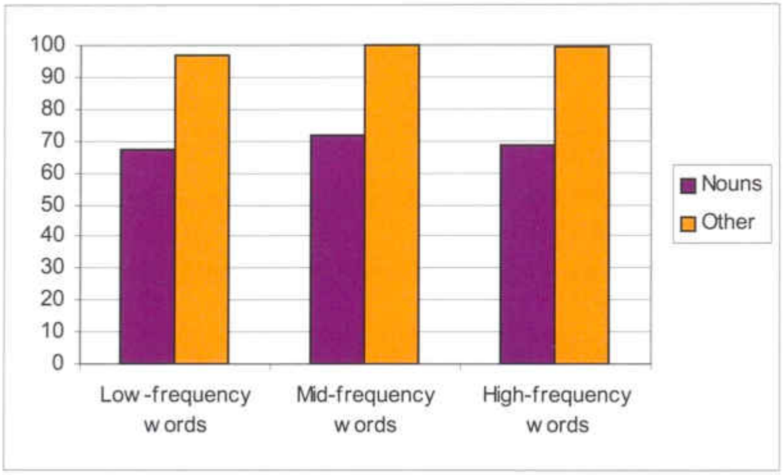
The steps taken in this discriminant analysis reveal a real interaction between phonological and semantic information. Thus, unlike the analysis with types, the first steps of the discriminant analysis with tokens are not exclusively dominated by semantic variables. On the contrary, semantic and phonological variables are merged throughout the various steps. All four phonological variables were used in the analysis, but their positions at the different steps range from step 1 (*Sem2*) to the penultimate step (*Sem4*). As to phonological variables, they also occupy several positions throughout the analysis, and the variables include all kinds of phonological clusters (i.e. absence of phonological features, single-feature variables as well as clusters of two, three and four features).

As to the results found from this stepwise discriminant analysis, they were very similar to the ones obtained from the standard discriminant analysis using the 16 variables simultaneously. Thus, in the stepwise discriminant analysis with tokens, 59.7% of nouns were correctly classified and 99.8% of other open class words were correctly classified. There was an overall correct classification of 77.4% of tokens, Wilks $\lambda = 0.923$, $\chi^2 = 774.084$, $p < 0.001$.

6.5.3.3. Frequency groups

The effects that word frequency had on word categorization when phonological and semantic information was considered were also tested by performing discriminant analyses with each of the frequency groups using both phonological and semantic cues in combination. The results from these analyses are illustrated in figure 6.25. For low-frequency words, 67.6% of nouns were correctly classified, and 97.1% of other open class words were correctly classified. There was an overall correct classification of 80.2% of words, Wilks $\lambda = 0.477$, $\chi^2 = 2421.385$, $p < 0.001$. Among mid-frequency words, there were 71.9% of correctly classified nouns and 99.8% of correctly classified other open class words, with an overall correct classification of 83.7% of words, Wilks $\lambda = 0.529$, $\chi^2 = 2584.557$, $p < 0.001$. Last, among high-frequency words, there was a total of 68.5% of correctly classified nouns and 99.3% of correctly classified other open class words, with an overall correct classification of 83.5% of words, Wilks $\lambda = 0.760$, $\chi^2 = 620.964$, $p < 0.001$.

Figure 6.25: Percentage of correct classification of noun tokens and other tokens for phonological and semantic variables across different frequency groups.



When these results were cross-validated using the leave-one-out method, very similar results were obtained. In the low-frequency word leave-one-out analysis, there were 65.8% of correctly classified nouns and 97.1% of correctly classified other open class words, Wilks $\lambda = 0.477$, $\chi^2 = 2421.385$, $p < 0.001$. For mid-frequency words, the results obtained from the leave-one-out cross-validation were identical to those obtained from the standard analysis. The same cross-validation with high-frequency words gave also almost identical results, with only a minimal change in completeness scores: 68.4% of the nouns were correctly classified and 99.3% of other open class words were correctly classified as well, Wilks $\lambda = 0.760$, $\chi^2 = 620.964$, $p < 0.001$.

In the second cross-validation using 50% of randomly selected cases, results were very similar as well. In the cross-validation with low-frequency words, there were 66.3% of correctly classified nouns and 98.3% of correctly classified other open class words, Wilks $\lambda = 0.490$, $\chi^2 = 1168.025$, $p < 0.001$. The same cross-validation with mid-frequency words gave a total of 74.4% of correctly classified nouns and 99.1% of correctly classified other open class words, Wilks $\lambda = 0.543$, $\chi^2 = 1214.220$, $p < 0.001$. With high-frequency words, the same cross-validation gave a total of 70.2% of correctly classified nouns and 98.9% of correctly classified other open class words, Wilks $\lambda = 0.739$, $\chi^2 = 342.186$, $p < 0.001$.

Last, a stepwise discriminant analysis was also performed with each of the frequency groups in order to analyze the weight that each of the variables had across different word frequency ranges. The number of steps and variables that were employed with each of the frequency groups is summarized in table 6.20.

The kind of variables that the three different stepwise analyses have taken at each step replicate the patterns obtained so far with the stepwise analyses with types and tokens. Thus, semantic variables appear to be especially powerful when word frequency is not considered (i.e. in the type analysis or in the low-frequency word analysis), but as word frequency increases, semantic variables become less prominent and phonological variables are increasingly found to contribute more to the classificatory system. Therefore, although the four semantic variables are entered in the three analyses, they occupy different step positions in each of them, ranging from the very first steps with low-frequency words, to the antipenultimate step with high-frequency words. The kind of phonological variables employed in each of the analyses varies as well, but it is mostly phonological features connected to word length and final voicing the ones that are found to contribute most to the discriminant functions.

Table 6.20: Statistical weight of phonological and semantic cues combined in stepwise discriminant analyses with all frequency groups.

Step	Low-frequency words	Mid-frequency words	High-frequency words
1.	“Common count nouns” (***Sem2***)	“Common count nouns” (***Sem2***)	“Common count nouns” (***Sem2***)
2.	“Proper nouns” (***Sem1***)	“Proper nouns” (***Sem1***)	“Common mass nouns” (***Sem3***)
3.	“XBLOC words” (***Sem4***)	“Common mass nouns” (***Sem3***)	“No phonological features” (***Phon0***)
4.	“Common mass nouns” (***Sem3***)	“No phonological features” (***Phon0***)	“Final voiced consonant” (***Phon2***)
5.	“Two syllables or more + trochaic stress patter” (***Phoncomb14***)	“Two syllables or more + final voiced consonant + trochaic stress pattern” (***Phoncomb124***)	“Proper nouns” (***Sem1***)
6.	“Two syllables or more” (***Phon1***)	“Low stressed vowel” (***Phon3***)	“Final voiced consonant + low stressed vowel” (***Phoncomb23***)
7.	“Two syllables or more + low stressed vowel” (***Phoncomb13***)	“Final voiced consonant” (***Phon2***)	“Two syllables or more + trochaic stress patter” (***Phoncomb14***)
8.	“Two syllables or more + low stressed vowel + trochaic stress pattern” (***Phoncomb134***)	“Two syllables or more + final voiced consonant + low stressed vowel + trochaic stress pattern” (***Phoncomb1234***)	“Two syllables or more + final voiced consonant + trochaic stress pattern” (***Phoncomb124***)
9.	--	“XBLOC words” (***Sem4***)	“Low stressed vowel” (***Phon3***)
10.	--	“Two syllables or more + trochaic stress patter” (***Phoncomb14***)	“XBLOC words” (***Sem4***)
11.	--	“Two syllables or more + final voiced consonant” (***Phoncomb12***)	“Two syllables or more” (***Phon1***)
12.	--	“Final voiced consonant + low stressed vowel” (***Phoncomb23***)	“Two syllables or more + final voiced consonant” (***Phoncomb12***)
13.	--	“Two syllables or more + low stressed vowel + trochaic stress pattern” (***Phoncomb134***)	
14.	--	“Two syllables or more” (***Phon1***)	
15.	--	“Two syllables or more + final voiced consonant + low stressed vowel” (***Phoncomb123***)	
16.	--	“Two syllables or more + low stressed vowel” (***Phoncomb13***)	

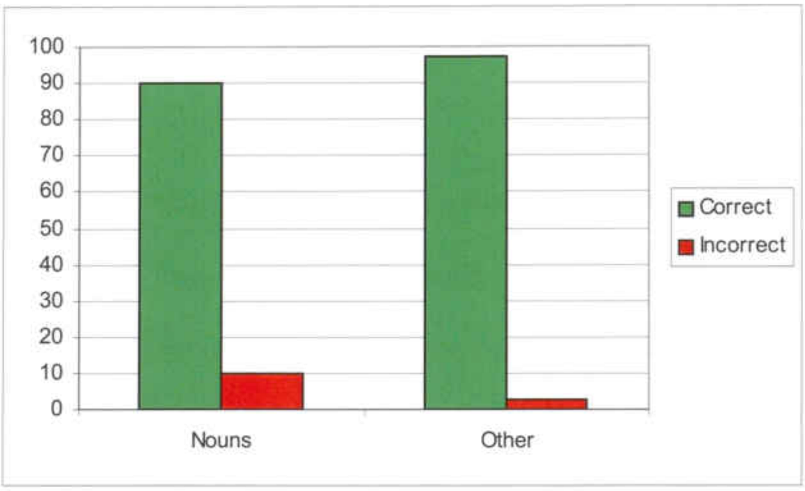
As to the results obtained from these three stepwise discriminant analyses, they were also very similar to the results found in all previous analyses with frequency groups using phonological and semantic variables. The low-frequency word analysis gave a total of 65.8% of correctly classified nouns and 98.5% of correctly classified other open class words, Wilks $\lambda = 0.478$, $\chi^2 = 2415.134$, $p < 0.001$. The mid-frequency word analysis gave exactly the same correct classification scores as the standard analysis, Wilks $\lambda = 0.529$, $\chi^2 = 2584.557$, $p < 0.001$. In fact, the stepwise discriminant analysis with mid-frequency words is the only one in which no variables were discarded and the 16 variables were introduced. Thus, this analysis used exactly the same information as the corresponding discriminant analysis using the standard method, the only difference being that each variable was introduced individually at each step in the stepwise analysis, while all variables were introduced simultaneously in the standard analysis. As to the stepwise analysis with high-frequency words, results were very similar to the ones from the standard analysis as well, with 67.9% of correctly classified nouns and 99.4% of correctly classified other open class words, Wilks $\lambda = 0.763$, $\chi^2 = 613.699$, $p < 0.001$.

6.5.4. Distributional, phonological and semantic cues

6.5.4.1. Types

Finally, the interaction between the three kinds of cues towards word categorization was investigated. In a first analysis, all the types from the four child corpora were used. The results of this analysis are illustrated in figure 6.26. When the 31 variables were introduced simultaneously, (i.e. the four semantic variables and the twelve phonological variables and the fifteen distributional variables) the discriminant analysis with types provided a total of 90.1% of correctly classified nouns and 97.0% of correctly classified other open class words. There was an overall correct classification of 93.1% of correctly classified types, Wilks $\lambda = 0.308$, $\chi^2 = 11323.821$, $p < 0.001$.

Figure 6.26.: Classification of noun types and other types for distributional, phonological and semantic variables combined.



In the cross-validation analysis using the leave-one-out method, the results found were exactly the same as those obtained from the standard discriminant analysis. In the cross-validation analysis using 50% of randomly selected cases, the results obtained were very similar as well, with a slight improvement in both accuracy and completeness scores. Thus, this second cross-validation gave a total of 90.4% of correctly classified nouns and 97.2% of correctly classified other open class words, with an overall correct classification of 93.4% of words, Wilks $\lambda = 0.311$, $\chi^2 = 5572.810$, $p < 0.001$.

A stepwise discriminant analysis was also run with types using all cues in combination, in order to see the exact weight that each kind of cue had within the classificatory system. Out of the 31 possible variables, the stepwise discriminant analysis with types took 20 variables in 20 different steps, and discarded the other 11. The variables that were chosen at each step and the order in which they were entered was the following:

1. “Common count nouns” (***Sem2***)
2. “Proper nouns” (***Sem1***)
3. “{∅} + x” (***Syn0a***)
4. “Common mass nouns” (***Sem3***)
5. “XBLOC words” (***Sem4***)
6. “{*the*} + x ” (***Syn2a***)
7. “{*a, an*} + x ” (***Syn1***)
8. “{POSSESSIVE} + x” (***Syn4a***)
9. “{∅} + x + *-(e)s*” (***Syn0b***)
10. “{*the*} + x + *-(e)s*” (***Syn2b***)
11. “Two syllables or more + trochaic stress pattern” (***Phoncomb14***)
12. “Two syllables or more + final voiced consonant + trochaic stress pattern” (***Phoncomb124***)
13. “{WH- ELEMENT} + x” (***Syn7a***)
14. “Two syllables or more” (***Phon1***)
15. “Two syllables or more + stressed low vowel” (***Phoncomb13***)
16. “Two syllables or more + final voiced consonant + stressed low vowel” (***Phoncomb123***)
17. “Two syllables or more + stressed low vowel + trochaic stress pattern” (***Phoncomb134***)
18. “Two syllables or more + final voiced consonant” (***Phoncomb12***)
19. “{QUANTIFIER} + x + *-(e)s*” (***Syn5b***)
20. “{QUANTIFIER} + x” (***Syn5a***)

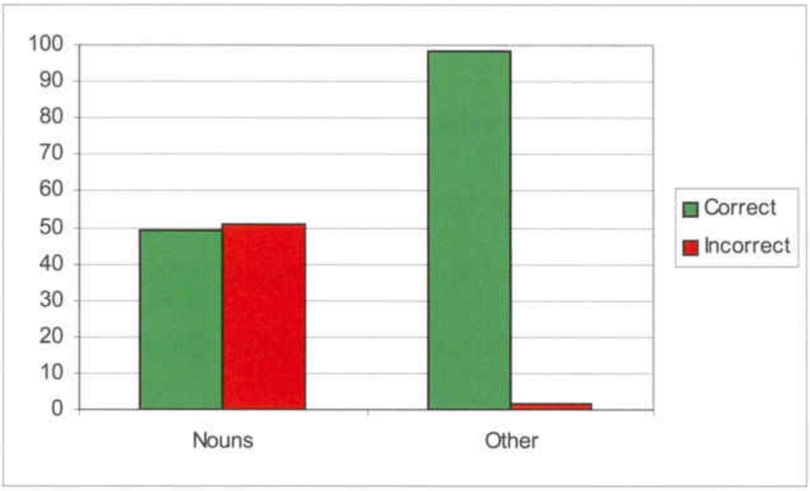
As can be seen, the variables which were found to contribute most to the discriminant function were both semantic and distributional variables in interaction. As far as semantic variables are concerned, all four variables were entered, and their positions ranged from the first step to the fifth. As to distributional variables, nine out of fifteen were introduced, the rest were discarded. They also occupied different positions, ranging from the third position to the last. Most of the distributional variables which were used were bigram contexts. Regarding phonological variables, seven out of twelve were used, and they mainly occupied the last positions of the stepwise discriminant analysis, ranging from step 11th to the 18th. Among the phonological variables considered, the most recurrent feature was word length.

As far as the results from this stepwise discriminant analysis are concerned, they were identical to those found from the standard discriminant analysis with types when all 31 cues were entered simultaneously, with 90.1% of correctly classified nouns and 97.0% of correctly classified other open class words, Wilks $\lambda = 0.308$, $\chi^2 = 11314.892$, $p < 0.001$. This indicates that the 11 cues that were discarded did not statistically contribute to the discriminant function in a significant way, nor did they contribute towards successful word categorization.

6.5.4.2. Tokens

The interaction between the three kinds of cues towards word categorization was also investigated with all tokens from the four corpora together. The results from the standard discriminant analysis with all tokens using all 31 variables together are illustrated in figure 6.27. There was a total of 49.1% of correctly classified nouns and 98.3% of correctly classified other open class words, with an overall correct classification of 70.7% of tokens, Wilks $\lambda = 0.894$, $\chi^2 = 1076.783$, $p < 0.001$.

Figure 6.27.: Classification of noun tokens and other tokens for distributional, phonological and semantic variables combined.



The cross-validation discriminant analysis with tokens using the leave-one-out method gave exactly the same results as the corresponding standard discriminant analysis with tokens also, except for minimally lower completeness scores (i.e. in the leave-one-out discriminant analysis there were 48.9% of correctly classified nouns). However, in the cross-validation discriminant analysis using 50% of randomly selected cases, successful categorization improved slightly: there were 52.7% of correctly classified nouns and 97.4% of correctly classified other open class words, with overall correct classification reaching 72.4% of tokens, Wilks $\lambda = 0.881$, $\chi^2 = 606.629$, $p < 0.001$.

A stepwise discriminant analysis was also run with tokens using all of the cues in combination, in order to see the exact weight that each kind of cue had within the classificatory system. Out of the 31 possible variables, the stepwise discriminant analysis with tokens took 19 variables and discarded the other 12. The variables that were chosen at each step and the order in which they were entered were the following:

1. "Common count nouns" (***Sem2***)
2. "{∅} + x" (***Syn0a***)
3. "{QUANTIFIER} + x + *-(e)s*" (***Syn5b***)
4. "Proper nouns" (***Sem1***)
5. "Common mass nouns" (***Sem3***)
6. "XBLOC words" (***Sem4***)
7. "{∅} + x + *-(e)s*" (***Syn0b***)
8. "{*the*} + x + *-(e)s*" (***Syn2b***)
9. "{*the*} + x " (***Syn2a***)
10. "{POSSESSIVE} + x" (***Syn4a***)
11. "Two syllables or more + trochaic stress pattern" (***Phoncomb14***)
12. "Final voiced consonant + low stressed vowel" (***Phoncomb23***)
13. "{POSSESSIVE} + x + *-(e)s*" (***Syn4b***)
14. "Low stressed vowel" (***Phon3***)
15. "{PREPOSITION} + x" (***Syn6a***)
16. "Two syllables or more + final voiced consonant + low stressed vowel" (***Phoncomb123***)
17. "{*a, an*} + x " (***Syn1***)
18. "{DEMONSTRATIVE} + x" (***Syn3a***)
19. "Two syllables or more" (***Phon1***)

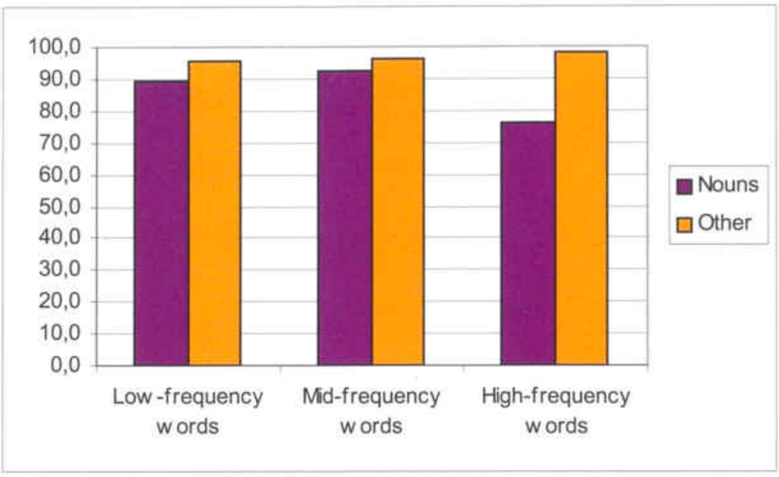
In the stepwise discriminant analysis with tokens, semantic and distributional variables were also the ones which were found to be statistically more salient. All of the semantic variables were entered, and their positions range from the first to the sixth step. Thus, the first steps of the discriminant analysis were shared by both semantic and distributional variables. A total of ten different distributional variables were entered, and they were widely distributed along the whole stepwise analysis, ranging from the second position to the penultimate step. As to phonological variables, only five out of the twelve possible variables were introduced, and they were mainly relegated to the last positions of the stepwise analysis, ranging from the 11th step to the last step. They mostly refered to features connected to word length and vowel quality.

