Distributional patterns in German child-directed speech and their usefulness for acquiring lexical categories

– A case study –

Distributional patterns in German child-directed speech and their usefulness for acquiring lexical categories

– A case study –

Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der Philologischen Fakultät

der Albert-Ludwigs-Universität

Freiburg i. Br.

vorgelegt von

Jens-Holger Keibel

aus Kiel

SS 2005

Keibel, J.-H. (2007). *Distributional patterns in German child-directed speech and their usefulness for acquiring lexical categories: A case study.* (Doctoral dissertation 2005, Albert-Ludwigs-Universität Freiburg, Germany). Retrieved from http://www.freidok.uni-freiburg.de/volltexte/2988/

Erstgutachter: Prof. Dr. Jürgen Dittmann

Zweitgutachter: Prof. Dr. Ad Aertsen

Drittgutachter: Prof. Dr. Bernd Kortmann

Vorsitzender des Promotionsausschusses der Gemeinsamen Kommission der Philologischen, Philosophischen und Wirtschafts- und Verhaltenswissenschaftlichen Fakultät: Prof. Dr. Hermann Schwengel

Datum der Fachprüfung im Promotionsfach:

24. März 2006

Acknowledgements

Writing a dissertation can at times make you feel like you have no social life, you are incredibly poor, and entirely lost in space. But looking back you realize there were many people, various places, and even some money that played important roles along the way.

First of all, I wish to thank Jürgen Dittmann who supervised this dissertation within the DFG research training group *Menschliche und maschinelle Intelligenz* ("Human and machine learning") at the University of Freiburg. Moreover, I would like to thank Ad Aertsen and Bernd Kortmann who took on the burden of serving as secondary reviewers. Many thanks go also to the Center for Cognitive Science (Freiburg) for generous support and for providing me with office space and technical equipment.

Much of the work presented here was pursued during several extended visits to the Center for Research in Language at the University of California, San Diego, which were in large part made possible by a DAAD fellowship. The center provided me with a stimulating scientific environment, and I thank all the people who made my stays a memorable time. Special thanks go to Jeff Elman whom I am greatly indebted to in a number of ways. Without his input und guidance, I could not have done this research. He taught me a lot about computational approaches in linguistics and about general research strategies, and he has been an inspiring role model, a rich source of advice on academic and non-academic issues, and simply great fun to work with.

The basic idea to this project, I owe to Mike Tomasello and Elena Lieven who also contributed other crucial suggestions especially during the early stages. Moreover, I thank the MPI for Evolutionary Anthropology in Leipzig for making available the *Leo* corpus without which this study would not have been possible.

How can I possibly thank my dear friend and colleague Dagmar Frohning? Her constantly available advice and our countless discussions, however related to the subject, have shaped a great deal of this dissertation, and she has been the most involved

and supportive figure throughout, which I am deeply grateful for. Moltissimo gagaschön also for all the converging evidence.

I would also like to thank several people for specific help and very useful suggestions. Heike Behrens helped me to make myself familiar with the corpus, and she drew my attention to certain facts about German and its acquisition that motivated some of my explorations. Several enriching discussions with Rik Belew were a great help for optimizing the formal evaluation tools. To Doug Roland I owe the specific idea of category scenarios which proved very useful for discussing both the method and the results. Finally, the feedback from Ewa Dąbrowska helped to refine the level of analysis for nouns and verbs.

There are many others who have made a difference. Jelena Jovanovic and Tona Rodriguez Nikl hosted me patiently for several weeks while I was apartment hunting. Heidi Markham as well as Verena Reisemann and Stefan Gößling-Reisemann took me with them on their vacations. And Daniel Panusch (almost) saved me from a frozen-pizza diet. Many thanks to you all, and to everyone who had to put up with my increasing mental absence during the final stages of writing the dissertation. Most of all, I am very thankful to my parents, Karin Keibel and Wolfgang Keibel, and to Gabi Keibel-Passmann, for their constant and loving support.

Freiburg, September 2005
Mannheim, March 2007

Contents

List of figures

List of tables

Abbreviations

1st	First person
2nd	Second person
3rd	Third person
acc.	Accusative case
ADJ	Adjective category
ADV	Adverb category
CDS	Child-directed speech
CONJ	Conjunction category
dat.	Dative case
DET	Determiner category
fem.	Feminine gender
gen.	Genetive case
imp.	Imperative verb form
inf.	Infinitive verb form
INTG	Interrogative category
INTJ	Interjection category
masc.	Masculine gender
N	Noun category
NP	Noun phrase
nom.	Nominative case
neut.	Neuter gender
past part.	Past participle form
pl.	Plural
PP	Prepositional phrase
PREP	Preposition category
PRON	Pronoun category
PTCL	Particle category
SCO vector	Standardized co-occurrence vector
sg.	Singular
V	Verb category

Introduction

How do children acquire the lexical categories (such as noun or verb) of their first language from the language sample they are exposed to? In particular, what is the role that distributional patterns of highly local lexical co-occurrence relations in the linguistic input play in this acquisition? As a research strategy, it is useful to decompose this basic question into the following subquestions.

(1) What are the salient distributional regularities in the input and how informative are they about lexical categories?
(2) Are children in principle sensitive to these regularities?
(3) To what extent do they in fact make use of these regularities in developing their lexical categories?
(4) When doing so, on which particular learning mechanism do they rely?

As the main objective of this dissertation, I address question (1) in isolation and for the particular case of German. No direct contribution will be made towards (2) – (4); but with a better understanding of the first question one could conceive a more targeted study of these other questions. All analyses are based on a very rich corpus of *child-directed speech* (CDS) utterances that were addressed to one German child over the course of three years. These data allow for a very detailed investigation — but being a case study, the results have to be interpreted with caution as it remains to be demonstrated that they are representative of the input to German children in general.

A second objective was to test and refine computational tools for extracting and evaluating distributional regularities in (electronically available) language samples. Of a number of possible tools that were initially considered, the majority was excluded for theoretical reasons, and only a few were actually implemented and tested empirically.

The dissertation is organized as follows. The first chapter provides the theoretical, empirical, and methodological background and sets the stage for several more specific goals pursued in this study. In chapter 2, I describe the corpus and the set of lexical

categories against which the distributional regularities in this corpus were evaluated. The methodological and linguistic results of the study can be found in the next two chapters. While the relevant computational tools are introduced and tested in chapter 3, chapter 4 presents detailed analyses of distributional patterns in German CDS data with respect to category development. In the final chapter, some implications of these results are discussed, focusing on the strengths and limitations of the type of distributional regularities considered here.

Chapter 1

Lexical categories and linguistic input

The first two sections of this chapter emphasize from several perspectives why the study of lexical categories and their acquisition are important fields of investigation. While the first section further discusses relevant linguistic and psycholinguistic aspects of how the adult endpoint of category acquisition might look like, controversial views on the potential role of biology and experience during this acquisition are reviewed in the second section. Among the possible ways in which the linguistic input may contribute to category acquisition, the dissertation focuses on only one type of information available, namely distributional regularities. The final section therefore summarizes previous research on distributional information in input data with respect to the acquisition of lexical categories. It concludes with presenting the particular goals pursued in this dissertation.

1.1 The status of lexical categories

Although there has been much controversy in modern linguistics about which particular lexical categories are to be distinguished, how they look like, and how they should be employed in the description of human languages, the notion of lexical categories as such and their general descriptive value are not disputed (subsection 1.1.1). Furthermore, these categories are not purely descriptive concepts; they are generally assumed to be also cognitively real, i.e., to play a role in language processing. A sample of the large body of empirical evidence supporting this assumption is listed in 1.1.2. Finally, I touch upon some fundamental issues of how lexical categories might be organized and represented in the mind (1.1.3).

1.1.1 Linguistic description

There is a long tradition of grouping the words of a language into categories that capture similarities and differences in grammatical behavior of the words (a concise summary of the history of these categories can be found in Trask, 1999). The descriptive value of these *lexical categories*[1] lies in enabling linguists to economically describe the grammatical regularities underlying the respective language in terms of general rules, rather than to enumerate all sentences that are licensed in that language. While this basic motivation for positing lexical categories is uncontroversial, there is considerable variation in how these categories are construed in linguistics; and in this subsection, I describe three sources of such variation. The first concerns the general design of the system of categories as a whole; the second pertains to the kinds of linguistic criteria by which the individual categories are to be defined; and the third asks about the role that lexical categories play in different theories of grammar.

Linguists appear to be far from agreeing on one particular system of lexical categories, even for a single language; instead, there exists much variation with respect to the number and composition of the lexical categories to be distinguished. But despite these discrepancies, current proposals for lexical category systems do not differ at arbitrary degrees. Several fundamental category labels recur across these proposals and differences between categorizations arise mainly from the fact that no given category is homogeneous on the inside — no matter how small it is, its members never have entirely identical properties. Thus, if a category system were to reflect all differences in linguistic properties, it would most likely end up with a very large number of fairly small categories (cf. Culicover, 1999; Croft, 2001).[2] The solution generally taken is to distinguish top-level categories which are further subdivided into more specific subclasses. But the problem remains to decide for any linguistically identified set of words whether it constitutes a top-level category or just a subclass within some larger category (cf. the *subclass problem*, Haspelmath, 2001), and it is these decisions in which competing category systems mainly differ. Overall, since the goal is to capture the regularities underlying the target language in an elegant and economical fashion,

---

[1] In Chomskyan generative grammar, the term *lexical category* is traditionally restricted to categories of content words, primarily noun, verb, and adjective (e.g., Chomsky, 1981:48). I use this term more generally to refer to any class of words that is determined at the hierarchical level just below the broad content–function word distinction. Various other labels have been used to denote these categories; for instance, *parts of speech*, *word classes*, and *form classes*.

[2] This is particularly relevant for closed classes which appear to be rather loosely connected sets of words insofar as virtually each of their members has distinct properties (Culicover, 1999).

there is an intrinsic tradeoff between generalization and exception. The fewer top-level categories are distinguished, the fewer and the more general grammatical rules are necessary to describe the language. On the other hand, the more general these rules become, the more exceptions to the rules have to be listed (cf. Trask, 1999).