As to the results obtained from this stepwise discriminant analysis, they were very similar to the ones obtained from the standard analysis using all 31 variables: there was a total of 48.0% of correctly classified nouns and 98.3% of correctly classified other open class words, with overall correct classification reaching 70.2% of tokens, Wilks $\lambda = 0.896$, $\chi^2 = 1056.996$, $p < 0.001$. This indicates that the set of twelve variables that were discarded (i.e. five distributional and seven phonological) only contributed with a slight improvement in completeness scores.

6.5.4.3. Frequency groups

The effect that word frequency had on word categorization when all cues were considered was also tested by performing discriminant analyses with each of the frequency groups using all of the 31 variables combination. The results from these analyses are illustrated in figure 6.28. For low-frequency words, 89.3% of nouns and 95.5% of other open class words were correctly classified. There was an overall correct classification of 91.9% of words, Wilks $\lambda = 0.351$, $\chi^2 = 3422.374$, $p < 0.001$. Among mid-frequency words, there were 92.7% of correctly classified nouns and 96.2% of correctly classified other open class words, with an overall correct classification of 94.2% of words, Wilks $\lambda = 0.389$, $\chi^2 = 3822.950$, $p < 0.001$. Last, among high-frequency words, there was a total of 75.9% of correctly classified nouns and 97.9% of correctly classified other open class words, with an overall correct classification of 86.6% of words, Wilks $\lambda = 0.681$, $\chi^2 = 865.211$, $p < 0.001$.

Figure 6.28: Percentage of correct classification of noun tokens and other tokens for all variables across different frequency groups.



The respective cross-validation discriminant analyses using the leave-one-out method gave almost identical results to the respective standard discriminant analyses in all cases. For low frequency words, 89.3% of nouns and 95.3% of other open class words were correctly classified. Among mid-frequency words, 92.6% of nouns and 96.1% of other open class words were correctly classified. For high-frequency words, correct classification reached 75.3% in the case of nouns and 97.6% in the case of other open class words.

Results were almost identical as well in the second cross validation analyses where 50% of cases selected at random were used. The low-frequency word cross-validation analysis gave a total of 89.5% of nouns and 95.2% of other open class words which were correctly classified, Wilks $\lambda = 0.354$, $\chi^2 = 1694.136$, $p < 0.001$. The same cross-validation with mid-frequency words resulted in 91.2% of nouns and 96.0% of other open class words which were correctly classified, Wilks $\lambda = 0.389$, $\chi^2 = 1867.920$, $p < 0.001$. The same cross-validation analysis with high-frequency words resulted in 75.8% of nouns and 98.2% of other open class words which were correctly classified, Wilks $\lambda = 0.649$, $\chi^2 = 484.969$, $p < 0.001$.

Last, a stepwise discriminant analysis was also performed with each of the frequency groups using all of the 31 variables, in order to analyze the weight that each kind of variable had across different word frequency ranges. The number of steps and variables that were employed with each of the frequency groups is summarized in table 6.21. The distribution of each of the steps across different word frequency groups shows that the three kind of variables work differently towards word categorization in each of the frequency groups.

Thus, for low-frequency words, there were as many semantic variables as distributional and phonological variables (i.e. four of each). Proportionally, this entails that semantic variables were more salient, since all of them were entered in the analysis and they also occupied the first positions. Distributional and phonological variables occupied the middle and last positions respectively, although proportionally, distributional variables were probably less salient, since only four variables out of fifteen were included. Phonological variables gained saliency among mid-frequency words, six out of the sixteen steps taken involved a phonological feature or a cluster. They also shared middle and last step positions with distributional variables, while semantic variables still occupied the first steps. Last, among high-frequency words is where distributional variables appear to be the strongest, with nine out of sixteen steps being connected to a distributional context. Distributional variables also shared first and middle positions with semantic variables, which were shown to be less powerful among high-frequency words. Phonological variables were the weakest among high-frequency words, with only three variables considered, and only entered in the last steps.

Table 6.21: Statistical weight of all cues combined in stepwise discriminant analyses with all frequency groups.

Step	**Low-frequency words**	**Mid-frequency words**	**High-frequency words**
1.	“{∅} + x” (***Syn0a***)	“{∅} + x” (***Syn0a***)	“Common count nouns” (***Sem2***)
2.	“Proper nouns” (***Sem1***)	“Proper nouns” (***Sem1***)	“{∅} + x” (***Syn0a***)
3.	“Common count nouns” (***Sem2***)	“Common count nouns” (***Sem2***)	“Proper nouns” (***Sem1***)
4.	“XBLOC words” (***Sem4***)	“Common mass nouns” (***Sem3***)	“{QUANTIFIER} + x + *-(e)s*” (***Syn5b***)
5.	“Common mass nouns” (***Sem3***)	“XBLOC words” (***Sem4***)	“Common mass nouns” (***Sem3***)
6.	“{DEMONSTRATIVE} + x” (***Syn3a***)	“{∅} + x + *-(e)s*” (***Syn0b***)	“XBLOC words” (***Sem4***)
7.	“{PREPOSITION} + x” (***Syn6a***)	“{DEMONSTRATIVE} + x” (***Syn3a***)	“{POSSESSIVE} + x” (***Syn4a***)
8.	“{QUANTIFIER} + x” (***Syn5a***)	“Two syllables or more + final voiced consonant + trochaic stress pattern” (***Phoncomb124***)	“{*the*} + x ” (***Syn2a***)
9.	“Two syllables or more + trochaic stress pattern” (***Phoncomb14***)	“{QUANTIFIER} + x” (***Syn5a***)	“{POSSESSIVE} + x + *-(e)s*” (***Syn4b***)
10.	“Two syllables or more” (***Phon1***)	“{POSSESSIVE} + x + *-(e)s*” (***Syn4b***)	“{*the*} + x + *-(e)s*” (***Syn2b***)
11.	“Two syllables or more + low stressed vowel” (***Phoncomb13***)	“{DEMONSTR.} + x + *-(e)s*” (***Syn3b***)	“{PREPOSITION} + x” (***Syn6a***)
12.	“Final voiced consonant + low stressed vowel” (***Phoncomb23***)	“Final voiced consonant” (***Phon2***)	“Low stressed vowel” (***Phon3***)
13.	--	“Two syllables or more + final voiced consonant + low stressed vowel + trochaic stress pattern” (***Phoncomb1234***)	“{*a, an*} + x ” (***Syn1***)
14.	--	“Absence of phonological features” (***Phon0***)	“Final voiced consonant + low stressed vowel” (***Phoncomb23***)
15.	--	“Low stressed vowel” (***Phon3***)	“Two syllables or more + trochaic stress pattern” (***Phoncomb14***)
16.	--	“Final voiced consonant + low stressed vowel” (***Phoncomb23***)	“{∅} + x + *-(e)s*” (***Syn0b***)

As to the results obtained from these stepwise discriminant analyses, they were also very similar to the ones obtained from the previous analyses using all cues in each of the frequency groups. The low-frequency word stepwise discriminant analysis gave a total of 89.0% of nouns and 95.6% of other open class words which were correctly classified, Wilks $\lambda = 0.353$, $\chi^2 = 3406.863$, $p < 0.001$. The mid-frequency word stepwise analysis resulted in a total of 92.6% of nouns and 95.5% of other open class words which were correctly classified, Wilks $\lambda = 0.390$, $\chi^2 = 3817.116$, $p < 0.001$. The same analysis with high-frequency words resulted in a total of 75.3% of nouns and 98.1% of other open class words which were correctly classified, Wilks $\lambda = 0.685$, $\chi^2 = 855.628$, $p < 0.001$.

6.5.5. Summary of results with cues in interaction

The analyses performed with different kinds of cues in interaction have shown that the three types of cues contribute differently toward word categorization. In the interaction between distributional and phonological cues, there was a slight improvement in overall correct word classification, indicating that accuracy scores were kept as high as with distributional cues alone, and were not affected by the low accuracy scores obtained from phonological cues alone. At the same time, the high completeness scores obtained with distributional and phonological cues alone were improved, so that completeness scores were even higher using both kinds of cues in combination. In general, distributional cues dominated all the analyses and phonological cues served at improving several of the completeness scores. As far as word frequency is concerned, the pattern obtained from distributional cues alone was kept, since overall correct classification with distributional and phonological cues together improved as word frequency increased.

As to the interaction between semantic and distributional cues, a similar pattern emerges. Semantic cues contributed in that high accuracy scores were kept, but the low completeness scores obtained from semantic cues alone were compensated this time with distributional information. Therefore, the combination of semantic and distributional variables gave very high accuracy and completeness scores. Semantic variables appeared to dominate all the analyses, although their categorization strength decreased as word frequency increased.

Overall correct classification also benefited from the interaction between phonological and semantic information, since the high accuracy scores obtained from semantic cues were kept in all cases and compensated for low accuracy scores obtained with phonological information only. At the same time, low completeness scores obtained from semantic cues alone were improved by the contribution of phonological information, giving much higher completeness scores than semantic information alone. However, overall correct classification using phonology and semantics was not as high as the one obtained from the interaction between semantics and syntax.

Finally, as far as the interaction between all three types of variables are concerned, overall correct classification improved in all cases, since it is in the last analyses where the highest accuracy and completeness scores are obtained for all word groups. Furthermore, different kinds of cues were seen to contribute differently, since semantic and phonological cues appeared to be the most powerful when word frequency was not considered (i.e. in the type analysis or with low- and mid- frequency words). However, as word frequency increased, the categorization strength of semantic and phonological variables diminished while that of distributional variables improved. A summary of all correct classification scores obtained in all the analysis is presented in table 6.22.

Table 6.22.: Correct classification scores obtained from all the analyses with all possible variable combinations.

	Types			Tokens			Low-frequency words			Mid-frequency words			High-frequency words		
	Noun	**Other**	**Overall**	**Noun**	**Other**	**Overall**	**Noun**	**Other**	**Overall**	**Noun**	**Other**	**Overall**	**Noun**	**Other**	**Overall**
Syn	72.8%	88.8%	**79.9%**	44.6%	96.5%	**67.4%**	70.1%	85.6%	**76.7%**	78.5%	88.0%	**82.5%**	70.2%	94.8%	**82.8%**
Phon	40.1%	71.6%	**53.9%**	84.9%	22.1%	**57.3%**	31.4%	76.7%	**50.7%**	57.2%	55.6%	**56.5%**	81.4%	36.4%	**59.5%**
Sem	67.9%	100%	**82.0%**	39.9%	100%	**66.3%**	61.6%	100%	**77.9%**	71.5%	100%	**83.6%**	63.8%	100%	**81.4%**
Syn+Phon	72.1%	88.9%	**79.5%**	46.4%	95.6%	**68.0%**	70.1%	85.6%	**76.7%**	79.3%	87.4%	**82.8%**	71.6%	94.7%	**83.4%**
Syn+Sem	90.1%	96.8%	**93.0%**	47.9%	98.3%	**70.1%**	89.9%	94.6%	**91.9%**	92.5%	95.4%	**93.8%**	75.2%	97.9%	**86.2%**
Phon+Sem	67.9%	100%	**82.0%**	60.4%	99.1%	**77.5%**	67.6%	97.1%	**80.2%**	71.9%	99.8%	**83.7%**	68.5%	99.3%	**83.5%**
Syn+Phon+Sem	90.1%	97.0%	**93.1%**	49.1%	98.3%	**70.7%**	89.3%	95.5%	**91.9%**	92.7%	96.2%	**94.2%**	75.9%	97.9%	**86.6%**

Chapter 7:
Discussion

The main objective of this dissertation was to analyze whether the linguistic samples to which young English-learning children are exposed contain information which is sufficient and reliable enough so that children could use it to form an abstract grammatical category for English nouns without the need to postulate any *a priori* linguistic knowledge about grammatical categories. As seen in chapter 4, three different sources of information (i.e. distributional, phonological and semantic information) might well contribute to the grammatical categorization of English nouns, be it on an individual basis (i.e. research question 1) or as the product of the combination of two or more sources of information together (i.e. research question 2). The results presented in chapter 6 suggest that distributional, phonological and semantic cues can be useful sources of information for the formation of the grammatical category of nouns in English, and that, when they are in interaction with one another, such sources of information provide an even more accurate representation of the English nominal grammatical category.

This chapter discusses the results obtained from the corpus analysis using each of the sources of information under consideration, in light of the evidence obtained from previous studies with similar objectives. The first section will be devoted to the results and implications from the analyses using individual cues, whereas the second section will be devoted to the results and implications from the analyses with cues in combination. Final remarks will follow in which general theoretical implications are discussed, as well as the limitations of the present study and issues for further research.

7.1. Individual cues

7.1.1. Distributional cues

As seen in chapter 5, the template that was taken to define distributional contexts in the present study took the form of either a bigram (i.e. one categorizing element plus one intervening element to be categorized, either to the left or to the right), or frame (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, Chater & Christiansen, 2005), or just any kind of context word (Redington, Chater & Finch, 1998; Mintz, 2003).

As a result, the number of distributional contexts that emerged in the present analysis 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 (see appendix A).

Significance tests showed that differences between nouns and other open class words obtained from the corpora under analysis 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. The only exception to this pattern was the two distributional contexts in which the categorizing context word was either a preposition or a *wh-* element. In those cases, significant differences between nouns and other open class words were found as well, but with an advantage of other open class words over nouns. This advantage only indicates that these elements are more likely to precede other open class words than nouns, while determiners in general are more likely to precede nouns than 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 accuracy 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 53.0% of correctly classified nominal tokens (Redington, Chater & Finch, 1998) to 62.4% of correctly classified tokens (Monaghan, Chater & Christiansen, 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, further analyses were carried out with tokens considering them separately in each of the frequency groups established. When correct classification of nominal elements was assessed in interaction with word frequency, completeness scores improved significantly. Thus, 70.1% of low-frequency nouns, 78.5% of mid-frequency nouns and 70.2% of high-frequency nouns were correctly classified. Therefore, given the fact that completeness scores were significantly high among each of the word-frequency groups, especially high-frequency words, distributional cues to nominal categorization in English child-directed speech are at no risk of being missed by young language learners.

Furthermore, the analysis with all fifteen distributional cues using all the types from the corpora provided 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; 1998; Croft, 2000; Maratsos, 2000; Marchman & Bates, 1994; Ravid, 1995) 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.

Criticisms to distributional analysis approaches (e.g. Chomsky, 1975; Pinker, 1984; 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, Chater & Christiansen, 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 analyze 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 on noun phrases, the prediction regarding miscategorization was that accuracy scores would be high, that is, that the percentage of other open class words wrongly misclassified as nouns on the basis of the selected distributional variables would be low.

This prediction was also confirmed 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 misclassified as nouns.

Accuracy scores obtained from the analyses with types or each of the three different word frequency groups 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 across different frequency groups ranged from 85.6% among low-frequency words to 88.0% among mid-frequency words and 94.8% among high-frequency words. Therefore, on the basis of the evidence provided by these data, the kind of distributional contexts established in the present dissertation for the purpose of the grammatical categorization of English nouns can be claimed to be accurate enough as not to overcategorize and subsume elements other than nouns, which would ultimately avoid wrong assumptions and miscategorization of certain elements on the part of young language learners.

More interestingly, when distributional cues alone are considered, overall successful categorization scores (i.e. taking into account both accuracy and completeness measures altogether) have been seen to improve as word frequency increases. Thus, overall successful categorization scores among low-frequency words reached a total of 76.7%, while successful classification reached a total of 82.5% among mid-frequency words and 82.8% among high-frequency words. These results replicate the findings of recent similar studies (i.e. Monaghan, Chater & Christiansen, 2005; Monaghan, Christiansen & Chater, 2007) where the same interaction between word frequency and accuracy of distributional information was also found: correct classification was very accurate among high-frequency words and dropped away for low-frequency words[16]. Crucially, given the fact that highly frequent words are more salient in the input and more likely to catch the learner's attention, these findings give more power to distributional information, given the fact that it is precisely with the most salient subsample of the input (i.e. very frequent words) that distributional information becomes more reliable.

Another point of interest in the present study was to compare the categorization strength of bigrams and frames separately. Frame-based distributional contexts with two context words at each side of an intervening word in the middle have been reported to be very useful for categorization. However, previous corpus studies (i.e. Cartwright & Brent, 1997; Mintz, 2003) show that frame contexts provide very high accuracy scores, since frame contexts are more restrictive than bigrams, but very low completeness scores, since grammatical categories are very unlikely to have all their members always framed by the two same context words. Thus, if a word occurs in the middle of a three-word frame, it is very likely to be a member of a particular grammatical category. However, the greater specificity of the syntactic frame entails that fewer instances of that word will occur in that frame.

On the other hand, bigram-based contexts achieve higher scores of completeness, but at the expense of getting lower scores in accuracy. Corpus studies

[16] In the present study, completeness scores among high-frequency words were a bit lower than for mid-frequency words. However, accuracy scores among high-frequency words were better than for mid-frequency words, which gave higher mean overall scores in general for high-frequency words as well.

which analyze the usefulness of distributional information strictly based on bigram contexts (e.g. Monaghan & Christiansen, 2004; Monaghan, Chater & Christiansen, 2005) provide evidence that bigram contexts are capable of categorizing many more words, and exhibit greater completeness scores, but bigram contexts are more vulnerable to categorization errors. Thus, for example, there are many more nouns which are found to be prececed by *the*, and there are far fewer nouns which are found in the central position of the frame *the__is*. However, there are also many other words which are not nouns and which can be preceded by *the* (e.g. most modifying adjectives), but the likelihood with which these other open class words can be found in the central position of the frame *the__is* is rather small.

The prediction regarding the categorization strength of the frames considered in the present study was that results would be very similar to other frame-based analyses, with perhaps even greater accuracy scores and lower completeness scores, since the frames considered here are basically restricted to English nominal plural morphology as the only element to the right of each frame. The prediction was confirmed by the results obtained. The combination of the six possible frames that were included in the present study (i.e. variables *Syn2b-Syn7b*) gave vey high accuracy scores in the type analysis as well as in the token analysis, since there were 100% of correctly classified other open class words. This indicates that there was absolutely no interference on the part of non-nominal elements to be included in nominal frames, and there was no risk that any word which is not a noun might be preceded by a determiner and followed by plural morphology. Bigram analyses, however, gave slightly lower accuracy scores, with 88.7% of correctly classified other open class word types, and 96.3% of correctly classified other open class word tokens. Thus, although accuracy was still high among bigrams, the accuracy scores obtained with frames outperformed those obtained by bigrams.

However, as initially predicted, results with frames were not good as far as completeness is concerned. In the type analysis using only frames, just 17% of the nominal types were correctly classified. The token analysis gave very similar results, with only 16.7% of correctly classified nominal tokens on the basis of frames only. On the other hand, bigram completeness scores were far higher, although they were not as high as the scores obtained from the anlayses using all distributional contexts together (i.e. both frames and bigrams). Thus, for the bigram analysis, correct classification reached a total of 62.5% among types and 32.4% among tokens.

The characteristics of English nominal inflectional morphology make it impossible to expand the set of morphological frames to include a wider variety of contexts, which indicates that the categorization system based exclusively on plural morphology is not valid or is incomplete as far as nominal categorization is concerned. However, similar analyses in languages with richer morphology and more regular nominal endings might provide better results. I will discuss this point in the last section of this chapter.

As to the particular kind of distributional contexts that appeared to be statistically the strongest, stepwise discriminant analyses with types as well as with tokens showed that variable *Syn1* (i.e. precedence of indefinite article), variables *Syn2a-Syn2b* (i.e. precedence of definite article, both with words in singular and plural) and variables *Syn4a-Syn4b* (i.e. precedence of possessive determiner both with words in singular and plural) were the ones that were found to contribute most to the classifying analysis with types, tokens in general, and all tokens in each of the three frequency groups. Therefore, in general terms, the inventory of English articles (i.e. both indefinite and definite), as well as possessive determiners appear to be the context words which accounted for most nouns in the linguistic subsamples from the four corpora under analysis. A similar categorization strength has been attributed to English articles in previous studies (e.g. Mintz, 2003; Monaghan, Chater & Christiansen, 2005), as well as for articles in other languages like French or Dutch (Monaghan, Christiansen & Chater, 2007) or Spanish (Feijóo et al., 2007). These findings are particularly interesting in the case of English articles since, unlike French, Dutch or Spanish articles, English articles do not agree in gender or number with the noun they precede. Therefore, English articles are morphologically invariable in terms of gender and number agreement features[17]. This makes them particularly helpful for young language learners as far as noun categorization is concerned given that, not only are they statistically strong and reliable, but they are fixed elements which are not subject to any morphosyntactic change triggered by the context in which they appear.

On the other hand, variables *Syn3a-Syn3b* (i.e. precedence of demonstrative determiners, both with singular and plural words), variables *Syn6a-Syn6b* (i.e. precedence of prepositions, both with singular and plural words) and variables *Syn7a-Syn7b* (i.e. precedence of *wh-* interrogative elements, both with singular and plural words) appeared to be the weakest contexts in all the analyses. In fact, variable *Syn7b* was the only one which was never entered in any of the stepwise discriminant analyes that were carried out (see chapter 6). This indicates that demonstratives as well as prepositions and *wh-* interrogative words are not very likely to be found immediately preceding nouns in the kind of linguistic exposure that English-learning infants experience, which yields those contexts not particularly useful or reliable for word categorization tasks. In fact, significance tests already revealed that variables *Syn3* were

[17] Unlike the definite article *the*, the English indefinite article has two forms *a/an*, but their distribution and use corresponds to phonological criteria, not morphological criteria.

non-significant sometimes, and variables *Syn6* and *Syn7* were more likely to describe other English open class words than nouns.

7.1.2. Phonological cues

Several previous studies have already pointed out the usefulness of phonological cues for grammatical categorization (see chapter 3 for a review). The literature suggests that, taken individually, each phonological cue does not yield very useful categorization information but, when taken together, combinations of several phonological cues have been shown to give better successful categorization results. In this line, sucessful categorization results with phonological cues have been reported from combinations of up to sixteen phonological features (Monaghan, Chater & Christiansen, 2005).

In the present dissertation, four different kinds of phonological features were taken to map words into their corresponding category (i.e. word length, consonant voicing, vowel quality and the position of stress). These four features have been reported to be the strongest for word categorization tasks. Unlike previous studies, instead of simply combining them all, a series of phonological variables were created in which all the possible combinations with the four phonological features were established. Thus, combinations ranged from total absence of any of the four phonological features, presence of just one phonological feature, combinations of two features, combinations of three features, and a final variable where all four phonological features converged. The initial prediction was that phonological variables would become stronger as long as they included longer combinations of phonological features. Then, phonological variables which only included one feature should be the weakest.

This prediction was in part confirmed by the results obtained from the analysis. Significance tests revealed that the variable which subsumed all four phonological features (i.e. *Phoncomb1234*) was found to be significant in all cases except among low- and mid-frequency words. All variables with three-feature combinations (i.e. *Phoncomb123*, *Phoncomb124* and *Phoncomb134*) were also found to be highly significant in all cases, with nouns being more likely than other open class words to contain three or four of the selected phonological features simultaneously.

However, the clusters of two phonological features, or the variables with only one phonological appear to be equally relevant for nouns as well as for other open class words. Thus, differences between nouns and other open class words with simple combinations were only significant when the phonological feature related to the position of stress was involved (i.e *Phoncomb14*, which described disyllabic trochees). This was due to the fact that most single-feature variables were found to be non-significant in all the analyses.

Thus, word length (*Phon1*) and consonant voicing in final position (*Phon2*) were never significant. Vowel quality (*Phon3*) revealed significant differences between nouns and other open class words only among types and mid-frequency words. Even then, when significant differences were found, mean proportions revealed that it was other open class words that were more likely than nouns to be described by this variable. With low-frequency words and high-frequency words, the same variable gave non-significant differences. However, when all tokens were introduced simultaneously, regardless their frequency, significant differences were restored between nouns and other open class words as fas as vowel quality is concerned[18].

[18] Note that variable *Phon4* does not exist, since the fourth phonological feature could not possibly be found alone with the kind of identification that phonological variables had in the present study: the fourth phonological feature described words with a trochaic stress pattern, while the first phonological feature described words with at least two syllables. Given a word which met the requirements of *Phon4* (i.e. a trochee), it would automatically meet the requirements of *Phon1* as well. Thus, such words were grouped under variable *Phoncomb14*.

As a result, the variable with the two-feature combination in which word length and final consonant voicing were involved (i.e. *Phoncomb12*) did not give statistically significant differences. Similarly, all the two-feature variables whose combinations contained the feature describing vowel quality (i.e *Phoncomb13* and *Phoncomb23*) did not give significant differences among high-frequency words either. Stepwise discriminant analyses also confirmed these tendencies, with three- and four-feature variables being introduced more frequently than single-feature variables.

Furthermore, the mean proportions obtained from the noun group and the other open class word group for each of the variables revealed that nouns are more likely than other open class words to be described by clusters of three or four features, while other open class words are more likely to contain just one of the selected features or none of them. This indicates that, when combined, the four phonological features selected in the present study tend to account for a greater proportion of nouns than any other word group in the English lexicon, while the opposite is true of a situation where none of the selected features are present (i.e. variable *Phon0*). The only exception to this pattern was variable *Phoncomb124*, which described disyllabic trochees whose final consonant was voiced and, as a variable containing a cluster of three features, it was nevertheless more likely to describe other open class words than nouns.

As mentioned in chapter 6, an advantage of other open class words over nouns as far as this variable is concerned might be due to the kind of naturalistic data that was used in the present study. Previous analyses which describe the phonological characteristics of nouns and other open class words are mainly based on lexical roots (Kelly, 1992; 1996). As mentioned before, the present study takes naturalistic data as a starting point, with fully inflected forms as each child would hear them. Within the other open class word group, there was a higher proportion of verbs than of adjectives and adverbs. Verbs are more likely than nouns to receive syllabic inflections, which would make word length increase in verbs, but not in nouns. Furthermore, those inflections tend to contain a final voiced consonant and, what is more, morphological inflections are always weak and they never take the primary stress of the word, which indicates that the lexical root that receives the inflection is very likely to have a trochaic stress pattern.

Take, for example, some verbal forms such as *bite*, *write*, *beat*, *eat*, *speak*, *take* or *break*. In their bare infinitive form, they are all monosyllabic, which rules out the possibility of getting a score in *Phon1* and *Phon4* simultaneously; they all end in a voiceless consonant (i.e. /t/ or /k/, respectively), which makes it impossible for them to score under *Phon2*; their stressed vowel is never low, since they either contain a diphthong in their stressed syllable (i.e. /aɪ/ or /eɪ/) or the high front vowel /i:/, which makes them not meet the requirements of *Phon3*. Therefore, all of them would be grouped under variable *Phon0*, following the tendency of most other open class words. However, when these verbs are inflected and take their *–ing* form (i.e. *biting*, *writing*, *beating*, *eating*, *speaking*, *taking* and *breaking*), or even their past participle form (i.e. *bitten*, *written*, *beaten*, *eaten*, *spoken*, *taken* and *broken*), they suddenly become disyllabic words with the stress on their first syllable, and their final consonant is voiced. This makes them be re-grouped under variable *Phoncomb124*, once they are inflected. Cases like these ones, and the fact that inflected forms are more frequent than non-inflected ones, might have been the cause of the advantage of other open class words over nouns as far as variable *Phoncomb124* is concerned.

Regarding the overall success with which phonological variables together would help to identify English nouns in their corresponding grammatical category, the present study predicted lower correct classification scores among phonological features than among the distributional contexts described above, suggesting that, although they are useful, phonological cues are weaker than distributional cues. This prediction was also validated by the results described in the previous chapter: when the analysis was run with types, overall correct classification using phonological cues was 53.9%, while it reached 79.9% with distributional cues. A similar pattern emerged among all tokens together, with an overall correct classification of 57.3% of tokens when phonological cues were used, and an overall correct classification of 67.4% of tokens when distributional cues were used.

However, unlike the results with distributional cues, and although overall correct classification with phonological cues reached scores above chance levels with types as well as with tokens, the distribution of correct classification was uneven among both groups under examination (i.e. nouns vs. other open class words). Thus, the analysis with types gave an overall correct classification above chance levels mainly because there were relatively high accuracy scores (i.e. 71.6% of other open class word types were correctly classified), but there were quite low completeness scores (i.e. only 40.1% of noun types were correctly classified). The opposite pattern was true of tokens, which reached an overall correct classification above chance levels mainly because there were very high completeness scores (84.9% of noun tokens were correctly classified) but there were very low accuracy scores (i.e. only 22.1% of other open class word tokens were correctly classified as such, and the rest were misclassified as nouns).

This indicates that, when word frequency is ignored and types are considered, the classificatory system established in the present analysis based on the selected phonological cues classifies few words of all word classes in general. This is good for other open class words since, given a set of variables which is meant to define the noun category, most other open class words should ideally fall out of this noun category and should not respond to any of the selected variables. Therefore, few other open class words being correctly classified as such by the selected variables means very high accuracy scores as far as the noun group is concerned. This is what happened and what gave such good accuracy scores. However, not only few other open class words were correctly classified, which makes the analysis successful in part, but also very few nouns were equally correctly classified, which gives low completeness scores.

The opposite is true when word frequency is considered and the analysis is run with tokens. In this case, many words of all word categories in general are correctly classified. This is good for the noun group, as it indicates that the selected phonological variables indeed gather around most of the nominal elements in the English lexicon. This gives very good completeness scores, indicating that very few nouns would be misclassified or would fall out of a noun category based on phonological information as described by the selected variables. However, the selected phonological variables would account for many of the other open class words as well, and many of them would be incorrectly misclassified as nouns, which gave very low accuracy scores.

This pattern was confirmed when the same analysis using phonological cues was run among each of the three different frequency groups. The results with low-frequency words were very similar to the results obtained from the type analysis, when token frequency was not considered: there were very high accuracy scores (i.e. 76.7% of low-frequency other open class word tokens were correctly classified), but there were quite low completeness scores (i.e. only 31.4% of low-frequency nouns were correctly classified). As word frequency increased, accuracy scores progressively decreased, while completeness scores progressively improved. Thus, while there was an advantage of accuracy scores over completeness scores among low-frequency words, these results were levelled among mid-frequency words, with nearly identical results in both accuracy and completeness (i.e. there were 57.2% of mid-frequency nouns and 55.6% of mid-frequency other open class words which were correctly classified). Among high-frequency words, the pattern obtained with low-frequency words is finally reversed, with very good completeness scores (i.e. there were 81.4% of correctly classified high-frequency nouns) but very low accuracy scores (i.e. there were only 36.4% of correctly classified and highly frequent other open class words).

Despite the low accuracy scores among high-frequency words or the low completeness scores among low-frequency words, overall correct classification using phonological information was always above chance levels for all frequency groups. As reported in the previous chapter, there was an overall correct classification of 50.7% of low-frequency words, 56.5% of mid-frequency words and 59.5% of high-frequency words. Therefore, phonological information can still be claimed to be useful for the grammatical categorization of English nouns, although to a lesser degree than distributional information.

Furthermore, phonological information shows a pattern which is similar to the one found with distributional information across different frequency groups, since overall correct classification increases as word frequency also does. Thus, phonological information seems to be more reliable for grammatical categorization with high-frequency nouns. These results are parallel to previous studies which report the usefulness of phonological information for word class identification only with highly frequent words (i.e. Kelly, 1992; 1996; Soreno & Jongman, 1990). Besides, stepwise discriminant analyses also confirmed the fact that individual phonological variables are more useful among high-frequency words, since single-feature variables were only entered in the stepwise analysis with high-frequency words, while clusters of multiple-feature variables were entered in the other stepwise analyses.

However, in a study using a comparable set of phonological variables than in the present analysis (i.e. one phonological variable connected to word length, one to vowel height, one to consonant quality and one connected to word stress position), Durieux & Gillis (2001) found the opposite pattern, with poorer performance of phonological cues for high-frequency items. They attributed these better results among low-frequency words to the fact that there were proportionally more nouns in their low-frequency group than in the other frequency sets, and nouns were classified with greatest accuracy by the kind of phonological information they considered in their analyses. This uneven distribution of words from different categories in each frequency group was due to the fact that their sample selection was carried out at random.

However, in the present analysis, groups were weighted so as to keep an even proportion of nouns and other open class words in each of the frequency groups as well as in the whole corpus (see table 6.5 in section 6.2.2.3). When this happens and grammatical categories have approximately the same size across different frequency groups, phonological information becomes again more useful for the categorization of high-frequency items than low-frequency items.

Monaghan, Chater & Christiansen (2005) also found that phonological information was more useful for the categorization of low-frequency items than high-frequency items. Furthermore, they claimed that their results were not due to the greater frequency of nouns in the lower-frequency groups, as suggested by Durieux & Gillis (2001), as their groups were compensated in their analysis as well. Furthermore, they also found an opposite pattern regarding distributional information, with better correct classification scores among high-frequency words. Therefore, they claimed that phonological and distributional information contribute differently towards grammatical categorization, since each source of information is useful with a different set of words:

when phonological information is accurate, distributional information is not, and viceversa.

Evidence for such an interaction between distributional and phonological information was not found in the present study. Rather, the results obtained from the present analysis reveal that distributional and phonological information do not contribute differently towards grammatical categorization, but their contribution is parallel and both work in the same direction (i.e. towards more successful categorization of high-frequency words). Thus, while the results regarding the interaction between distributional information and word frequency in the present analysis are similar to those in Monaghan, Chater & Christiansen (2005), the results regarding the interaction between word frequency and phonological information are not.

This might be due to the fact that the methodology employed in both analyses is not directly comparable. While the distributional variables that were entered in the discriminant analyses in both studies are very similar, the phonological variables are not. Monaghan, Chater & Christiansen (2005) used a different number of phonological features (i.e. 16, and not 4) and the phonological variables were defined in a different way (i.e. they did not create multiple-feature variables). The scoring system was also different.

Furthermore, their analysis was aimed at testing the successful contribution of phonological information towards the grammatical categorization of nouns and verbs and, for this purpose, they used phonological features which are typically associated to nouns, as well as a set of phonological features which are typically associated to verbs. Thus, their results among low-frequency items include both accuracy scores among nouns and verbs simultaneously.

However, in a further study with exactly the same variables and the same encoding system, Christiansen & Monaghan (2006) aimed at establishing whether phonological and distributional cues were differentially useful for nouns and verbs. They found that phonological cues contributed to greater accuracy of classifying verbs, whereas distributional cues were more reliable among nouns. They claimed that verbs had more variation in the kind of syntactic contexts in which they could occur, which required greater consistency in the phonological information that related to their grammatical category.

Therefore, the interaction between phonological and distributional information as far as word frequency is concerned which was found in Monaghan, Chater & Christiansen (2005) might have been affected by the presence of verbs and verb-categorizing phonological variables in that study, with consequently overall better classification scores, especially among low-frequency words. Thus, the success with which verbs appear to be classified by phonological information might have affected these results. When only nouns, and the phonological variables that categorize them are considered, as in the present study, phonological variables become more reliable as word frequency increases, and phonological and distributional information cohere with a similar categorization pattern.