An alternative solution that sidesteps this tradeoff is to abandon the notion of categories as discrete classes with sharp boundaries and instead consider them to be continuous in nature such that category membership is a matter of degree, with the most prototypical members in the category's center and the less prototypical ones in the periphery. This was proposed by researchers who transferred the prototype framework from semantic classes (such as words for colors, furniture, or ways of causing) to the issue of grammatical distinctions such as lexical categories (e.g., Taylor, 1989). Strong empirical support for this position — for the particular case of the English categories noun, verb, and adjective — is found in work by Ross (1972) who analyzed a number of grammatical phenomena suggesting a quasi-continuous transition from prototypical verbs to prototypical adjectives and further to prototypical nouns.[3]

These architectural differences are closely related to the second source of variation which concerns the criteria by which the categories are to be defined — i.e., the criteria by which words are assigned to categories, or, in the case of continuous categories, the dimensions along which the transitions are observed. The proposed criteria are generally formulated in terms of the words' semantic, pragmatic, syntactic, or morphosyntactic properties. Semantic criteria are usually brought forward to define the broad distinction between *content words* (which have an independent lexical meaning) and *function words* (which have no meaning when used in isolation but mark grammatical functions when used in context). Although in practice, this distinction is not always as clear-cut as it may seem, linguists operate with it quite routinely. Traditionally, semantic notions have also played a crucial role for defining the major content word categories noun, verb, and adjective. This can still be observed in many contemporary dictionaries and school books, the characterization of nouns still typically involves a reference to living beings, places, and objects; for verbs to actions and events; and for adjectives to qualities. These semantic characterizations have been highly criticized because, for any given language, they constitute neither necessary nor sufficient conditions for lexical categories (cf. Sasse, 1993:648). For instance, many English words that would typically

[3] Ross' work was published before prototype theory became popular, but his notion of *category squish* is fully compatible with a more general account of prototype phenomena (cf. Taylor, 1989:188f).

be considered verbs in fact do not refer to actions or events (e.g., *to belong*) while, on the other hand, many words denoting actions or events are not verbs (e.g., *departure*). If, however, the semantic definition of verbs is extended to also include experiences, states, etc., the verb category becomes largely indistinguishable from nouns and adjectives.

Many researchers have therefore abandoned semantics as the basis of lexical categories and concluded from these issues that categories must instead be defined in terms of formal (syntactic and morphological) criteria. Probably the most influential proposal in this direction is *distributional analysis* — a general method introduced by American structural linguists — which defines lexical categories as *substitution classes*, grouping together words with roughly identical distributions, i.e., words that can be substituted for each other in sentence frames or inflectional frames (e.g., Harris, 1946). However, this method is not free of problems either. Because very few, if any, words have precisely identical distributions, an uncompromising distributional categorization would generate a large number of small categories as described earlier (cf. Croft, 2001:36). Within this framework, the general solution has been to group all those words together that have identical distributions just across a selected set of frames (Harris, 1946:163f,177).[4] But as there is no objective way of selecting such frames, the method is likely to produce different category systems whenever it is applied. This problem is further aggravated by words that show critical distributional properties of more than one category.[5]

But even if these problems are overcome, the categories derived by distributional analysis are necessarily language-specific. Therefore, attempts to capture lexical categories — at least the major categories noun, verb, and adjective — from a cross-linguistic perspective have motivated a renaissance of semantic-pragmatic accounts.[6] One such approach proposes to overcome the before-mentioned problems of semantic

---

[4] One particularly sophisticated approach of this type was presented by Fries (1952/1957) who identified for American English four large categories — corresponding to nouns (together with pronouns), verbs, adjectives, and adverbs — and 15 smaller categories of *function words*. Fries treated content word and function word categories in isolation, discussing them in separate chapters, and labeling them by two different notations. However, he explicitly refused to make this distinction, presumably because it would involve semantic properties (ibid.:88).

[5] Such ambiguous words can most likely be found for any nontrivial category system. This becomes apparent, for instance, in the collection of lexico-syntactic idiosyncrasies presented by Culicover (1999).

[6] Such attempts were motivated by (implicit and explicit) claims about universal categories and, likewise, by typological reports about certain languages lacking particular categories that are distinguished in English. Without a clear cross-linguistic characterization of categories, a typological fieldworker would not know how to even look for potential nouns, verbs, etc.

definitions by resorting to prototype theory, and to characterize only prototypical nouns, verbs, and adjectives as referring to physical objects, intentional physical actions and properties of objects, respectively (e.g., Bates & MacWhinney, 1982). In an alternative account with a stronger emphasis on pragmatics, Hopper and Thompson (1984) point out that the degree to which a given word is, for instance, a prototypical noun, does not so much depend on its inherent semantic properties as a lexical entry but rather on the semantics of how it is used in the discourse context. They describe the prototypical pragmatic function of nouns as introducing a manipulable participant into the discourse, and that of verbs as reporting an actual event. Abstracting from this account, Croft (1991, chap. 2) characterizes the prototypical usage of nouns, verbs, and adjectives in terms of the notion of *typological markedness*, linking semantic classes of lexical concepts with pragmatic classes of speech acts. For instance, nouns denoting objects are considered typologically unmarked when used in propositional acts encoding reference. Correspondingly, verbs denoting actions are unmarked when used to express predication, and adjectives denoting properties are unmarked when used in acts of modification.

Focusing on the cognitive aspects underlying lexical categories, Langacker (1987) develops an elaborate notion of semantics (couched within his Cognitive Grammar framework) that treats the meaning of a particular linguistic unit not as some objective property of the unit per se, but as its conceptualization and the abstract cognitive operations involved in it. Abstracting away from prototypical nouns and verbs, Langacker derives schematic conceptualizations intended to universally characterize all nouns (as regions in cognitive domains that are interpreted as static entities) and all verbs (as relations that are cognitively scanned through time).[7]

It is important to note that these semantic-pragmatic and semantic-cognitive approaches do not actually provide criteria by which words in a particular language can be assigned to categories. Rather, they explain from a cross-linguistic perspective why languages do have such a thing as lexical categories in the first place, and why certain categorial distinctions like those between noun, verb, and adjective may exist. But languages differ with respect to the degree and grammatical means by which they discriminate these categories; and in the extreme, a few languages do not distinguish

[7] Note, however, that according to Langacker's characterization, nonfinite verb forms (i.e., infinitives and participles) are not considered verbs but rather a special class of words denoting *atemporal relations* (Langacker, 1987:75f; also see discussion on pp. 217f).

them at all.[8] In consequence, the semantic characterizations can serve as a heuristic for finding these categories — or for testing for their existence — in a particular language; but they then have to be defined and analyzed in terms of formal criteria which are necessarily language-specific (cf. Sasse, 1993:651-657).