7.1.3. Semantic cues

As seen in previous chapters, a further objective in the present analysis was to analyze the potential usefulness of semantic information as far as the categorization of English nouns is concerned. Traditional accounts on noun categorization based on semantic information have put forward the idea that young language learners might group all nouns together under the semantic label of "object" and all verbs together under the semantic label of "action" (Pinker, 1984). The fact that most nouns refer to common objects and their subsequent imageability based on notional grounds has been argued to be the reason why nouns are learned before verbs or before words encoding actions and relations in language development (Bassano, 2000; Gentner, 1982).

However, more recent findings show that nominal elements denoting common objects dominate children's early vocabularies only as far as type productivity is concerned (Gopnik & Choi, 1995; Nelson, 1995; Nelson et al., 1993). However, reports on early vocabulary production show that there are more tokens of non-nominal expressions (i.e. verbs and relational words such as *there*, *up* or *no*) than of nouns, and that this tendency is not only true of the English language (Gopnik & Choi, 1995).

Thus, as these authors suggest, the makeup of children's early lexicons might not be the result of there being a more "learnable" or more "imaginable" category in semantic terms, but it might be a reflection of children's actual linguistic experience, in such a way that children's first words might be instantiations of the kind of words that their parents used with them. This is also coherent with the results obtained from the input analysis undertaken in the present study. As seen in the previous chapter (see table 6.1), the descriptive data that was obtained reveal that the kind of linguistic experience to which the four children under consideration are exposed contains far more nominal types than non-nominal types (i.e there were overall 5,388 noun types and 4,233 other open class word types). However, when it comes to tokens, other open class words exceed nominal tokens by far (i.e. the four corpora together contained 51,577 nominal tokens, but 88,047 other open class word tokens).

The predominance of noun types over other open class word types has been explained by the fact that many very complex and abstract entities are realized as nouns in the adult language (Nelson, 1995). This raises the question of how many of these nominal types are actually "easily learnable". If children engage in word categorization tasks guided by the fact that all words that refer to common objects are grouped together under the noun category, then how many of the overall 5,388 noun types in this study can actually be described by these semantic features and how salient is that proportion in statistical terms?

Previous analyses of linguistic input addressed to young language learners have shown that common object nouns represent only a very small proportion of all the noun repertory that children hear (Nelson, 1995; Nelson et al., 1993). The initial prediction as far as this study was concerned was that many of the nouns to which very young English-learning children are exposed refer to basic-level common objects, and will be subsumed by variable *Sem2*. Other nouns were expected to either refer to proper names of people (i.e. *Sem1* nouns) or to non-discrete mass entities (i.e. *Sem3* nouns). Neither of them would pose any learning problems either. However, an important number of nominal elements were also predicted to lack any of the above-mentioned semantic characteristics (i.e. *Sem4* nouns) and lack the necessary semantic components that would guide children in their categorization tasks, provided they perform such tasks on the basis of semantic information.

These predictions are confirmed by the data obtained in the analysis. The descriptive data from the study show that only approximately half of the noun types as well as half of the nominal tokens refer to basic-level object nouns and were described by variable *Sem2*, while the other half belonged to the other three semantic subsets. Within those, about a third of the noun types and a quarter of the noun tokens were subsumed by variable·*Sem4*, which was the one which grouped together all nouns which did not have any of the semantic features that would foster correct grammatical categorization.

Nevertheless, in terms of significance, the results from the test within the noun group (i.e. the Wilcoxon signed-rank test) showed that, in spite of there being several instantiations of XBLOC nouns or *Sem4* nouns, differences between nouns which had appropriate semantic content (i.e. *Sem2* nouns) and nouns which did not (i.e. *Sem4* nouns) were still highly significant. Mean scores of basic-level object nouns were higher than non-basic-level object nouns in the analysis with types as well as with tokens. Within frequency groups, differences were significant as well, but higher means were obtained from the non-basic-level object noun group among mid-frequency words, indicating that within this frequency range, words which lack appropriate semantic content are actually more typical than basic-level object nouns.

A further objective in this study was to see, in the same way as with distributional and phonological information, whether there was an overlap between nouns and other open class words as far as semantic information is concerned, and to test whether the overlap was significant enough so as to bring about a considerable misclassification of elements. Provided that, according to Pinker (1984), children engage in a semantic analysis of the input and make the hypothesis that all words that denote actions belong to the grammatical category of verbs, what do children do when they encounter action words which are not verbs? And what do they do with verbs that do not denote actions? Does the input offer a high proportion of contradictory information of this kind?

The present analysis shows that, indeed, there is a slight degree of semantic overlap between nouns and other open class words, since some of the XBLOC nouns have certain semantic features which are typical of verbs (i.e. mainly nouns that denote actions). At the same time, obviously not all of the other open class words denoted actions since, besides non-action verbs, adjectives and adverbs were also included in this group. However, taking the semantic feature of "action", differences were still highly significant between nouns and other open class words, with action words being better represented by the other open class word group than by the noun group.

Thus, in the case of semantic information, the kind of overlap between nouns and other open class words seems to be one way, that is, some of the nouns might lack the corresponding nominal semantic features, or might have verbal semantic features, and might therefore be misclassified as other open class words (i.e. mainly misclassified as verbs). However, the same risk does not seem to hold for any of the other open class words, since they are very unlikely to contain any of the semantic features associated to nouns, and thus be misclassified as such on the basis of semantic information. Empirical analyses of children's early vocabularies also suggest that, even when the very same word can be both a noun and a verb (e.g. English *kiss*, *hug*, *call*, *help*) most children assign those action words exclusively to the verb category, regardless of their parents' frequent use of them as nouns (Nelson, 1995).

This makes word classification based on semantic information different from classification based on distributional or phonological information. As seen so far, distributional and phonological analyses reveal that there is a two-way overlap between nouns and other open class words: some nouns lack their corresponding phonological features or distributional context and are misclassified as other open class words (i.e. completeness scores are affected) while, at the same time, some other open class words are found within nominal distributional contexts or contain nominal phonological features and are misclassifed as nouns (i.e. accuracy scores are affected). If word classification based on semantic information produces one-way overlap across different categories only (i.e. nouns overlap with other open class words but other open class words do not overlap with nouns), then completeness scores are expected to suffer, but not accuracy scores.

The discriminant analyses using semantic variables that were performed on all types and tokens as well as frequency groups confirm this, since perfect accuracy scores were obtained in all cases (i.e. there were always 100% of correctly classified other open class words). However, completeness scores were affected in all cases, due to the fact that some of the nouns in the corpora did not contain the corresponding nominal semantic features (i.e. they were classified under variable *Sem4*, since they did not conform to the basic-level object description and were not proper nouns or mass nouns either) or they contained semantic features which are associated to verbs[19].

Thus, for the type analysis, only 67.9% of the nominal types were correctly classified, and these completeness scores dropped dramatically in the token analysis, with only 39.9% of correctly classified nominal tokens. This analysis with all tokens together using semantic variables alone gave, in fact, the worst completeness scores obtained in the whole of the present study. Completeness scores within frequency groups showed a slight improvement, with 61.6% of correctly classified low-frequency tokens and 71.5% of correctly classified mid-frequency tokens. But completeness dropped again to 63.8% among high-frequency tokens.

The fact that correct classification scores were far better with types than with tokens confirms the tendency described above in connection to early vocabulary production, where higher productivity of nominal types is observed, while tokens from other categories outnumber nominal tokens. Thus, children's early word production can

[19] Recall that, for discriminant analysis purposes, variables *Sem4a* (i.e. action words) and *Sem4b* (i.e. other XBLOC words) were united under a unique *Sem4* variable, which grouped together all words which did not contain any of the selected nominal semantic features (i.e. neither proper nouns, nor basic-level objects or mass nouns). Thus, as far as discriminant analyses with semantic variables are concerned, variable *Sem4* stands here as variables *Syn0a* or *Phon0* did in distributional and phonological analyses, respectively.

be seen as a direct reflection of the kind of linguistic environment that they have experienced. However, nouns cannot be claimed to be "easily learnable" on the basis of their semantic association to basic-level objects alone, since statistical analyses where the diagnosticity of such semantic classification was tested provided a considerable proportion of misclassified nouns. In the case of tokens, correctly classified nouns were merely at chance levels and completeness scores were very low.

In any case, the basic-level association is still strong and can account for many of the English nouns to which children are exposed. In fact, variable *Sem2*, which corresponded to basic-level common count nouns, occupied the first position in all stepwise analyses with semantic variables. However, in spite of it being the strongest of all semantic variables from a statistical point of view, it is not sufficient in terms of the overall correct classification of English nouns. These findings suggest that semantic information may not always be the only factor which is used to determine the assignment of words to their grammatical category. The next section discusses the way in which semantic information and other sources of information interact with one another towards word categorization.

7.2. Cues in interaction

The empirical evidence discussed so far suggests that the linguistic input to which English-learning children are exposed contains reliable cues for the grammatical categorization of a high proportion of nouns on the basis of either syntax or phonology or semantics alone. However, learners never hear speech with just a single kind of cue to word categorization in isolation. In children's natural linguistic environment, there are multiple redundant language-specific cues to word category membership. How statistically strong is each source of information for word categorization tasks, as opposed to the other sources of information which are also available?

As stated in chapter 4, when different kinds of cues were put to interact with one another, overall successful categorization scores were expected to improve when compared to the results obtained with cues in isolation. Furthermore, different sources of information were predicted to interact with one another in such a way that each of them would complement the other two.

These predictions are also confirmed by the results obtained from the present study. The analyses performed with cues in interaction have shown that the three types of cues available contribute differently toward word categorization. Thus, in the interaction between distributional and phonological information, overall correct word classification was more successful than with distributional or phonological cues alone. Therefore, both accuracy and completeness scores were kept high, since the low completeness scores or the low accuracy scores obtained in some of the analyses with phonological cues alone were improved by the contribution of distributional cues. In this way, overall correct classification using both phonology and syntax improved and scores were even higher than with each of the cues alone. There was also a positive interaction with word frequency, since the success of the combination of syntax and phonology was greater as word frequency increased, and the differences across each of the frequency groups were greater using both kinds of cues than with distributional cues alone.

The interaction between semantic and distributional cues also gave more successful results than each of these kinds of cues in isolation. This time, semantic information contributed in providing successful accuracy scores, while distributional cues contributed in providing greater completeness scores, so that the low completeness scores obtained from semantic cues alone were improved.

Noun categorization also benefited from the interaction between phonology and semantics in all the analyses performed. In this case, semantic information contributed with very high accuracy scores, which were never affected by any of the low accuracy scores obtained from the analysis using phonological cues in isolation. In turn, phonological cues contributed by improving the low completeness scores obtained in the analyses with semantic cues in isolation, so that the product of the interaction resulted in much better completeness scores.

The best scores in the whole of the present study were obtained when distributional, phonological and semantic variables were put to interact with one another. The outcome of the interaction of the three possible sources of information gave almost perfect scores in the analysis with types with an accuracy score of 97.0% and a completeness score of 90.1% of nouns. Accuracy scores were better in the analysis with all tokens (i.e. there were 98.3% of correctly classified other open class words), but completeness scores dropped to only 49.1% of correctly classified nouns. However, as mentioned before, regular patterns from words are more likely to be extracted and learned once they are found to be recurrent enough as far as type frequency is concerned, rather than token frequency (Bybee, 1995; Marchman & Bates, 1994). In this way, the high completeness scores obtained from the type analysis would compensate for the low completeness scores with tokens.

Furthermore, completeness scores also improved considerably when different frequency ranges were considered. The total amount of correctly classified nouns reached 89.3% among low-frequency words, 92.7% among mid-frequency words and 75.9% among high-frequency words. Accuracy scores were also very high, ranging from 91.9% among low-frequency words, to 97.0% among high-frequency words. Then, with frequency groups as well, the combination of cues across the three different modalities provided better results overall than using either type of cue in isolation.

The contribution of each cue type was not always additive, given the fact that, as stepwise discriminant analyses revealed, some of the cues that were discarded in the combined discriminant analyses were nevertheless entered in the separate analyses, while some other cues were only entered in the combined analyses and not in the separate ones. This is especially true of phonological variables. For example, variables *Phoncomb134* (i.e. the combination of word length, vowel height and stress position) or *Phoncomb1234* (i.e. the combination of all four phonological features) were entered in most of the discriminant analyses using phonological cues alone, but were only entered in one of the combined discriminant analyses (i.e. the one with mid-frequency words). On the other hand, phonological variables such as *Phoncomb14* (i.e. the combination of word length and stress position) were hardly ever used in any of the analyses with phonological cues alone, but were often entered in the combined analyses.

In the rest of the cases, the cues contributed additively in the present analysis. Furthermore, the stepwise discriminant analyses using all three different kinds of cues provided evidence suggesting that each cue type contributed differently towards categorization with different groups of words. This shows that distributional, phonological and semantic information did not completely overlap, but they complemented one another. The overall sucessful categorization scores obtained with each of the frequency groups showed that distributional and phonological information improved their categorization strength as word frequency increased, while the opposite is true of semantic information, with lower successful categorization scores with high-frequency words.

This pattern is confirmed by considering the types of variables that were entered at each step in the different stepwise analyses with each frequency group. For low-frequency words, semantic variables dominated the analysis, while phonological and distributional variables showed a smaller contribution. Proportionally, distributional variables were the ones that contributed less: only four out of the fifteen possible distributional variables were entered in this low-frequency word stepwise analysis (i.e. a proportion of 0.26), while four out of the twelve possible phonological variables were used (i.e. a proportion of 0.33).

With mid-frequency words, a similar pattern emerges, although more steps are used and better accuracy and completeness scores are obtained. As to the kind of cues that were found to be statistically strong enough so as to contribute to the classificatory system, semantic variables still dominated the analysis, while phonological and distributional variables were weaker, albeit there is an increasing strength of phonological and distributional variables, with six variables of each kind introduced.

With high-frequency words, although all four semantic variables are still used, their statistical classificatory strength regarding the discriminant function diminishes in favour of distributional variables, which then share the first steps of the classificatory system with semantic variables. Furthermore, the increasing strength exhibited by distributional variables with high-frequency words is not only indicated by the order in which they appear in the stepwise analysis, but also in the number of distributional variables that are found to be statistically strong enough so as to contribute to the discriminant function, since a total of nine distributional variables were used. Phonological variables, on the other hand, show their weakest contribution with high-frequency words, since only three variables were entered, and they occupied the last positions in the analysis.

Monaghan, Chater & Christiansen (2005) have already proposed the *Phonological-Distributional Coherence Hypothesis* in their analysis with distributional and phonological cues. According to them, both sources of information contribute differently towards word classification, since better classification for high-frequency items is achieved thanks to distributional contexts, while better classification for the low-frequency items was found in the analyses with phonological cues. Recall that, although these results are not directly comparable to the present study, such a correlation between phonological cues and word frequency was not found in the present study. Unlike in the study by Monaghan, Chater & Christiansen (2005), accurate classification of items using phonological cues did not decrease as word frequency increased, but rather the opposite: phonological cues showed a classification pattern which was parallel to that of distributional cues, since classification scores became better as word frequency increased (for a detailed discussion, see section 7.1.2. in this chapter).

However, the evidence provided by the stepwise analyses using the three possible sources of information in the present analysis revealed that, indeed, although successful classification of phonological cues did not decrease among high-frequency words, the contribution of phonological cues in the combined analysis using all variables with high-frequency words was minimal, as it was restricted to the mere presence of three cues in the last steps of a total sequence of sixteen steps. Therefore, the results from the present study provide evidence for the complementarity and the different contribution of distributional, phonological and semantic information.

7.3. General discussion

The data presented in this study provide empirical evidence against the *Argument of the Poverty of the Stimulus* as far as the categorization of English nouns is concerned. The analysis has undertaken a close examination of a subsample of the Manchester corpus (Theakston et al., 2001) available from the CHILDES database, which currently stands as the best available approximation to children's linguistic environment. The main conclusion that can be drawn from this examination is that the linguistic input to which English-learning children are exposed contains information which is reliable and sufficient enough so that children can build a formal category for nouns without the need of assuming any *a priori* innate linguistic knowledge about the nature of grammatical categories.

In particular, the nouns in the linguistic subsamples analyzed in this study contain enough inherent internal properties (i.e. both semantic and phonological features) as well as external properties (i.e. they reliably and consistently occur in a series of recurrent distributional morphosyntactic contexts) that bind them together. Therefore, given a learner who is sensitive to all these features, a noun grammatical category can be formed on the basis of these three sources of information.

Furthermore, the three sources of information available have been shown to interact with one another in such a way that they complement each other and happen to categorize different elements. Thus, while word-internal properties like semantic and phonological features appear to be more successful with low-frequency words, word-external properties like morphosyntax and distributional information appear to be better for higher-frequency words. This poses a difficult chicken-and-egg problem as far as word categorization is concerned: do children use the external distributional properties of words for initial categorization in order to further exploit word-internal phonological and semantic regularities, or do they use phonology and semantics to perform an initial classification of speech elements upon which distributional analyses are conducted?

As seen in chapter 3, proponents of the *Semantic Bootstrapping* approach argue that the initial categorization that children carry out is performed on the basis of semantic information (Brown, 1973; Grimshaw, 1981; Macnamara, 1982; Pinker, 1984). Under this approach, a set of words is initially classified according to their meaning. After that, children analyze the grammatical properties of these words and use distributional information to categorize all words which do not conform to the semantic pattern. Therefore, morphosyntactic distributonal contexts are later on gradually adopted as classifying devices (Macnamara, 1982).

However, as seen before, an initial categorization analysis based entirely on semantic information alone would bring about many miscategorization errors on the part of language learners, due in part to the problem of multicategoriality of many English words (Nelson, 1995; Nelson et al., 1993) and also to the nature of grammatical categories themselves (see Labelle, 2005 for a review). The empirical evidence obtained in this study also reveal that semantic information alone can account for a subset of English nouns, but not all of them.

Furthermore, unlike previous studies like Grimshaw (1981) or Pinker (1984), the present study does not assume that semantic concepts and the grammatical categories into which they map are innate and, therefore, available to the children from the onset of the language learning process. Rather, the literature suggests that the cognitive principles that enable children to make the most out of the semantic cues available in the input mostly become accessible during their second year of life (Clark, 2009; Golinkoff et al., 1994; Karmiloff & Karmiloff-Smith, 2001; Tomasello, 2003). As Demuth (1992) suggests, "[a]ccess to the semantics of the system becomes available only at later stages of development, whereas early overgeneralizations are normally of a phonological nature" (Demuth 1992: 630). Naigles (2002) holds a similar view, suggesting that learning form is easy for children, but the mapping of form and meaning is harder and develops later.

Nevertheless, sensitivity to distributional patterns in the linguistic input has been shown to be present in infants by the second half of their first year of life. At about the age of eight months, children have been shown to be able to track distributional statistical patterns both from artificial languages (Aslin et al., 1998; Saffran et al., 1996a; Thiessen & Saffran, 2003) as well as natural languages (Pelucchi et al., 2009). Such abilities have been reported to be useful for word categorization (Gerken, Wilson & Lewis, 2005). Besides, sensitivity to the kind of phonological features considered in the present study (i.e. word length, individual phoneme recognition or stress patterns) has been found in very young infants in their first year of life as well (Johnson & Jusczyk, 2001; Jusczyk, 1997; Kuhl, 2000; Shi, Werker & Morgan, 1999).

Thus, children are more likely to establish an initial rudimentary classification of linguistic items in terms of the internal phonological makeup of words as well as the external distributional contexts in which each word type occurs, especially considering the fact that such sources of information (i.e. phonology and syntax) are especially reliable among recurrent higher-frequency words. This initial categorization might guide further subsequent analyses and might actually help children make the most out of other sources of information like semantics.

Studies in languages with grammatical gender distinctions have also shown that sucessful word categorization can proceed on the basis of phonological and distributional information only, even in cases when grammatical gender assignment is not governed by word meaning and, therefore, semantic information becomes irrelevant for the task. Such preference for morphophonological cues over semantic cues has been found in artificial language learning studies (Brooks et al., 1993; Frigo & McDonald, 1998) as well as in natural languages such as French (Desrochers, Paivio, & Desrochers, 1989; Matthews, 2010), Spanish (Pérez-Pereira, 1990), German (Mills, 1986), Italian (Bates et al., 1995), Hausa (Corbett, 1991) or Bantu languages (Suzman, 1996).

Other researchers have also suggested that children's initial categorization might be guided by distributional and phonological information as well, and that "(...) perceptual analyses of linguistic input help to bootstrap semantic analyses, which in turn bootstrap syntactic analyses. The child passes through a series of stages marked by use of increasingly diverse and grammatically sophisticated forms of information". (Morgan, Shi & Allopenna 1996: 280)

Given the kind of reliability that each source of information has been shown to exhibit in the present study, children's initial noun category might primarily consist of a small subsample of highly frequent nouns which might have been learned on the basis of distributional information with the help of phonological information. Although distributional cues were found to be more reliable than phonological cues in this study, phonological cues should not be discarded as they might guide word categorization to a great extent.

To start with, the results obtained from the analyses using both distributional and phonological cues were better than the ones obtained from the analyses with distributional cues alone. Besides, recall that the kind of distributional contexts which were selected for the present study were exclusively made up of closed-class items. As seen in chapter 3, phonological cues have been shown to be very useful for initial word categorization and the distinction between open class words (i.e. nouns, verbs, adjectives and adverbs) and closed class items (i.e. determiners, prepositions, conjunctions, etc.). In fact, as far as the open vs. closed class distinction is concerned, phonological information has been reported as far more useful and accurate than distributional information (Cutler, 1993; Monaghan, Chater & Christiansen, 2005; Shi, 1995; Shi, Morgan & Allopena, 1998).

Thus, phonological information might guide the initial parsing of speech between lexical vs. grammatical elements, which might help children become familiar with the kind of words that make up the kind of distributional contexts selected here. With this information, children might be able to exploit distributional cues to a great extent and they might be able to use them for the categorization of open class words like nouns. Additional phonological features which correlate with nouns might further assist this initial categorization. Subsequently, once the category might have been created with a subset of high-frequency nouns, analogical extension on the part of the language-learning children might gradually guide the further classification of other lower-frequency items. For this task, semantic information might be considerably useful, as semantic cues were found to be especially reliable among mid-frequency words. Such final interaction of all three sources of information might ultimately help children to successfully categorize most, if not all of the nouns to which they are exposed.

7.4. Limitations and further research

The present analysis provides evidence on the kind of information which can be found in the input and from which children might be able to infer a noun category without assuming any innate knowledge that might guide children's noun categorization. However, nouns are not the only word class that needs to be built in order for syntax acquisition to proceed. Further research will be needed in order to see the way the results obtained in the present study for nouns can be generalized to other grammatical categories. Crucially, verbs have been said to occur in a greater variety of distributional contexts, and the kinds of words that precede verbs are less predictable than those that precede nouns. Conversely, verbs have been shown to be more accurately described by a set of phonological features than nouns (Christiansen & Monaghan, 2006; Monaghan, Chater, & Christiansen, 2005). The kind of semantic information that verbs can carry is also more complex than that of nouns. A study in which verbal distributional, phonological and semantic variables are analyzed will ultimately reveal the accuracy and usefulness of each of these sources of information for the categorization of verbs, as well as the way in which they interact.

Furthermore, the objective of the present study was merely to work out the overall degree of accuracy with which nouns are likely to be categorized on the basis of the input alone. With the data that has been obtained, the conclusion can be drawn that there is enough evidence in the input from which to classify nominal elements into their right grammatical category, and this evidence has been shown to be reliable, sufficiently represented in the input, and available to the child learner.

However, the mere fact that cues to grammatical categories are available and consistently represented in the input does not prove that children are actually sensitive to such cues and use them when categorizing words and, ultimately, when learning their language. As seen before, a number of studies have already shed light into early child sensitivity to both phonological and distributional cues to perform linguistic tasks such as word segmentation (see section 7.3). However even if children are sensitive to distributional and phonological cues in general, and to the kind of distributional and phonological cues selected in this study in particular, there is no evidence that they actually use them for the task of word categorization. And if they do, additional research is needed in order to analyze the actual way in which they use them.

Are all three sources of information available and readily used simultaneously from the beginning? Is one used after another? If so, what is the exact developmental pattern which children undergo? I have hypothesized that distributional and phonological cues might be initially useful for children, and word categorization might later be improved and expanded by means of using semantic information. However, this is only a hypothesis. A further experiment will be needed in which 2½-year-old English-learning children are trained with the variables under consideration and then tested on the way they deal with such variables and the way they treat novel words which contain the set of selected cues.

Last, but not least, further research is needed in order to see the way in which the results obtained here for English nouns generalize to other languages as well. If language structure is not innate, then children must be equipped with the relevant learning mechanisms to make the most out of the linguistic input to which they are exposed and extract the relevant information from which to infer grammatical structure. Such learning mechanisms should be equally applicable to any kind of data and should enable children to cope with stimuli from just any natural language, not only English.

Several previous studies have examined the usefulness of certain phonological cues (Durieux & Gillis, 2001; Fisher & Tokura, 1996; Monaghan, Christiansen & Chater, 2007; Shi, Morgan & Allopenna, 1998) as well as distributional cues (Chemla et al., 2009; Feijóo et al., 2008; Monaghan, Christiansen & Chater, 2007; Redington et al., 1995) for word categorization in languages other than English. However, only Monaghan, Christiansen & Chater (2007) have looked at the interaction between distributional and phonological cues for word categorization across different languages, and no study has undertaken the interaction between phonology, syntax and semantics for word categorization in different languages. Thus, while semantic features might be somehow universal (although the actual mapping of semantic features into certain grammatical categories might be language-specific), phonological and distributional cues are certainly different across different languages, and so is the way in which those cues interact with one another as well as the way they interact with word frequency.

Crucially, for languages with richer inflectional morphology than English, the kind of morphosyntactic frames considered in the present study might be much more useful than the combination of determiners and the single plural morpheme in English. Furthermore, if distributional cues appear to be more powerful in other languages than in English, then it might follow that phonological features in those languages are less regular and worse predictors of grammatical category membership. On the other hand, a language with weaker distributional contexts might show consistent and reliable phonological correlates to grammatical categories. Further research will be needed in order to test these hypotheses.

Chapter 8:
Conclusion

This dissertation provides empirical evidence against the *Argument from the Poverty of the Stimulus*, given that the data obtained suggest that the linguistic environment to which English-learning children are exposed contains enough information to assign nouns to their corresponding grammatical categories. In particular, English child-directed speech contains a constellation of distributional, phonological and semantic cues which reliably encode the necessary information to allow syntactically naïve learners to group all and only the nouns they hear into the same category on the basis of their formal similarities. Consequently, it is certainly unwise to completely dismiss the linguistic input as a potential source of information for the acquisition of the noun category in English.

As stated in chapter 1, proponents of representational innateness within the generative grammar approach suggest that the core aspects of the grammatical structure of all natural languages are part of what they label as *Universal Grammar*. Such a view entails two main assumptions. First, the grammatical structure of natural languages is claimed to be completely modular and independent from the phonological inventory of language or even from their semantic or pragmatic constraints. The architecture of syntax is generally represented in terms of rules, and it is understood as being autonomous and independent from phonology and the lexicon. Second, the set of rules that make up the syntactic structure of natural languages is claimed to be universal (i.e. common to all linguistic systems) and innately specified, given the fact that these rules could not possibly be learned from experience, mainly because grammatical features are underrepresented in speakers' output.

The present dissertation provides evidence against both assumptions. To start with, syntactic aspects of language might not be completely independent from phonology or the lexicon. The present study provides compelling evidence that phonological and semantic information contribute to a great extent to a very basic syntactic procedure, namely the grammatical categorization of English nouns. Therefore, according to the evidence provided in this study, the English nominal category should be best defined according to a series of syntactic, phonological and semantic criteria together.

Then, the data obtained here from the analysis of parental speech do not give support to the rule-based nominal category which is conceived within generative grammar approaches, where nouns are defined as [+N, –V] elements in a strictly formal and binary manner. It is very difficult to see the way in which the elements from the real language samples analyzed in the present study would fit into the [+N, –V] formal rule. As we have seen, a number of nominal elements to which English-learning children are exposed are syntactically marked as nouns (i.e. they are preceded by articles) but they are semantically marked as verbs (i.e. they denote actions). Some others lack the relevant nominal semantic features, while they can still be classified as nouns on the basis of distributional information. Some others have a phonological form which is most typically found among other open class words, and still some others are not found within the prototypical distributional contexts, although they can still be classified as nouns on the basis of semantic and/or phonological information. Such data would never be totally captured by the two-feature algebraic formula proposed by generativists. As Kelly (1992) suggested:

> [g]rammatical categories should be defined probabilistically rather than discretely through rule-based criteria. In such a view, the categories of nounhood and verbhood would be organized around prototypes that are formed from semantic, morphological, and phonological predictors of grammatical class (...). Hence, the noun and verb classes would grade into one another, with no sharp boundary distinguishing them. (Kelly 1992: 359)

Under such an account, grammatical categories do not have rigid boundaries, but they rather exhibit a prototype structure (Kelly, 1992; Taylor, 1989). Thus, words do not need to carry all of the properties of a category in order to be correctly classified. For example, words lacking some of the relevant phonological properties may nevertheless be accurately classified in terms of semantics and/or syntax and viceversa. Furthermore, if grammatical categories are not understood as binary groups but as probabilistically defined word classes, and if this is the kind of word class that language learners should ultimately come up with, the set of features that make up such word classes is clearly and reliably represented in input speech.

This takes us back to the idea of representational innateness defended by generative grammarians. As said before, nativist approaches claim that the core aspects of grammatical structure of natural languages (i.e. those aspects which belong to *Universal Grammar*) are innate, since it is impossible to learn them on the basis of the impoverished input that constitutes children's linguistic environment. If there is something universal at all within all existing natural languages, this must be the fact that all languages have nominal elements. Thus, the grammatical category of noun is probably the best candidate for children's presumably existing genetic endowment. If so, and if the *Argument of the Poverty of the Stimulus* were right, then the set of features that make up the noun category as a whole should be underrepresented in parental speech adressed to children. However, such a claim does not hold for the kind of data considered in the present study, suggesting that the English nominal category can, in fact, be learned from input-driven data without the need of postulating any *a priori* innately specified linguistic knowledge about the nature of such category.

Consequently, if there is anything innate at all in the development of grammatical categories such as the noun category, it must be the kind of learning mechanisms that allow children to process, benefit from and deal with the kind of linguistic cues that are present in the input and which make up the set of defining criteria for nouns. This brings us again to the question formulated in chapter 1, with which I started the present discussion: what does "innate" really mean?

Nativist accounts of linguistic skills outside representational innateness approaches claim that, although actual grammatical categories and structure might be learned, they are learned by innate mechanisms which, as such, are both domain-specific and species-specific. Thus, as stated in the introduction, the main claim under these assumptions is that language learning is mainly accomplished by means of mechanisms which are exclusively devoted to the linguistic domain and are uniquely found among the human species (Chomsky, 1975; Fodor, 1983; Lenneberg, 1967).

However, empirical evidence from research does not provide support for such a claim either. As seen in previous chapters, from the very early stages of language development children already exhibit statistical learning abilities as well as sensitivity to phonological and semantic features. Such skills allow children to exploit the linguistic cues in the input for grammatical categorization purposes. If such learning mechanisms are innate, then they must be unique of the linguistic domain on the one hand, and of the human species on the other hand.

Nevertheless, a number of studies provide evidence that such learning mechanisms are used by children in tasks other than language learning and, furthermore, they are used at the same stages of cognitive development, parallel to the language learning tasks (Bates, 1994; Saffran, 2002; Saffran & Thiessen, 2007). For example, empirical research on cognitive development show that learning in such different domains as both music and language is, at least in part, subserved by the same domain-general learning processes or mechanisms (McMullen & Saffran, 2004; Saffran et al., 1999). Similar results have been found when comparing linguistic skills with domain-general cognitive skills using visual stimuli (Fiser & Aslin, 2001; Kirkham, Slemmer & Johnson, 2002; Younger, 1993).

What is more, similar learning mechanisms have been found not to be specifically human. In particular, when nonhuman primates (e.g. cotton-top tamarin monkeys) are tested using similar procedures, they show the same statistical learning abilities as human infants (Hauser, Newport & Aslin, 2001; Ramus et al., 2000). Bonobos have been shown to possess impressive word recognition skills as well (Shanker, Savage-Rumbaugh & Taylor, 1999). Furthermore, sensitivity to phonological cues and prosodic patterns which parallel human abilities have also been found among other species, like rats (Toro, Trobalon & Sebastián-Gallés, 2005), or chinchillas (Kuhl & Miller, 1975).

Thus, animals might possess far more sophisticated learning abilities than were previously hypothesized. Therefore, human infants are not a privileged species neither at computing transitional probabilities of sequential units in language, nor at detecting linguistic prosodic regularities. On the contrary, the kind of learning mechanisms which are needed to perform one of the fundamental tasks within language development, namely word categorization, does not seem to be domain- or species-dependent. Therefore, an important component of linguistic development does not seem to have its roots in any kind of biological adaptation, which entails that grammar might not be, after all, the product of human philogeny. Consequently, the phylogenetic components of human linguistic abilities should be reconsidered.

The present dissertation has provided sound evidence against one of the most important theoretical arguments according to which human language is presumably unlearnable (i.e the *Argument of the Poverty of the Stimulus*). Further research on children's actual learning abilities for linguistic categorization, as well as for categorization in other cognitive domains, will provide a better understanding of the *logical* problem of language acquisition.

References:

Anderson, S. R. & D. W. Lightfoot. 1999. "The human language faculty as an organ". *Annual Review of Physiology* 62: 697-722.

Anderson, S. R. & D. W. Lightfoot. 2002. *The Language Organ: Linguistics as Cognitive Physiology*. Cambridge: Cambridge University Press.

Aslin, R. N., J. R. Saffran & E. L. Newport. 1998. "Computation of conditional probability statistics by 8-month-old infants". *Psychological Science* 9: 321-324.

Aslin, R. N., J. Z. Woodward, N. P. LaMendola & T. G. Bever. 1996. "Models of word segmentation in fluent maternal speech to infants". In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Morgan, J.L. & K. Demuth (eds.). Mahwah, NJ: Erlbaum. 117-134.

Baldwin, D. A. 1991. "Infants' contribution to the achievement of joint reference". *Child Development* 62: 875-890.

Baldwin, D. A. 1993. "Infants' ability to consult the speaker for clues to word reference". *Journal of Child Language* 20: 395-418.

Bassano, D. 2000. "Early development of nouns and verbs in French: Exploring the interface between lexicon and grammar". *Journal of Child Language* 27: 521-559.

Bates, E. 1994. "Modularity, domain specificity and the development of language". *Discussions in Neuroscience* 10: 136-149.

Bates, E., A. Devescovi, L. Pizzamiglio, S. D'Amico & A. Hernandez. 1995. "Gender and lexical access in Italian". *Perception and Psychophysics* 58: 992-1004.

Bates, E., V. Marchman, D. Thal, L. Fenson, P. Dale, J. S. Reznick, J. Reilly & J. Hartung. 1994. "Developmental and stylistic variation in the composition of early vocabulary". *Journal of Child Language* 21: 85-124.

Baker, M. C. 2001. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. New York: Basic Books.

Baker, M. C. 2003. *The Lexical Categories: Verbs, Nouns, and Adjectives*. Cambridge: Cambridge University Press.

Benedict, H. 1979. "Early lexial development: comprehension and production". *Journal of Child Language* 6: 183-200.

Bernstein Ratner, N. & B. Rooney. 2001. "How accessible is the lexicon in Motherese?". In *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Weissenborn, J. & B. Höhle (eds.). Amsterdam: John Benjamins. 71-78.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, Massachusetts: MIT Press.

Bloom, P. 2001. "Roots for word learning". In *Language Acquisition and Conceptual Development*. Bowerman, M. & S. C. Levinson (eds.). Cambridge: Cambridge University Press. 159-184

Bouchard, D., C. Dubuisson & A. M. Parisot. 2005. "Categories in Quebec Sign Language: Reflections on categorization across modalities". In *Handbook of Categorization in Cognitive Science*. Cohen, H. & C. Lefebvre (eds.). Amsterdam: Elsevier. 381-399.

Bowerman, M. & S. Choi. 2001. "Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories". In *Language Acquisition and Conceptual Development*. Bowerman, M. & S. C. Levinson (eds.). Cambridge: Cambridge University Press. 475-510.