As the third source of variation, lexical categories play somewhat different roles in competing theories of grammar. To illustrate this, I briefly compare their status in two such theories in which it differs fundamentally. First, within all variants of Chomskyan generative grammar, lexical categories have the status of atomic units which, by virtue of phrase structure rules, are successively combined into more complex linguistic units, up to complete sentences. Each of these categories can be characterized in terms of the syntactic positions it can fill according to the grammatical theory; but generativists typically do not define categories by explicit and precise criteria independent of the theory. Instead they simply assume them as universal abstract entities of their own right.

This theoretical primacy of lexical categories contrasts sharply with their status in a more recently introduced variant of the family of construction grammars, namely *radical construction grammar* (RCG; Croft, 2001, 2005). Here, the basic unit is that of a construction, a notion which is meant to capture linguistic expressions at any degree of complexity and abstraction, ranging from concrete individual words to complex and highly abstract schemas that roughly correspond to the phrase structure rules in the generative paradigm. The schematic portions of a construction consist of slots that can be filled by particular sets of words (or other constructions); and it is these sets of slot-fillers that constitute lexical categories (and more complex syntactic categories) which therefore exist only relative to the construction. Thus, within this theory, lexical categories are derived (instead of assumed), they are construction-specific (instead of existing independently of grammatical rules and structures) and therefore necessarily language-specific (rather than universal).[9]

In sum, the theoretical status, the defining criteria, and the precise composition of lexical categories are debated issues, for particular languages as well as cross-

[8] The currently predominant view on universal categories seems to be that all languages distinguish the two broad categories of content words and function words and that most languages distinguish the two categories noun and verb (cf. Sasse, 1993; Haspelmath, 2001).

[9] For instance, the schematic intransitive construction for English is coded in terms of two category-specific categories [INTRSUBJ INTRVERB] whereas the transitive active construction is described as [TRSUBJ TRVERB TROBJ]. In particular, there is no global verb category in RCG. A category comprising all verbs can be built by generalizing across construction-specific categories INTRVERB, TRVERB, etc.; but this step involves morphological constructions such that the resulting category is again construction-specific (Croft, 2005). Thus, the resulting category may have the same extension as the traditional verb category but a very different theoretical status.

linguistically. But despite this variation, the general value of lexical categories as terms of linguistic description is not questioned; and at least for Indo-European languages, traditional categories like noun, verb, preposition etc. — but also broader category distinctions like that between content words and function words — all are associated with an intuitively clear core notion such that native speakers of a particular language would typically agree on some prototypical members of each of these categories. It is therefore warranted to use these category labels when only such core notions are intended. For the purpose of this dissertation, confusion is avoided by considering only a limited subset of words and explicitly specifying their category (cf. section 2.2).

1.1.2 Cognitive correlates

Despite these controversies in linguistic theory, it is generally agreed that lexical categories as such are cognitively real and not merely theoretical terms of linguistic description. The mere fact that there are *any* abstract categories quite obviously has to be cognitively real — otherwise, native speakers could not create and understand new utterances that they had never heard before. But theoretically, these cognitive categories need not correspond to the ones that were identified by any linguistic theory. However, a rich body of empirical evidence from psycholinguistics, aphasiology, and cognitive neuroscience converges on the view that at least some fundamental category distinctions of linguistics play a role in human language processing — most clearly for the contrast between the categories noun and verb but also for the distinction between content and function words. Here, I summarize some of the core phenomena.

In everyday situations, but also experimentally-induced, speakers can experience so-called *tip-of-the-tongue* (TOT) states in which they fail to retrieve a particular word although they have the impression it is just about to come out. In such states, speakers can typically access partial information about the word including its lexical category and often even more fine-grained grammatical properties such as noun gender (e.g., Brown & McNeill, 1966; Vigliocco, Antonini, & Garrett, 1997). Moreover, TOT states almost exclusively occur for content words, and here by far most frequently for proper names and common nouns (for review see Vigliocco, 2001).

More remarkable category effects can be observed when the words are retrieved quite fluently but errors occur during this process. For instance, a retrieved word may not be the one that was intended (so-called *word substitutions*), e.g., “*on my elbow*” instead of “*on my knee*” (examples taken from Dell & Reich, 1981); or two words may

be retrieved simultaneously and compete to fill the same single slot in an utterance which often results in a fusion of both words (*word blends*), e.g., "*hail a tab*" instead of "*hail a taxi/cab*". But even if the right number of words and only the intended words are retrieved, some of them may switch their syntactic slots (*word exchanges*), e.g., "*writing a mother to my letter*" instead of "*writing a letter to my mother*". It has been noted repeatedly that the vast majority of such word-level speech errors involve words of the same lexical category (e.g., Meringer & Mayer, 1895; Bierwisch, 1970/1981; Fromkin, 1971; Garrett, 1975; Dell & Reich, 1981). Other types of speech errors can occur at the level of sounds or morphemes which result in words that do not necessarily exist. However, when they do, these also tend to belong to the same category as the word that was actually intended (Garrett, 1975).