Bowerman, M. & S. C. Levinson. 2001. *Language Acquisition and Conceptual Development*. Cambridge: Cambridge University Press.

Braine, M. D. S., R. E. Brody, P. J. Brooks, V. Sudhalter, J. A. Ross, L. Catalano & S. M. Fisch. 1990. "Exploring language acquisition in children with a miniature artificial language. Effects of item and pattern frequency, arbitrary subclasses, and correction". *Journal of Memory and Language* 29: 591-610.

Brent, M. R. 1994. "Acquisition of subcategorization frames using aggregated evidence from local syntactic cues". *Lingua* 92: 433-470.

Brent, M. R. 1996. "Advances in the computational study of language acquisition". *Cognition* 61: 1-38.

Brent, M. R. & T. A. Cartwright. 1996. "Distributional regularity and phonotactic constraints are useful for segmentation". *Cognition* 61: 93-125.

Brooks, P. B., M. D. S. Braine, L. Catalano, R. E. Brody & V. Sudhalter. 1993. "Acquistion of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning". *Journal of Memory and Language* 32: 79-95.

Brown, R. 1957. "Linguistic determinism and part of speech". *Journal of Abnormal and Social Psychology* 55: 1-5.

Brown, R. 1973. *A First Language: The Early Stages*. Cambridge, Massachusetts: Harvard University Press.

Bruner, J. 1983. *Child's Talk*. New York: Norton.

Bybee, J. 1995. "Regular morphology and the lexicon". *Language and Cognitive Processes* 10: 425-455.

Bybee, J. 1998. "The emergent lexicon". In *Proceedings of the Chicago Linguistics Society* 34. Gruber, M. C., D. Higgins, K. S. Olson, & T. Wysocki (eds.). Chicago: Chicago Linguistics Society. 421-435.

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

Caselli, M. C., P. Casadio & E. Bates. 1999. "A comparison of the transition from first words to grammar in English and Italian". *Journal of Child Language* 26: 69-111.

Cassidy, K. W. & M. H. Kelly. 1991. "Phonological information for grammatical category assignments". *Journal of Memory and Language* 30: 348-369.

Cassidy, K. W. & M. H. Kelly. 2001. "Children's use of phonology to infer grammatical class in vocabulary learning". *Psychonomic Bulletin and Review* 8: 519-523.

Chemla, E., T. H. Mintz, S. Bernal & A. Cristophe. 2009. "Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquistion strategies". *Developmental Science* 12(3): 396-406.

Chomksy, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1975. *The Logical Structure of Liguistic Theory*. Plenum Press, New York.

Chomsky, N. 1981. "Lectures on government and binding: The pisa lectures". Dordrecht: Foris Publications.

Chomsky, N. 1993. *Language and Thought*. London: Moyer Bell.

Christiansen, M. H. & P. Monaghan. 2006. "Discovering verbs through multiple-cue integration". In *Action Meets Word: How Children Learn Verbs*. Golinkoff, R. M. & K. Hirsh-Pasek (eds.). New York: Oxford University Press. 88-107.

Clark, E. V. 2009. *First Language Acquisition*. Cambridge: Cambridge University Press.

Cohen, H. & C. Lefebvre (eds.). 2005. *Handbook of Categorization in Cognitive Science*. Amsterdam: Elsevier.

Cole, R. A. & J. Jakimik. 1980. "A model of speech perception". In *Perception and Production of Fluent Speech*. Cole, R. A. (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates. 133-163.

Corbett, G. 1991. *Gender*. Cambridge, UK: Cambridge University Press.

Crain, S. & D. Lillo-Martin. 1999. *An Introduction to Linguistic Theory and Language Acquisition*. Oxford: Blackwell.

Crain, S. & P. Pietroski. 2001. "Nature, nurture and Universal Grammar". *Linguistics and Philosophy* 24: 139-186.

Croft, W. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Croft, W. 2003. *Typology and Universals*. Cambridge, UK: Cambridge University Press.

Culicover, P. W. 1999. *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford: Oxford University Press.

Cutler, A. 1990. "Exploiting prosodic probabilities in speech segmentation". In *Cognitive Models of Speech Processing*. Altmann, G. T. M. (ed.). Cambridge, MA: MIT Press. 105-121.

Cutler, A. 1993. "Phonological cues to open- and closed-class words in the processing of spoken sentences". *Journal of Psycholinguistic Research* 22: 109-131.

Cutler, A. 1996. "Prosody and the word boundary problem". In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Morgan, J.L. & K. Demuth (eds.). Mahwah, NJ: Erlbaum. 87-99.

Cutler, A. & D. M. Carter. 1987. "The predominance of strong initial syllables in the English vocabulary". *Computer Speech and Language* 2: 133-142.

Deacon, T. W. 1998. *The Symbolic Species: The Co-evolution of Language and the Human Brain*. London: Penguin.

Demirdache, H. & L. Matthewson. 1995. "On the universality of syntactic categories". In *Proceedings of the Northeastern Linguistics Association* (NELS 25). GLSA, Amherst, MA. 79-94.

Demuth, K. 1992. "The acquisition of Sesotho". In *The Crosslinguistic Study of Language Acquisition*, vol. 3. Slobin, D. I. (ed.). Hillsdale NJ: Lawrence Erlbaum. 557-638.

Desrochers, A., A. Paivio, & S. Desrochers. 1989. "L'effect de la fréquence d'usage des noms inanimés et de la valeur predictive de leur terminaison sur l'identification du genre gramatical". *Revue Canadienne de Psychologie* 43: 62-73.

Dixon, R. M. W. 1977. "Where have all the adjectives gone?" *Studies in Language* 1: 1-80.

Dockrell, J. & J. McShane. 1990. "Young children's use of phrase structure and inflectional information in form-class assignments of novel nouns and verbs". *First Language* 10: 127-140.

Durieux, G. & S. Gillis. 2001. "Predicting grammatical classes from phonologiacl cues: An empirical test". In *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*, Weissenborn. J. & B. Höhle (eds.). Amsterdam: John Benjamins. 189-229.

Echols, C. H. 2001. "Contributions of prosody to infants' segmentation and representation of speech". In *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Weissenborn, J. & B. Höhle (eds.). Amsterdam: John Benjamins. 25-46.

Echols, C. H. & E. L. Newport. 1992. "The role of stress and position in determining first words". *Language Acquisition* 2: 189-220.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi & K. Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Farmer, T. A., M. H. Christiansen & P. Monaghan. 2006. "Phonological typicality influcences lexical processing". In *Proceedings of the National Academy of Sciences* 103: 12203-12208.

Feijóo, S. 2007. "Little Red Riding Hood and the chomskyan wolf: A debate on poverty of the stimulus arguments". *BELLS* 16: 1-14.

Feijóo, S., E. Serrat, C. Muñoz, J. Serrano. 2008. "Frequent frames and noun categorization in Spanish child-directed speech", poster presented at the XI International Congress for the Study of Child Language, Edinburgh, UK.

Fiser, J. & R. N. Aslin. 2001. "Unsupervised statistical learning of higher-order spatial structures from visual scenes". *Psychological Science* 12: 499-504.

Fisher, C. & H. Tokura. 1996. "Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence". *Child Development* 67: 3192-3218.

Fitneva, S. A., M. H. Christiansen & P. Monaghan. 2009. "From sound to syntax: Phonological constraints on children's lexical categorization of new words". *Journal of Child Language* 36: 967-997.

Fodor, J. 1983. *Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Frigo, L. & J. L. McDonald. 1998. "Properties of phonological markers that affect the acquisition of gender-like subclasses". *Journal of Memory and Language* 39: 218-245.

Gelman, S. A. & M. Taylor. 1984. "How two-year-old children interpret proper and common names for unfamiliar objects". *Child Development* 55: 1535-1540.

Gentner, D. 1982. "Why nouns are learned before verbs: Linguistic relativity vs. natural partitioning." In *Language Development: Language, Culture and Cognition*. S. A. Kuczaj II (ed.). Hillsdale, NJ: Erlbaum. 301-334.

Gerken, L. A. 1996. "Phonological and distributional information in syntax acquisition". In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Morgan, J.L. & K. Demuth (eds.). Mahwah, NJ: Erlbaum. 411-425.

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

Gerken, L. A., B. Landau, & R. E. Remez. 1990. "Function morphemes in young children's speech perception and production". *Developmental Psychology* 27: 204-216.

Gerken, L. A. & B. J. McIntosh. 1993. "Interplay of function morphemes and prosody in early language". *Developmental Psychology* 29: 448-457.

Gerken, L. A., R. Wilson & W. Lewis. 2005. "Infants can use distributional cues to form syntactic categories". *Journal of Child Language* 32: 249-268.

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

Gleitman, L. & H. Gleitman. 2001. "Bootstrapping a first vocabulary". In *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Weissenborn, J. & B. Höhle (eds.). Amsterdam: John Benjamins. 79-96.

Gleitman, L., & E. Wanner. 1982. "Language acquisition: The state of the art". In *Language Acquisition: The State of the Art*. Gleitman, L. & E. Wanner (eds.). New York: Cambridge University Press. 3-48.

Golinkoff, R. M., K. Hirsh-Pasek & M. A. Schweisguth. 2001. "A reappraisal of young children's knowledge of grammatical morphemes". In *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Weissenborn, J. & B. Höhle (eds.). Amsterdam: John Benjamins. 167-188.

Golinkoff, R. M., C. B. Mervis & K. Hirsh-Pasek. 1994. "Early object labels: A case for a developmental lexical principles framework. *Journal of Child Language* 21: 125-155.

Gopnik, A., & S. Choi. 1995. "Names, relational words, and cognitive development in English and Korean speakers: Nouns are not always learned before verbs". *In Beyond Names for Things: Young Children's Acquisition of Verbs*. Tomasello, M. & W. E. Merriman (eds.). Mahwah, NJ: Lawrence Erlbaum Associates. 63- 80.

Grimshaw, J. 1981. "Form, function and the language acquisition device". In *The Logical Problem of Language Acquisition*. Baker, C. L. & J. McCarthy (eds.). Cambridge, MA: MIT Press. 163-182.

Hanitriniaina, S. & L. Travis. 1998. "Unparsing and f-nominals in Malagasy". Paper presented at the CLA Conference (University of Ottawa).

Harnad, S. 2005. "To cognize is to categorize: cognition is categorization". In *Handbook of Categorization in Cognitive Science*. Cohen, H. & C. Lefebvre (eds.). Amsterdam: Elsevier. 19-43.

Hauser, M. D., E. L. Newport & R. N. Aslin. 2001. "Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins". *Cognition* 78: 53-64.

Hilferty, J. 2003. *In Defense of Grammatical Constructions*, Ph. D. Dissertation, Universitat de Barcelona.

Jackendoff, R. 1977. *X-Bar Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.

Jackson-Maldonado, D., D. Thal, V. Marchman, E. Bates & V. Gutierrez-Clellen. 1993. "Early lexical development in Spanish-speaking infants and toddlers". *Journal of Child Language* 20: 523-549.

Johnson, E. K. & P. W. Jusczyk. 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. 1997. *The Discovery of Spoken Language*. Cambridge, MA: The MIT Press.

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

Jusczyk, P. W. & R. N. Aslin. 1995. "Infants' detection of sound patterns of words in fluent speech". *Cognitive Psychology* 29: 1-23.

Jusczyk, P. W., A. Cutler & N. Redanz. 1993. "Preference for the predominant stress patterns of English words". *Child Development* 64: 675-687.

Jusczyk, P. W., D. M. Houston & N. Redanz. 1999. "The beginnings of word segmentation in English-learning infants". *Cognitive Psychology* 39: 159-207.

Kail, M. 2000. "Acquisition syntactique et diversité linguistique". In *L'acquisition du Langage*, vol. 2. Kail, M. & M. Fayol (eds.). Paris: Presses Universitaires de France. 9-44.

Karmiloff, K. & A. Karmiloff-Smith. 2001. *Pathways to Language: From Fetus to Adolescent.* Massachusetts: Harvard University Press.

Katz, N., E. Baker & J. Macnamara. 1974. "What's in a name? On the child's acquisition of proper and common names". *Child Development* 45: 269-273.

Kelly, M. H. 1988. "Rhythmic alternation and lexical stress differences in English". *Cognition* 30: 107-137.

Kelly, M. H. 1992. "Using sound to solve syntactic problems: The role of phonology in grammatical category assignments". *Psychological Review* 99: 349-364.

Kelly, M. H. 1996. "The role of phonology in grammatical category assignments". In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* Morgan, J.L. & K. Demuth (eds.). Mahwah, NJ: Erlbaum. 249-262.

Kelly, M. H. & J. K. Bock. 1988. "Stress in time". *Journal of Experimental Psychology: Human Perception and Performance* 14: 389-403.

Kelly, M. H. & S. Martin. 1994. "Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition and language". *Lingua* 92: 105-140.

Kelly, M. H. & S. Martin. 1995. "Phonological cues to grammatical class". Unpublished manuscript, University of Pennsylvania.

Kirkham, N. Z., J. A. Slemmer & S. P. Johnson. 2002. "Visual statistical learning in infancy: Evidence for a domain general learning mechanism". *Cognition* 83(2): 335-342.

Kuhl, P. K. 2000. "A new view of language acquisition". *Proceedings of the National Academy of Science* 97(22): 11850-11857.

Kuhl, P. K. & J. D. Miller. 1975. "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants". *Science* 190: 69-72.

Labelle, M. 2005. "The acquisition of grammatical categories: The state of the art". In *Handbook of Categorization in Cognitive Science*. Cohen, H. & C. Lefebvre (eds.). Amsterdam: Elsevier. 433-457.

Laurence, S. & E. Margolis. 2001. "The poverty of the stimulus argument". *British Journal of Philosophy of Science* 52: 217-276.

Laurence, S. & E. Margolis. 2002. "Radical concept nativism". *Cognition* 86: 25-55.

Lenneberg, E. 1967. *Biological Foundations of Language*. New York: Wiley.

Lieven, E. 2006. "Variation in first language development". In *Encyclopedia of Language and Linguistics*, (2nd ed.), Volume XX. Brown, K. (ed.). Oxford: Elsevier. 350-354.

Lyons. J. 1977. *Semantics*. Cambridge: Cambridge University Press.

McMullen, E. & J. R. Saffran. 2004. "Music and language: A developmental comparison". *Music Perception* 21(3): 289-311.

Macnamara, J. 1982. *Names for Things: A Study of Child Language*. Cambridge, Massachusetts: MIT Press.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Marantz, A. 1995. "The Minimalist Program". In *Government and Binding and the Minimalist Program*. Webelhuth, G. (ed.). Oxford: Basil Blackwell. 349-382.

Maratsos, M. 1998. "The acquisition of grammar". In *Handbook of Child Psychology*, vol. 2. Kuhn, D. & R. S. Siegler (eds.). New York: Wiley. 421-466.

Maratsos, M. 1999. "Some aspects of innateness and complexity in language acquisition". In *The Development of Language*. Barret, M. (ed.). Hove: Psychology Press. 191-228.

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.

Maratsos, M. & M. A. Chalkley. 1980. "The internal language of children's syntax: The ontogenesis and representation of syntactic categories". In *Children's Language*, vol. 2. Nelson, K. (ed.). New York: Gardner Press. 127-214.

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.

Matthews, C. 2010. "On the nature of phonological cues in the acquisition of French gender categories: Evidence from instance-based learning models". *Lingua* 120(4): 879-900.

Meyers, L. S., G. Gamst & A. J. Guarino. 2006. *Applied Multivariate Research*. London: SAGE Publications.

Miller, G. 1956. "The magical number seven, plus or minus two: some limits on our capacity for processing information". *Psychological Review* 63: 81-97.

Mills, A. E. 1986. *The Acquisition of Gender: A Study of English and German*. Berlin: Springer-Verlag.

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., E. L. Newport, & T. G. Bever. 2002. "The distributional structure of grammatical categories in speech to young children". *Cognitive Science* 26: 393-424.

Monaghan, P., N. Chater, & M. H. Christiansen. 2003. "Inequality between the classes: Phonological and distributional typicality as predictors of lexical processing". In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates. 963-968.

Monaghan, P., N. Chater, & M. H. Christiansen. 2005. "The differential contribution of phonological and distributional cues in grammatical categorisation". *Cognition* 96: 143-182.

Monaghan, P. & M. H. Christiansen. 2004. "What distributional information is useful and usable in language acquisition?" In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Monaghan, P., M. H. Christiansen & N. Chater. 2007. "The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition". *Cognitive Psychology* 55: 259-305.

Morgan, J. L. 1986. *From Simple Input to Complex Grammar: Learning, Development, and Conceptual Change*. Cambridge, MA: MIT Press.

Morgan, J. L. & K. Demuth (eds.) 1996. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah, NJ: Erlbaum.

Morgan, J. L., R. Shi & P. Allopenna. 1996. "Perceptual bases of rudimentary grammatical categories: toward a broader conceptualization of bootstrapping". In *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Morgan, J. L. & K. Demuth (eds.). Mahwah, NJ: Erlbaum. 262-293.

Naigles, L. R. 2002. "Form is easy, meaning is hard: Resolving a paradox in early child language". *Cognition* 86: 157-199.

Nelson, K. 1985. *Making Sense: The Acquisition of Shared Meaning*. New York: Academic Press.

Nelson, K. 1995. "The dual category problem in the acquisition of action words". In *Beyond Names for Things: Young Children's Acquisition of Verbs*. Tomasello, M. & W. E. Merriman (eds.). Mahwah, NJ: Lawrence Erlbaum Associates. 223-250.

Nelson, K., J. Hampson & L. K. Shaw. 1993. "Nouns in early lexicons: evidence, explanations and implications". *Journal of Child Language* 20: 61-84.

Newmeyer, F. J. 1998. *Language Form and Language Function*. Cambridge, MA: MIT Press (Bradford).

Pelucchi, B., J. F. Hay & J. R. Saffran. 2009. "Statistical learning in a natural language by 8 month-old infants". *Child Development* 80(3): 674-685.

O'Grady, W. 1997. *Syntactic Development*. Chicago: Chicago University Press.

Onnis, L., P. Monaghan, K. Richmond, & N. Chater. 2005. "Phonology impacts segmentation in speech processing". *Journal of Memory and Language* 53: 225-237.