A related phenomenon concerns free word associations. When presented with a stimulus word, adults most frequently respond with another word of the same lexical category (Ervin, 1961, also citing the earliest work on word associations; Deese, 1965; the chapters in Postman & Keppel, 1970). Interestingly, this is not the case for preschool children who commonly respond with content words that are syntagmatically, rather than paradigmatically, related with the stimulus word (Ervin, 1961; Nelson, 1977). This indicates that at least some category effects emerge as a result of linguistic experience.

Moving from single words to larger structures, it has been proposed that at any point during sentence processing, adult readers or listeners generate probabilistic expectations about subsequent words, based on the words and structure already encountered (e.g., Elman, 1991; MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994; Altmann, Nice, Garnham, & Henstra, 1998; Narayanan & Jurafsky, 2002; Gibson, 2006). There is an increasing body of evidence supporting this general hypothesis. In particular, a number of cross-modal priming experiments demonstrated that adult listeners and readers have expectations about the lexical category of the next word. Sentence fragments predicting the lexical category of the next word were found to have inhibitory priming effects in tasks of lexical decision (e.g., Wright & Garrett, 1984), word naming (e.g., Liu, 1997), and picture naming (e.g., Federmeier & Bates, 1997) when the target word was not of the predicted category. The latter two studies also found facilitative priming effects for target words of the category predicted by the sentence fragment. Both these studies used sentence fragments that strongly predict the next word's lexical category (either verb or noun) but only mildly constrain its semantic properties (e.g., "*I want to …*", "*Look at the …*"). This indicates

that the priming effects can be attributed to grammatical rather than semantic expectations.

Evidence from aphasiology includes a double dissociation between noun and verb processing that was documented cross-linguistically for certain groups of patients with language disorders who were significantly more impaired at producing or comprehending nouns than verbs, or vice versa (e.g., Damasio & Tranel, 1993; Daniele, Giustolisi, Silveri, Colosimo, & Gainotti, 1994; Chen & Bates, 1998). Furthermore, Broca's aphasics typically display a severe deficit in processing function words and bound morphemes relative to content words (cf. Friederici & Saddy, 1993).

Neuroimaging studies of intact brains, using techniques such as ERP (event-related potential), PET (positron emission tomography), and fMRI (functional magnetic resonance imaging), have shown different patterns of brain activity for the processing of nouns and verbs (e.g., Pulvermüller, Lutzenberger, & Preissl, 1999; Perani et al., 1999). Additional studies (e.g., Federmeier, Segal, Lombrozo, & Kutas, 2000; Shapiro, Pascual-Leone, Mottaghy, Gangitano, & Caramazza, 2001; Vinson & Vigliocco, 2002) appear to rule out the possibility that this neural dissociation could arise from semantic factors. Further differences in neural processing have been identified for the distinction between content and function words (for review see Pulvermüller, 1999).

In sum, empirical evidence from psycholinguistics, aphasiology, and cognitive neuroscience converges on the minimal conclusion that the storage or processing of words somehow involves their lexical category, among other properties. At least some lexical category distinctions are cognitively real.[10]

1.1.3 Issues of representation

Given that lexical categories are cognitively real, the next important question is how they are represented and organized in the mind. There are a number of proposals that can be inferred from more general accounts of the mental lexicon and language

[10] The data from aphasiology (relation between lesioned brain region and lexical category showing a processing deficit) and from neuroimaging studies with unimpaired speakers (relation between active brain region and lexical category being processed) have sometimes been interpreted to imply that particularly nouns and verbs are stored or processed in distinct brain regions. However, the relation between brain regions and lexical categories might be an indirect one; so far the data only show that different brain regions contribute differently to the processing of nouns and verbs (cf. Elman et al., 1996). Furthermore, the ERP study by Federmeier et al. (2000) demonstrated that neural effects of lexical categories differ in their nature and neural location, crucially depending on the context that words occur in and on the kind of task to be performed. Nevertheless, the results do show that there are *some* neural differences between these categories.

production and perception. In the following, I highlight three crucial contrasts in which these proposals differ. Although these contrasts are related to each other, each of them has distinct consequences on how one thinks about lexical categories.

Explicit vs. implicit categories

Adults may explicitly operate with lexical categories or just use language *as if* they were operating with them. The cognitive effects of lexical categories listed in the preceding subsection are consistent with either possibility — that is, they may be caused by processes that directly involve lexical categories or they may arise as epiphenomena of other processes. In the former perspective, lexical categories are typically construed as abstract symbols that can be stored in the mental lexicon and manipulated by cognitive processes. An instantiation of the latter view is found in connectionist models where categories can be inferred from complex processes and representations, for instance, as different regions in a high-dimensional space of mental states (Elman, 1990, 2004). While this view seems to weaken the role that lexical categories play in cognitive processes, it suggests the interesting possibility that on different occasions of language processing, the relevant categories might take on different shapes — corresponding to different angles from which the cognitive apparatus *looks* at that space.

Type vs. token

This points to a second crucial contrast which is closely related. Lexical categories may be stored and processed as properties of word *types*, or alternatively, they may get assigned to particular instantiations (i.e., *tokens*) of these words by virtue of the context they occur in. While an intuitive and prominent psycholinguistic view of the mental lexicon (e.g., Levelt, Roelofs, & Meyer, 1999) regards lexical categories as tags of word types (more specifically, *lemmas*), the alternative position is in line with several of the linguistic characterizations that were cited in 1.1.1.