Pérez-Pereira, M. 1990. "¿Cómo determinan los niños la concordancia de género?: Refutación de la teoría del género natural". *Infancia y Aprendizaje* 50: 73-91.

Peters, A. M. 1997. "Language typology, prosody, and the acquisition of grammatical morphemes". In *The Cross-Linguistic Study of Language Acquisition*, vol. 5. Slobin, D. I. (ed.). Mahwah, NJ: Lawrence Erlbaum Associates. 135-199.

Pinker, S. 1984. *Language Learnability and Language Development*. Cambridge, MA: MIT Press.

Pinker, S. 1987. "The bootstrapping problem in language acquisition". In *Mechanisms of Language Acquisition*. MacWhinney, B. (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates. 399-442.

Pinker, S. 1994. *The Language Instinct: How the Mind Creates Language*. New York: William Morrow and Company.

Pinker, S. 1997. "Acquiring language". *Science* 276: 1177-1181.

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

Pullum, G. K. 1996. "Learnability, hyperlearning, and the poverty of the stimulus". In *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society: Parasession on the Role of Learnability in Grammatical Theory*. Johnson, J., M. L. Luge, & J. L. Moxley (eds.). Berkeley, CA: Berkeley Linguistics Society. 498-513.

Pullum, G. K. and B. Scholz. 2002. "Empirical assessment of stimulus poverty arguments". *The Linguistic Review* 19: 9-50.

Quine, W. 1960. *Word and Object*. Cambridge, MA: Harvard University Press.

Radford, A. 1995. "Phrase structure and functional categories". In *The Handbook of Child Language*. Fletcher, P. & B. MacWhinney (eds.). Oxford: Basil Blackwell. 483-507.

Radford, A. 1996. "Towards a structure-building model of acquisition". In *Generative Perspectives on Language Acquisition*. Clahsen, H. (ed.). Amsterdam: John Benjamins. 43-89.

Ramus, F., M. D. Hauser, C. Miller, D. Morris & J. Mehler. 2000. "Language discrimination by human newborns and by cotton-top tamarin monkeys". *Science* 288: 349-351.

Ravid, D. 1995. *Language Change in Child and Adult Hebrew*. Oxford: Oxford University Press.

Reali, F., M. H. Christiansen, & P. Monaghan. 2003. "Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration". In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates. 970-975.

Redington, M., N. Chater & S. Finch. 1998. "Distributional information: a powerful cue for acquiring syntactic categories". *Cognitive Science* 22: 435-469.

Redington, M., N. Chater, C. Huang, L. P. Chang, S. Finch, & K. Chen. 1995. "The universality of simple distributional methods: Identifying syntactic categories in Chinese". *Proceedings of the Cognitive Science of Natural Language Processing*. Dublin.

Rispoli, M. 1999. "Functionalist accounts of the process of first language acquisition". In *Handbook of Child Language Acquisition*. Ritchie, W. C. & T. K. Bhatia (eds.). San Diego, CA: Academic Press. 221-246.

Rosch, E. 1978. "Principles of categorization". In *Cognition and Categorization*. Rosch, E. & Lloyd, B.B. (eds.). Hillsdale: Lawrence Erlbaum Associates. 27-48.

Rumelhart, D. E. & J. L. McClelland. 1987. "Learning the past tense of English verbs: Implicit rules or parallel distributed processing". In *Mechanisms of Language Acquisition*. MacWhinney, B. (ed.). Hillsdale, NJ: Lawrence Erlbaum. 195-248.

Sachs, J., R. Brown, & R. A. Salerno. 1976. "Adults' speech to children". In *Baby Talk and Infant Speech* (special issue). W. von Raffler-Engel & Y. LeBrun (eds.). *Neurolinguistics* 5: 240-245.

Saffran, J. R. 2002. "Constraints on statistical language learning". *Journal of Memory and Language* 47: 172-196.

Saffran, J. R., R. N. Aslin & E. L. Newport. 1996a. "Statistical learning by 8-month-old infants." *Science* 274: 1926-1928.

Saffran, J. R., E. Johnson, R. N. Aslin & E. L. Newport. 1999. "Statistical learning of tone sequences by human infants and adults". *Cognition* 70: 27-52.

Saffran, J. R., E. L. Newport & R. N. Aslin. 1996b. "Word segmentation: The role of distributional cues." *Journal of Memory and Language* 35: 606-621.

Saffran, J. R., & Thiessen, E. D. 2007. "Domain-general learning capacities". In *Handbook of Language Development*. Hoff, E. & M. Shatz (eds.). Cambridge: Blackwell. 68-86.

Sampson, G. 2002. "Exploring the richness of the stimulus". *The Linguistic Review* 19: 73-104.

Schachter, P. 1985. "Part-of-speech systems". In *Language Typology and Syntactic Description*. Shopen, T. (ed.). Cambridge: Cambridge University Press. 3-61.

Scholz, B. C. and G. K. Pullum. 2002. "Searching for arguments to support linguistic nativism". *The Linguistic Review* 19: 185-223.

Scholz, B. C. & G. K. Pullum. 2006. "Irrational nativist exuberance". In *Contemporary Debates in Cognitive Science*. Stainton, R. (ed.). Oxford: Blackwell. 59-80.

Shady, M. E. 1996. *Infants' Sensitivity to Function Morphemes*. Unpublished doctoral dissertation, State University of New York at Buffalo.

Shanker, S. G., S. Savage-Rumbaugh & T. J. Taylor. 1999. "Kanzi: A new beginning". *Animal Learning and Behaviour* 27(1): 24-25.

Sherman, D. 1975. "Noun-verb stress alternation: An example of lexical diffusion of sound change in English". *Linguistics* 159: 43-71.

Shi, R. 1995. *Perceptual Correlates of Content Words and Function Words in Early Language Input*. Ph. D. Dissertation, Brown University, Providence, RI

Shi, R., J. Morgan & P. Allopenna. 1998. "Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective". *Journal of Child Language* 25: 169-201.

Shi, R., J. F. Werker & J. L. Morgan. 1999. "Newborn infants' sensitivity to perceptual cues to lexical and grammatical words". *Cognition* 27: B11-B21.

Shipley, E., C. Smith & L. Gleitman. 1969. "A study in the acquisition of language: free responses to commands". *Language* 45: 322-342.

Slobin, D. I. 2001. "Form-function relations: how do children find out what they are?". In *Language Acquisition and Conceptual Development*. Bowerman, M. & S. C. Levinson (eds.). Cambridge: Cambridge University Press. 406-449.

Snow, C. E. & C. A. Ferguson. (eds.). 1977. *Talking to Children: Language Input and Acquisition*. Cambridge: Cambridge University Press.

Soja, N. N. 1992. "Inferences about the meaning of nouns: The relationship between perception and syntax". *Cognitive Development* 7: 29-45.

Soreno, J. A. & A. Jongman. 1990. "Phonological and form class relations in the lexicon". *Journal of Psycholinguistic Research* 19: 387-404.

Sperber, D. & D. Wilson. 1986. *Relevance: Communication and Cognition*. Cambridge, MA: Harvard University Press.

Suzman, S. M. 1996. "Acquisition of noun class systems in related Bantu languages". In *Children's Language*, vol. 9. Johnson, C. E. & J. H. V. Gilbert (eds.). Mahwah, NJ: Lawrence Erlbaum. 81-104.

Swingley, D. 2005. "Statistical clustering and the contents of the infant vocabulary". *Cognitive Psychology* 50: 86-132.

Taylor, J. R. 1989. *Linguistic Categorization: Prototypes in Linguistic Theory*. Oxford: Clarendon Press.

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.

Thiessen, E. D., E. A. Hill & J. R. Saffran. 2005. "Infant-directed speech facilitates word segmentation". *Infancy* 7(1): 53-71.

Thiessen, E. D. & J. R. Saffran. 2003. "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants". *Developmental Psychology* 39: 706-716.

Tomasello, M. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Cambridge, Massachusetts: Harvard University Press.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition.* Cambridge, MA: harvard University Press.

Tomasello, M., N. Akhtar, K. Dodson & L. Rekau. 1997. "Differential productivity in young children's use of nouns and verbs". *Journal of Child Language* 24: 373-387.

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

Tomasello, M. & W. E. Merriman. 1995. *Beyond Names for Things: Young Children's Acquisition of Verbs*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tomasello, M. & R. Olguin. 1993. "Twenty-three-month-old children have a grammatical category of noun". *Cognitive Development* 8: 451-464.

Toro, J. M., J. B. Trobalon & N. Sebastián-Gallés. 2005. "Effects of backward speech and speaker variability in language discrimination by rats". *Journal of Experimental Psychology* 31: 95-100.

Travis, L. 2005. "Lexical, functional, crossover, and multifunctional categories". In *Handbook of Categorization in Cognitive Science*. Cohen, H. & C. Lefebvre (eds.). Amsterdam: Elsevier. 320-346.

Valian, V. & S. Coulson. 1988. "Anchor points in language learning: The role of marker frequency". *Journal of Memory and Language* 27: 71-86.

Waxman, S. R. & A. E. Booth. 2001. "Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives". *Cognitive Psychology* 43: 217-242.

Weissenborn, J. & B. Höhle (eds.). 2001. *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition*. Amsterdam: John Benjamins.

Younger, B. A. 1993. "Understanding category members as 'the same sort of thing': Explicit categorization in ten-month infants". *Child Development* 64: 309-320.

Zangl, R. & D. L. Mills. 2007. "Increased brain activity to infant-directed speech in 6- and 13-month-old infants". *Infancy* 11: 31-62.

Appendix A:

Distributional contexts for the *Syn* variables in every child corpus

Table 1: Distributional contexts that emerged from the corpus of Anne.

Syn0	**a**	{∅} + x
	b	{∅} + x + *-(e)s*
Syn1	**a**	{*a, an*} + x
Syn2	**a**	{*the*} + x
	b	{*the*} + x + *-(e)s*
Syn3	**a**	{*this, that, these, those*} + x
	b	{*these, those*} + x + *-(e)s*
Syn4	**a**	{*her, his, my, our, their, your*} + x
	b	{*her, his, my, our, their, your*} + x + *-(e)s*
Syn5	**a**	{*all, another, any, enough, little, more, much, no, one, other, some*} + x
	b	{*all, any, eight, enough, few, four, little, many, more, no, other, some, three, two*} + x + *-(e)s*
Syn6	**a**	{*about, after, at, by, for, from, in, into, of, on, with*} + x
	b	{*about, after, for, in, into, of, on, with*} + x + *-(e)s*
Syn7	**a**	{*what, which, whose*} + x
	b	{*what, which, whose*} + x + *-(e)s*

Table 2: Distributional contexts that emerged from the corpus of Aran.

Syn0	**a**	{∅} + x
	b	{∅} + x + *-(e)s*
Syn1	**a**	{*a, an*} + x
Syn2	**a**	{*the*} + x
	b	{*the*} + x + *-(e)s*
Syn3	**a**	{*this, that, these, those*} + x
	b	{*these, those*} + x + *-(e)s*
Syn4	**a**	{*her, his, its, my, our, their, your*} + x
	b	{*her, his, its, my, our, their, your*} + x + *-(e)s*
Syn5	**a**	{*all, another, any, both, enough, every, five, little, more, much, no, one, other, some*} + x
	b	{*all, any, both, five, forty, four, little, many, more, no, other, six, sixty, some, three, twenty, two*} + x + *-(e)s*
Syn6	**a**	{*about, after, as, at, before, between, by, for, from, in, into, of, off, on, outside, over, with*} + x
	b	{*about, as, at, for, from, in, into, of, on, over, with*} + x + *-(e)s*
Syn7	**a**	{*what, which*} + x
	b	{*what, which*} + x + *-(e)s*

Table 3: Distributional contexts that emerged from the corpus of Becky.

Syn0	**a**	{∅} + x
	b	{∅} + x + *-(e)s*
Syn1	**a**	{*a, an*} + x
Syn2	**a**	{*the*} + x
	b	{*the*} + x + *-(e)s*
Syn3	**a**	{*this, that, these, those*} + x
	b	{*these, those*} + x + *-(e)s*
Syn4	**a**	{*her, his, its, my, our, their, your*} + x
	b	{*her, his, its, my, our, their, your*} + x + *-(e)s*
Syn5	**a**	{*all, another, any, both, either, enough, four, little, more, much, no, one, other, same, some, two*} + x
	b	{*all, any, both, enough, few, five, four, little, many, more, no, other, same, six, some, two*} + x + *-(e)s*
Syn6	**a**	{*about, at, by, for, from, in, of, on, with*} + x
	b	{*about, at, for, from, of, on, with*} + x + *-(e)s*
Syn7	**a**	{*what, which*} + x
	b	{*what, which*} + x + *-(e)s*

Table 4: Distributional contexts that emerged from the corpus of Carl.

Syn0	**a**	{∅} + x
	b	{∅} + x + -*(e)s*
Syn1	**a**	{*a, an*} + x
Syn2	**a**	{*the*} + x
	b	{*the*} + x + -*(e)s*
Syn3	**a**	{*this, that, these, those*} + x
	b	{*these, those*} + x + -*(e)s*
Syn4	**a**	{*her, his, its, my, our, their, your*} + x
	b	{*her, his, its, my, our, their, your*} + x + -*(e)s*
Syn5	**a**	{*all, another, any, few, little, many, more, no, one, other, some, two*} + x
	b	{*all, any, few, five, four, little, many, more, no, other, six, some, two*} + x + -*(e)s*
Syn6	**a**	{*about, at, behind, by, for, from, in, into, of, on, with*} + x
	b	{*about, at, by, for, in, into, of, on, with*} + x + -*(e)s*
Syn7	**a**	{*what, which, whose*} + x
	b	{*what, which, whose*} + x + -*(e)s*

Appendix B:

General data from each individual corpus

Table 1: Anne.

	Total Types	Total Tokens	Type/Token Ratio	Proportion of Types	Proportion of Tokens
Total corpus	2,761	103,457	0.027	--	--
Total selected	2,457	37,538	0.065	0.89	0.36
Total nouns	1,393	14,909	0.093	0.57	0.40
Total other	1,064	22,629	0.047	0.43	0.60

Table 2: Aran.

	Total Types	Total Tokens	Type/Token Ratio	Proportion of Types	Proportion of Tokens
Total corpus	3,432	118,469	0.029	--	--
Total selected	3,213	43,281	0.074	0.94	0.37
Total nouns	1,819	15,946	0.114	0.57	0.37
Total other	1,394	26,781	0.052	0.43	0.62

Table 3: Becky.

	Total Types	Total Tokens	Type/Token Ratio	Proportion of Types	Proportion of Tokens
Total corpus	2,222	65,411	0.034	--	--
Total selected	1,992	26,804	0.074	0.90	0.41
Total nouns	1,083	8,213	0.132	0.54	0.31
Total other	899	16,777	0.054	0.45	0.63

Table 4: Carl.

	Total Types	Total Tokens	Type/Token Ratio	Proportion of Types	Proportion of Tokens
Total corpus	2,266	76,859	0.029	--	--
Total selected	1,979	35,649	0.056	0.87	0.46
Total nouns	1,093	12,509	0.087	0.55	0.35
Total other	876	21,860	0.040	0.44	0.61

Appendix C:

Descriptive data for the distributional cue analysis from each individual corpus

Table 1(a): Nouns in the corpus of Anne.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{Ø} + x (Syn0a)	584	4,293	0.136	0.42	0.29
{Ø} + x + -(e)s (Syn0b)	268	1,346	0.199	0.19	0.09
{a, an} + x (Syn1)	383	1,593	0.240	0.27	0.11
{the} + x (Syn2a)	458	2,933	0.156	0.33	0.20
{the} + x + -(e)s (Syn2b)	178	594	0.300	0.13	0.04
{this, that, these, those} + x (Syn3a)	205	687	0.298	0.15	0.05
{these, those} + x + -(e)s (Syn3b)	46	85	0.541	0.03	0.01
{POSSESSIVE} + x (Syn4a)	229	1,132	0.202	0.16	0.08
{POSSESSIVE} + x + -(e)s (Syn4b)	70	291	0.241	0.05	0.02
{QUANTIFIER} + x (Syn5a)	191	588	0.325	0.14	0.04
{QUANTIFIER} + x + -(e)s (Syn5b)	121	361	0.335	0.09	0.02
{PREPOSITION} + x (Syn6a)	151	688	0.219	0.11	0.05
{PREPOSITION} + x + -(e)s (Syn6b)	57	83	0.687	0.04	0.01
{WH- ELEMENT} + x (Syn7a)	72	177	0.407	0.05	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	28	58	0.483	0.02	0.00

Table 1(b): Other open class words in the corpus of Anne.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	1,005	20,405	0.049	0.94	0.90
{∅} + x + -(e)s (Syn0b)	0	0	--	0.00	0.00
{a, an} + x (Syn1)	82	497	0.165	0.08	0.02
{the} + x (Syn2a)	50	251	0.199	0.05	0.01
{the} + x + -(e)s (Syn2b)	0	0	--	0.00	0.00
{this, that, these, those} + x (Syn3a)	80	500	0.160	0.08	0.02
{these, those} + x + -(e)s (Syn3b)	0	0	--	0.00	0.00
{POSSESSIVE} + x (Syn4a)	35	75	0.467	0.03	0.00
{POSSESSIVE} + x + -(e)s (Syn4b)	0	0	--	0.00	0.00
{QUANTIFIER} + x (Syn5a)	144	444	0.324	0.14	0.02
{QUANTIFIER} + x + -(e)s (Syn5b)	0	0	--	0.00	0.00
{PREPOSITION} + x (Syn6a)	75	278	0.270	0.07	0.01
{PREPOSITION} + x + -(e)s (Syn6b)	0	0	--	0.00	0.00
{WH- ELEMENT} + x (Syn7a)	28	179	0.156	0.03	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	0	0	--	0.00	0.00

Table 2(a): Nouns in the corpus of Aran.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	802	4,524	0.177	0.44	0.28
{∅} + x + -(e)s (Syn0b)	236	841	0.281	0.13	0.05
{a, an} + x (Syn1)	463	2,110	0.219	0.25	0.13
{the} + x (Syn2a)	570	3,297	0.173	0.31	0.21
{the} + x + -(e)s (Syn2b)	139	414	0.336	0.08	0.03
{this, that, these, those} + x (Syn3a)	302	941	0.321	0.17	0.06
{these, those} + x + -(e)s (Syn3b)	68	208	0.327	0.04	0.01
{POSSESSIVE} + x (Syn4a)	311	1,177	0.264	0.17	0.07
{POSSESSIVE} + x + -(e)s (Syn4b)	71	242	0.293	0.04	0.02
{QUANTIFIER} + x (Syn5a)	249	804	0.310	0.14	0.05
{QUANTIFIER} + x + -(e)s (Syn5b)	141	375	0.376	0.08	0.02
{PREPOSITION} + x (Syn6a)	226	687	0.329	0.12	0.04
{PREPOSITION} + x + -(e)s (Syn6b)	70	119	0.588	0.04	0.01
{WH- ELEMENT} + x (Syn7a)	77	194	0.397	0.04	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	10	13	0.769	0.01	0.00

Table 2(b): Other open class words in the corpus of Aran.

	OTHER				
	Total Types	Total Tokens	Type/Token Ratio	Type Proportion	Token Proportion
{∅} + x (Syn0a)	1,278	24,069	0.053	0.92	0.90
{∅} + x + -(e)s (Syn0b)	0	0	--	0.00	0.00
{a, an} + x (Syn1)	123	686	0.179	0.09	0.03
{the} + x (Syn2a)	88	389	0.226	0.06	0.01
{the} + x + -(e)s (Syn2b)	0	0	--	0.00	0.00
{this, that, these, those} + x (Syn3a)	170	490	0.347	0.12	0.02
{these, those} + x + -(e)s (Syn3b)	0	0	--	0.00	0.00
{POSSESSIVE} + x (Syn4a)	61	144	0.423	0.04	0.01
{POSSESSIVE} + x + -(e)s (Syn4b)	0	0	--	0.00	0.00
{QUANTIFIER} + x (Syn5a)	150	337	0.445	0.11	0.01
{QUANTIFIER} + x + -(e)s (Syn5b)	0	0	--	0.00	0.00
{PREPOSITION} + x (Syn6a)	149	446	0.334	0.11	0.02
{PREPOSITION} + x + -(e)s (Syn6b)	0	0	--	0.00	0.00
{WH- ELEMENT} + x (Syn7a)	30	220	0.136	0.02	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	0	0	--	0.00	0.00

Table 3(a): Nouns in the corpus of Becky.

	NOUNS				
	Total Types	Total Tokens	Type/Token Ratio	Type Proportion	Token Proportion
{∅} + x (Syn0a)	516	2,797	0.184	0.48	0.34
{∅} + x + -(e)s (Syn0b)	159	419	0.379	0.15	0.05
{a, an} + x (Syn1)	319	1,295	0.246	0.29	0.16
{the} + x (Syn2a)	326	1,240	0.263	0.30	0.15
{the} + x + -(e)s (Syn2b)	74	193	0.383	0.07	0.02
{this, that, these, those} + x (Syn3a)	98	229	0.428	0.09	0.03
{these, those} + x + -(e)s (Syn3b)	32	56	0.571	0.03	0.01
{POSSESSIVE} + x (Syn4a)	171	668	0.256	0.16	0.08
{POSSESSIVE} + x + -(e)s (Syn4b)	49	205	0.239	0.05	0.02
{QUANTIFIER} + x (Syn5a)	160	415	0.386	0.15	0.05
{QUANTIFIER} + x + -(e)s (Syn5b)	105	267	0.393	0.10	0.03
{PREPOSITION} + x (Syn6a)	114	257	0.444	0.11	0.03
{PREPOSITION} + x + -(e)s (Syn6b)	36	53	0.679	0.03	0.01
{WH- ELEMENT} + x (Syn7a)	27	111	0.243	0.02	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	8	12	0.667	0.01	0.00

Table 3(b): Other open class words in the corpus of Becky.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	852	14,885	0.057	0.95	0.89
{∅} + x + -(e)s (Syn0b)	0	0	--	0.00	0.00
{a, an} + x (Syn1)	68	271	0.251	0.08	0.02
{the} + x (Syn2a)	37	174	0.213	0.04	0.01
{the} + x + -(e)s (Syn2b)	0	0	--	0.00	0.00
{this, that, these, those} + x (Syn3a)	91	320	0.284	0.10	0.02
{these, those} + x + -(e)s (Syn3b)	0	0	--	0.00	0.00
{POSSESSIVE} + x (Syn4a)	26	50	0.520	0.03	0.00
{POSSESSIVE} + x + -(e)s (Syn4b)	0	0	--	0.00	0.00
{QUANTIFIER} + x (Syn5a)	109	436	0.250	0.12	0.03
{QUANTIFIER} + x + -(e)s (Syn5b)	0	0	--	0.00	0.00
{PREPOSITION} + x (Syn6a)	61	574	0.106	0.07	0.03
{PREPOSITION} + x + -(e)s (Syn6b)	0	0	--	0.00	0.00
{WH- ELEMENT} + x (Syn7a)	13	67	0.194	0.01	0.00
{WH- ELEMENT} + x + -(e)s (Syn7b)	0	0	--	0.00	0.00

Table 4(a): Nouns in the corpus of Carl.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
{∅} + x (Syn0a)	536	5,791	0.093	0.49	0.46
{∅} + x + -(e)s (Syn0b)	154	482	0.320	0.14	0.04
{a, an} + x (Syn1)	294	1,416	0.208	0.27	0.11
{the} + x (Syn2a)	353	1,844	0.191	0.32	0.15
{the} + x + -(e)s (Syn2b)	96	300	0.320	0.09	0.02
{this, that, these, those} + x (Syn3a)	123	435	0.283	0.11	0.03
{these, those} + x + -(e)s (Syn3b)	21	33	0.636	0.02	0.00
{POSSESSIVE} + x (Syn4a)	163	595	0.274	0.15	0.05
{POSSESSIVE} + x + -(e)s (Syn4b)	56	244	0.230	0.05	0.02
{QUANTIFIER} + x (Syn5a)	167	456	0.366	0.15	0.04
{QUANTIFIER} + x + -(e)s (Syn5b)	87	207	0.420	0.08	0.02
{PREPOSITION} + x (Syn6a)	94	339	0.277	0.09	0.03
{PREPOSITION} + x + -(e)s (Syn6b)	32	62	0.516	0.03	0.00
{WH- ELEMENT} + x (Syn7a)	46	287	0.160	0.04	0.02
{WH- ELEMENT} + x + -(e)s (Syn7b)	8	18	0.444	0.01	0.00

Table 4(b): Other open class words in the corpus of Carl.

	OTHER				
	Total Types	Total Tokens	Type/Token Ratio	Type Proportion	Token Proportion
{∅} + x (Syn0a)	820	19,673	0.042	0.94	0.90
{∅} + x + -(e)s (Syn0b)	0	0	--	0.00	0.00
{a, an} + x (Syn1)	73	555	0.132	0.08	0.03
{the} + x (Syn2a)	53	298	0.178	0.06	0.01
{the} + x + -(e)s (Syn2b)	0	0	--	0.00	0.00
{this, that, these, those} + x (Syn3a)	92	305	0.302	0.11	0.01
{these, those} + x + -(e)s (Syn3b)	0	0	--	0.00	0.00
{POSSESSIVE} + x (Syn4a)	20	60	0.333	0.02	0.00
{POSSESSIVE} + x + -(e)s (Syn4b)	0	0	--	0.00	0.00
{QUANTIFIER} + x (Syn5a)	125	465	0.269	0.14	0.02
{QUANTIFIER} + x + -(e)s (Syn5b)	0	0	--	0.00	0.00
{PREPOSITION} + x (Syn6a)	55	353	0.156	0.06	0.02
{PREPOSITION} + x + -(e)s (Syn6b)	0	0	--	0.00	0.00
{WH- ELEMENT} + x (Syn7a)	14	151	0.093	0.02	0.01
{WH- ELEMENT} + x + -(e)s (Syn7b)	0	0	--	0.00	0.00

Appendix D:

Descriptive data for the phonological cue analysis from each individual corpus

Table 1(a): Nouns in the corpus of Anne.

	NOUNS				
	Total Types	Total Tokens	Type/Token Ratio	Type Proportion	Token Proportion
No phonological cues (***Phon0***)	198	2,687	0.074	0.14	0.18
Two syllables or more (***Phon1***)	55	264	0.208	0.04	0.02
Final voiced consonant (***Phon2***)	211	2,697	0.078	0.15	0.18
Low stressed vowel (***Phon3***)	91	1,084	0.084	0.07	0.07
Two syllables or more + Final voiced consonant (***Phoncomb12***)	56	180	0.311	0.04	0.01
Two syllables or more + Low stressed vowel (***Phoncomb13***)	33	150	0.220	0.02	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	121	1,555	0.078	0.09	0.10
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	69	1,540	0.045	0.05	0.10
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	51	446	0.114	0.04	0.03
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	266	1,832	0.145	0.19	0.12
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	82	1,219	0.067	0.06	0.08
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	160	1,255	0.127	0.11	0.08

Table 1(b): Other open class words in the corpus of Anne.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	209	8,827	0.024	0.20	0.39
Two syllables or more (***Phon1***)	26	342	0.076	0.02	0.02
Final voiced consonant (***Phon2***)	182	4,396	0.041	0.17	0.19
Low stressed vowel (***Phon3***)	94	1,769	0.053	0.09	0.08
Two syllables or more + Final voiced consonant (***Phoncomb12***)	38	361	0.105	0.04	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	19	131	0.145	0.02	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	76	814	0.093	0.07	0.04
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	28	934	0.030	0.03	0.04
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	12	170	0.071	0.01	0.01
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	239	3,798	0.063	0.22	0.17
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	34	258	0.132	0.03	0.01
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	107	829	0.129	0.10	0.04

Table 2(a): Nouns in the corpus of Aran.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	259	2,954	0.088	0.14	0.19
Two syllables or more (***Phon1***)	81	384	0.211	0.04	0.02
Final voiced consonant (***Phon2***)	290	3,026	0.096	0.16	0.19
Low stressed vowel (***Phon3***)	124	1,367	0.091	0.07	0.09
Two syllables or more + Final voiced consonant (***Phoncomb12***)	109	489	0.223	0.06	0.03
Two syllables or more + Low stressed vowel (***Phoncomb13***)	32	176	0.182	0.02	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	173	1,125	0.154	0.10	0.07
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	90	1,031	0.087	0.05	0.06
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	62	373	0.166	0.03	0.02
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	318	1,750	0.182	0.17	0.11
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	93	1,207	0.077	0.05	0.08
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	188	2,064	0.091	0.10	0.13

Table 2(b): Other open class words in the corpus of Aran.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	233	9,372	0.025	0.17	0.35
Two syllables or more (***Phon1***)	72	436	0.165	0.05	0.02
Final voiced consonant (***Phon2***)	244	5,635	0.043	0.18	0.21
Low stressed vowel (***Phon3***)	103	2,748	0.037	0.07	0.10
Two syllables or more + Final voiced consonant (***Phoncomb12***)	104	507	0.205	0.07	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	17	175	0.097	0.01	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	98	1,058	0.093	0.07	0.04
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	39	1,847	0.021	0.03	0.07
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	23	89	0.258	0.02	0.00
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	300	4,101	0.073	0.22	0.15
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	39	178	0.219	0.03	0.01
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	122	635	0.192	0.09	0.02

Table 3(a): Nouns in the corpus of Becky.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	156	1,414	0.110	0.14	0.17
Two syllables or more (***Phon1***)	39	145	0.269	0.04	0.02
Final voiced consonant (***Phon2***)	193	1,625	0.119	0.18	0.20
Low stressed vowel (***Phon3***)	58	464	0.125	0.05	0.06
Two syllables or more + Final voiced consonant (***Phoncomb12***)	45	145	0.310	0.04	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	26	68	0.382	0.02	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	98	1,545	0.063	0.09	0.19
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	60	392	0.153	0.06	0.05
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	27	111	0.243	0.02	0.01
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	193	984	0.196	0.18	0.12
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	77	603	0.128	0.07	0.07
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	111	717	0.155	0.10	0.09

Table 3(b): Other open class words in the corpus of Becky.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	195	6,424	0.030	0.22	0.38
Two syllables or more (***Phon1***)	32	223	0.143	0.04	0.01
Final voiced consonant (***Phon2***)	150	4,316	0.035	0.17	0.26
Low stressed vowel (***Phon3***)	68	1,644	0.041	0.08	0.10
Two syllables or more + Final voiced consonant (***Phoncomb12***)	37	365	0.101	0.04	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	11	118	0.093	0.01	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	70	630	0.111	0.08	0.04
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	26	630	0.041	0.03	0.04
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	14	103	0.136	0.02	0.01
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	189	1,796	0.105	0.21	0.11
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	34	176	0.193	0.04	0.01
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	73	352	0.207	0.08	0.02

Table 4(a): Nouns in the corpus of Carl.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	146	1,939	0.075	0.13	0.16
Two syllables or more (***Phon1***)	25	90	0.278	0.02	0.01
Final voiced consonant (***Phon2***)	189	1,723	0.110	0.17	0.14
Low stressed vowel (***Phon3***)	75	798	0.094	0.07	0.06
Two syllables or more + Final voiced consonant (***Phoncomb12***)	45	171	0.263	0.04	0.01
Two syllables or more + Low stressed vowel (***Phoncomb13***)	28	74	0.378	0.03	0.01
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	105	1,000	0.105	0.10	0.08
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	62	2,475	0.025	0.06	0.20
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	27	149	0.181	0.02	0.01
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	209	1,453	0.144	0.19	0.12
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	72	1,417	0.051	0.07	0.11
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	110	1,220	0.090	0.10	0.10

Table 4(b): Other open class words in the corpus of Carl.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
No phonological cues (***Phon0***)	182	8,194	0.022	0.21	0.37
Two syllables or more (***Phon1***)	27	466	0.058	0.03	0.02
Final voiced consonant (***Phon2***)	154	4,728	0.033	0.18	0.22
Low stressed vowel (***Phon3***)	67	2,658	0.025	0.08	0.12
Two syllables or more + Final voiced consonant (***Phoncomb12***)	41	363	0.113	0.05	0.02
Two syllables or more + Low stressed vowel (***Phoncomb13***)	8	26	0.308	0.01	0.00
Two syllables or more + Trochaic stress pattern (***Phoncomb14***)	55	528	0.104	0.06	0.02
Final voiced consonant + Low stressed vowel (***Phoncomb23***)	27	825	0.033	0.03	0.04
Two syllables or more + Final voiced consonant + Low stressed vowel (***Phoncomb123***)	11	35	0.314	0.01	0.00
Two syllables or more + Final voiced consonant + Trochaic stress pattern (***Phoncomb124***)	202	3,366	0.060	0.23	0.15
Two syllables or more + Low stressed vowel + Trochaic stress pattern (***Phoncomb134***)	25	221	0.113	0.03	0.01
Two syllables or more + Final voiced consonant + Low stressed vowel + Trochaic stress pattern (***Phoncomb1234***)	77	450	0.171	0.09	0.02

Appendix E:

Descriptive data for the semantic cue analysis from each individual corpus

Table 1(a): Nouns in the corpus of Anne.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	144	1,573	0.092	0.10	0.11
Basic level count nouns (*Sem2*)	720	7,562	0.095	0.52	0.51
Basic level mass nouns (*Sem3*)	73	826	0.088	0.05	0.06
Action words (*Sem4a*)	71	453	0.157	0.05	0.03
Non-basic level words (*Sem4b*)	384	4,495	0.085	0.28	0.30

Table 1(b): Other open class words in the corpus of Anne.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	0	0	--	0.00	0.00
Basic level count nouns (*Sem2*)	0	0	--	0.00	0.00
Basic level mass nouns (*Sem3*)	0	0	--	0.00	0.00
Action words (*Sem4a*)	633	12,775	0.050	0.59	0.56
Non-basic level words (*Sem4b*)	13	202	0.064	0.01	0.01

Table 2(a): Nouns in the corpus of Aran.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	257	2,041	0.126	0.14	0.13
Basic level count nouns (*Sem2*)	809	8,209	0.099	0.44	0.51
Basic level mass nouns (*Sem3*)	93	862	0.108	0.05	0.05
Action words (*Sem4a*)	69	416	0.166	0.04	0.03
Non-basic level words (*Sem4b*)	594	4,418	0.134	0.33	0.28

Table 2(b): Other open class words in the corpus of Aran.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	0	0	--	0.00	0.00
Basic level count nouns (*Sem2*)	0	0	--	0.00	0.00
Basic level mass nouns (*Sem3*)	0	0	--	0.00	0.00
Action words (*Sem4a*)	853	14,320	0.060	0.61	0.53
Non-basic level words (*Sem4b*)	17	375	0.045	0.01	0.01

Table 3(a): Nouns in the corpus of Becky.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	145	1,615	0.090	0.13	0.20
Basic level count nouns (*Sem2*)	576	4,124	0.140	0.53	0.50
Basic level mass nouns (*Sem3*)	69	502	0.137	0.06	0.06
Action words (*Sem4a*)	42	250	0.168	0.04	0.03
Non-basic level words (*Sem4b*)	255	1,722	0.148	0.24	0.21

Table 3(b): Other open class words in the corpus of Becky.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	0	0	--	0.00	0.00
Basic level count nouns (*Sem2*)	0	0	--	0.00	0.00
Basic level mass nouns (*Sem3*)	0	0	--	0.00	0.00
Action words (*Sem4a*)	507	7,812	0.065	0.56	0.47
Non-basic level words (*Sem4b*)	15	353	0.042	0.02	0.02

Table 4(a): Nouns in the corpus of Carl.

	NOUNS				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	153	4,287	0.036	0.14	0.34
Basic level count nouns (*Sem2*)	570	5,474	0.104	0.52	0.44
Basic level mass nouns (*Sem3*)	53	465	0.114	0.05	0.04
Action words (*Sem4a*)	46	300	0.153	0.04	0.02
Non-basic level words (*Sem4b*)	274	1,983	0.138	0.25	0.16

Table 4(b): Other open class words in the corpus of Carl.

	OTHER				
	Total Types	**Total Tokens**	**Type/Token Ratio**	**Type Proportion**	**Token Proportion**
Proper nouns (*Sem1*)	0	0	--	0.00	0.00
Basic level count nouns (*Sem2*)	0	0	--	0.00	0.00
Basic level mass nouns (*Sem3*)	0	0	--	0.00	0.00
Action words (*Sem4a*)	527	11,623	0.045	0.60	0.53
Non-basic level words (*Sem4b*)	14	388	0.036	0.02	0.02