For instance, Hopper and Thompson (1984:747) conclude from their pragmatic-semantic characterization of prototypical nouns and verbs that although word types can show a "propensity or predisposition to become [nouns] or [verbs]", they are "in principle to be considered as LACKING CATEGORIALITY completely unless nounhood and verbhood is forced on them by their discourse functions" (capitalization by the authors). Within the RCG model by Croft (2001), lexical categories are by definition properties of word tokens. Since each category is defined by a particular construction, category membership of words can only be decided when they occur

within constructions. Psycholinguistic support for the alternative position comes from sentence processing studies which show that *categorially ambiguous* words (i.e., word types that can be used as members of more than one category) are in general processed quite smoothly when they occur in disambiguating context (for review see Altmann, 1998).[11]

Interestingly, the ERP study by Federmeier et al. (2000) provides empirical support for both views. The authors found category effects of word types that cannot be explained by context, and conversely category effects of context that cannot be explained by word types. This suggests a blend of both positions — some categorial properties are (explicitly or implicitly) stored with word types while some other categorial aspects only arise in context; and both types of information interact when a particular word token is processed along with its context. Again, connectionist models can be taken as illustrating one possibility of how such a blend might be realized in the mind (cf. Elman, 1990, 2004).

Discrete vs. continuous categories

Section 1.1.1 briefly touched on the position that category membership is a matter of degree, with more prototypical members in a category's center and a continuous transition into neighboring categories. In contrast, according to the traditional view, categories have precise boundaries and are defined in terms of necessary and sufficient criteria; they are therefore considered homogeneous in the sense that each member is an equally good representative of its category. The prototype effects that have been identified at various levels in adult language processing and child language acquisition (e.g., Taylor, 1989) appear to indicate that the mental representations of categories are continuous in nature. However, it is a conceivable possibility that the mental representations in fact involve discrete categories — which would presumably be large in number and small in size — and that prototype effects arise as a side-effect of how these representations are put to use in language processing, or how they are created in the process of language acquisition.

But the distinction between the two notions is in fact not as clear-cut as may seem to be implied. For prototype categories in general, prototypical members share

[11] In fact, this is typically studied within the broader issue of *lexical ambiguity*. This term refers to words that can express several different meanings — which may or may not involve multiple categories. Since multiple meanings within the same lexical category are not relevant within the scope of this dissertation, I use the more specific term *categorial ambiguity*.

properties with each other which can be taken as rough equivalents to the necessary and sufficient criteria required for defining discrete categories. In other words, the centers of different prototype categories are discrete because prototypical members of one category are not simultaneously prototypical of other categories (cf. Taylor, 1989:54). In this sense, discrete categories with well-defined boundaries may be interpreted as approximations of continuous categories with prototypical centers, and vice versa.

These three contrasts, especially the former two, are closely related to each other, but they are nonetheless logically independent. They may correlate in the positions advanced by psycholinguists — a notion of discrete categories being explicitly encoded with the word types in the mental lexicon probably represents the prevalent view whereas researchers defending a notion of implicit categories typically assume these categories to arise only in the usage context of particular word tokens. But it is theoretically possible to construe categories as implicit properties of word types (e.g., when word types are represented as unclassified nodes in a lexical network or as points in a vector space such that category membership has to be inferred from their location within the network or vector space). Conversely, it is possible to think of categories as being explicitly represented but only arising for particular word tokens in context (e.g., the construction-specific categories in RCG). And in both cases, the categories may be regarded as discrete or continuous (cf. above).

1.2 Lexical categories in language acquisition

In the first section I summarized evidence suggesting that lexical categories play an important (though possibly indirect) role in the cognitive processes underlying the language usage of adults. It is therefore reasonable to ask how these adult categories emerge in the process of language acquisition. And in this section, I review some crucial issues and claims pertaining to this question. It is important to emphasize here that the study is concerned with the acquisition of lexical categories and not with vocabulary acquisition. Theoretically, a child may learn hundreds of nouns without ever realizing that they all belong to the same category or that there is such a thing as categories in the first place. It is quite plausible, however, that these two aspects of acquisition closely interact with each other. A child may first have to acquire a certain minimum number of

words of a given category before being able to discover that category (cf. the *critical mass* effects of lexical development on acquisition of grammar, proposed by Bates and Goodman, 1997). On the other hand, having discovered a set of initial proto-categories may in turn facilitate the acquisition of new lexical items; and there is evidence that this is indeed the case (Elman, 2004; Borovsky & Elman, 2006).

The section begins by reviewing three types of arguments for the claim that at least some lexical category distinctions are innate (subsection 1.2.1). Irrespective of such claims, a child's linguistic experience necessarily plays a crucial role in category development; and I summarize various potential sources of information in the linguistic input that have been proposed to be exploited by the child during this process (1.2.2).

1.2.1 Claims about innate categories

Much debate in the study of language acquisition revolved around the issue whether infants approach language with an innate predisposition for specific aspects of human language, including the fundamental distinctions between lexical categories. On the one hand, there is consensus among contemporary researchers that genetic endowment plays a crucial role in language development. Undeniably, all human beings are in principle capable of acquiring a language while, at least so far, not a single individual of any other species has been found to acquire a communicative system even remotely as complex as human languages. But this distinctively human predisposition may be rather general in nature and need not involve the genetic encoding of any specific grammatical content, such as lexical categories. These two opposing views on what is innate can be captured by the distinction between *general nativism* and *linguistic nativism* (cf. Scholz & Pullum, 2002).[12]

Most prominently associated with linguistic nativism are generative approaches to language acquisition which commonly assume lexical categories to be innately prespecified, in the form of abstract symbols — which in some accounts are supplemented by innate strategies for how instances of these categories can be identified in any language. At least three different kinds of arguments have been brought forward to justify this assumption: typological evidence, considerations about the best research strategy, and finally, logical arguments questioning the learnability of lexical categories

[12] Other general arguments for linguistic nativism are summarized by Jackendoff (2002:94-101); and criticisms of these arguments can be found in Tomasello (1995) and Elman et al. (1996:371-391).

from the input. Below, I briefly review each of these arguments in turn, along with objections against their logic or premises.

First, the typological evidence concerns universal properties of human languages. From early on, generative grammar accounts of language acquisition have closely linked the logically independent notions of linguistic universals and innateness. That is, properties that are universally found across all languages, are by default taken to arise from innate prespecification (e.g., Chomsky, 1965:27) and should therefore be part of a theory of *universal grammar* (UG) which describes the initial state of a child before being exposed to language (cf. Jackendoff, 2002). At the time this theory was first described, there were several lexical categories of English considered good candidates to be linguistic universals; and thus, they entered UG. However, when typologists converged on the view that in fact only the content–function word distinction, and possibly the categories noun and verb are universal (cf. 1.1.1), UG still retained the assumption of innate lexical categories; but the child is now thought to be equipped with some device to determine which of these categories are instantiated in the particular language she is exposed to (cf. Culicover, 1999).

Innate linguistic constraints, however, are only one possible explanation for linguistic universals. And the generative tradition has been criticized for too readily equating (apparent and genuine) universals with innate predisposition. Alternative accounts propose that universals might arise from general cognitive and social constraints as well as language evolution (Tomasello, 1995), or as an epiphenomenon of the general information-theoretic problem of mapping nonlinear messages onto linear speech that any communicative system has to solve (Elman et al., 1996). Some proponents of UG concede that alternative explanations of linguistic universals should be preferred in principle but remain skeptical whether the proposed alternatives can fully account for the phenomena without invoking innate linguistic constraints (e.g., Jackendoff, 2002:74-82).

The second argument in favor of innate categories involves the so-called *continuity assumption*, an influential formulation of which was advocated by Pinker (1984). By this assumption, children — at any point of their linguistic development — operate with the same linguistic mechanisms as adults (including the same syntactic categories and the same types of grammatical rules). In consequence, these mechanisms are assumed to be innate. Pinker argues that theories of language acquisition should be constrained by specific assumptions about innate prerequisites of the child in order to derive testable predictions with respect to the course of language development (ibid.:6-10). To this end,

the continuity assumption serves as the null hypothesis of his research strategy; and it has to be rejected only if the acquisition theory he proposes generates predictions that are inconsistent with developmental facts.

One immediately obvious fact that comes to mind, of course, is that children initially do not show the linguistic proficiency of adults in their observable language behavior; and several proposals have been made to explain why this might be the case despite the assumed adult-like competence of infants. Tomasello (2000) critically evaluated these proposals against a variety of empirical evidence and found them to be inconsistent with the data — except for lexicalist accounts which posit that children first acquire substantial syntactic facts about individual lexical items before they can access the linguistic generalizations encoded in UG. However, as Culicover (1999:197) points out, if children are ascribed the capacities to extract some syntactic properties of words, there is no a-priori reason to deny them the capacities to generalize across these properties and derive the categories themselves. Culicover (ibid.:36-41) demonstrates further that, given the variety of syntactic idiosyncrasies found in natural languages, children have to solve very similar learning tasks, with or without innate categories. Thus, while Tomasello's and Culicover's analyses cannot ultimately falsify the continuity assumption, they converge on the conclusion that innate categories would provide no advantage to the child.

Nevertheless, the research strategy to take the explicit assumptions about the child's innate predisposition as a null hypothesis seems indeed justified by the logical considerations put forward by Pinker. On the other hand, it may invite to resort to innate constraints as the *default explanation* for any developmental mystery, thereby conveniently passing all burden of proof on to genetics and providing little, if any, insight into the matter (cf. Bohannon, MacWhinney, & Snow, 1990; Berman, 1991). In Pinker's terms, adding new assumptions about innate constraints amounts to testing another theory. While explicit assumptions *do* constrain every single theory and make it more testable, a general practice of resorting to innateness by default would appear rather unconstrained. An alternative research strategy would therefore rather test null hypotheses about learning — about the underlying cognitive mechanisms and the available evidence in the environment, the latter of which can be investigated more directly.

However, it is precisely this available evidence that the third and probably most influential type of argument for innate categories (and for UG in general) is targeted at. In its essence, it asserts that children's linguistic input is inadequate for reliably

acquiring the complete and accurate grammar from it. But children do arrive at such grammars; therefore, so the argument, acquisition must be constrained by an innate UG. This general *inadequacy* argument comes in various different flavors (a historical survey is provided by Thomas, 2002), and the most prominent versions have been extensively debated in the literature (e.g., Pullum & Scholz, 2002, and response articles in the same volume). The argument is typically not spelled out for lexical categories — but they are implicit in many of its proposed variants because if adult grammar is construed as a generative grammar, one can hardly posit any innate constraints on this grammar (beyond very general principles such as structure dependency) that do not involve its primitives; hence, these must be innate in the first place. As Jackendoff (2002:77) puts it within an even broader context, "one cannot construct a language without them. Chomsky therefore wishes to attribute them to the brain's prespecification".

One particular version of the argument that is more immediately relevant for the acquisition of lexical categories concerns *negative evidence*, i.e., information regarding conceivable structures that are ungrammatical. Over the course of language development, children indisputably generalize beyond their input in that they learn to produce and comprehend utterances that they have never heard before. In this process, children might easily overgeneralize and miss the various exceptions that grammatical regularities in any natural language tend to have — but children appear to avoid or recover from possible overgeneralizations (cf. Bowerman, 1988; Sokolov & Snow, 1994). The argument concludes that this can only be explained by innate constraints because the input to children is asserted to lack the right kind of negative evidence to help them detect the exceptions to apparent rules (e.g., Crain, 1991; Tracy, 1990). This line of reasoning has received further support from formal learnability theory which developed around Gold's (1967) famous theorems on *language identification in the limit* (e.g., Pinker, 1979; Wexler & Culicover, 1980).

There have been several objections against this broad argument. For instance, although children might generally receive very little overt negative evidence (in the form of explicit marking or even corrections of ungrammatical utterances), they appear to receive a variety of *implicit* or *indirect* negative evidence, be it in the form of a systematic absence of overgeneralizations from the input (e.g., Bowerman, 1988:98, note 5; Elman, 1991; Schlesinger, 1991), or in the form of statistical patterns in adults' responses to errors in the child's own productions (cf. Sokolov & Snow, 1994). Particularly the results from learnability theory have been criticized for building on

formal assumptions that in crucial ways do not apply to the learning situation of the child (e.g., Braine, 1988:218; Elman et al., 1996:385). Further, if the assumptions about the child and her input are made only slightly more realistic, positive evidence can in principle suffice for recovering from overgeneralizations — as more recent results from learnability theory (Chater & Vitányi, 2004) and connectionist modeling (Elman, 2002) have demonstrated. But even if Gold's mathematical paradigm is accepted and innate constraints are taken to be the only possible conclusion, they need not be specific to language but could also be general constraints on cognitive development, as Scholz and Pullum (2002:196) point out. And finally, it is questionable whether the issue of overgeneralization can be exploited as an argument for innate constraints at all because it is problematic for nativist accounts as well (Bowerman, 1988; Schlesinger, 1991).

In sum, none of the existing arguments for innate categories stands up to scrutiny. On the other hand, opponents of the innateness assumption may have refuted these arguments but not the assumption itself — so far, no conclusive evidence has been presented for the counterclaim that lexical category distinctions are *not* innately predetermined. However, ultimately, this question might bear little relevance on the puzzle of language acquisition because, as was pointed out earlier, innate categories would probably not even make the child's acquisition task significantly easier (cf. Culicover, 1999; Tomasello, 2000).

Nevertheless, assuming innate categories has consequences on how one views the child's starting point and her operative learning strategies. To illustrate this, it is useful to subdivide the acquisition of lexical categories into the two tasks (i) to identify how many and which categories there are to be distinguished in the target language (*discovery of categories*); and (ii) to learn which words belong to which category (*mapping words onto categories*). According to the innateness assumption, the first task would be constrained by genetic endowment. With current UG accounts, however, the child has to determine which of the innate categories provided by UG are actually instantiated in her particular target language. Logically, this challenge belongs to task (i) but it can also be construed as being solved by (ii). By contrast, empiricist accounts typically assume that, at least in the early stages, both tasks are mastered in combination such that one task should benefit from progress made in the other one.

Because different languages do not have identical — nor even considerably overlapping — vocabularies, no-one would dispute that at least the mapping task (ii) is essentially accomplished by learning from linguistic experience. But there has been disagreement concerning the precise kinds of information found in the input that are

deemed relevant for category acquisition; and concerning the time course in which each of these different kinds of information become available or useful. And these differences are closely associated with the discovery task (i) and the issue of innateness. In this sense, the conflicts were in part carried over to the other side — that is, from arguments about learnability to arguments about learning.

1.2.2 Sources of information in the input

The general idea of how the input can be informative about lexical categories is that words of the same lexical category (or at last some subset of these words) tend to share certain characteristic properties which the child can reasonably be assumed to have access to. Following these observable properties might therefore take the child some way into the adult categories. However, such cues are often no deterministic predictors of lexical category; and when they are, they generally only capture small homogeneous subclasses of the intended categories. The child thus has to deal with imperfect correlations between cues and categories — a situation which parallels the severe difficulties which linguists encountered in attempting to identify necessary and sufficient criteria for defining these categories (cf. 1.1.1).

This subsection begins by specifying various sources of information in the input that have been proposed as providing the child with useful cues about lexical categories. I then briefly sketch views concerning the role and interplay between these different sources of information during the process of category acquisition.

Semantic content

A very common proposal is that children can exploit semantic notions such as *concrete physical object*, *intentional physical action*, *stative attribute*, and *spatial relation* that characterize prototypical nouns, verbs, adjectives, and prepositions, respectively, to develop initial proto-categories (Grimshaw, 1981; Bates & MacWhinney, 1982; Pinker, 1984). Beyond these canonical associations between lexical and semantic categories, Braine (1992) points to a wider range of ontological notions such as *place*, *time*, *event*, *experience*, *proposition*, etc., each of which appears to be correlated with a particular lexical category and may thus lead the child to discovering some smaller semantic clusters of words which are good candidates for belonging to the same categories. However, as was remarked in 1.1.1, such correlations are not perfect, and it is generally

agreed that some additional source of information has to be used by children to eventually arrive at the adult categories.

A closely related possibility is that children capitalize on pragmatic-semantic cues to lexical categories. Inspired by Hopper and Thompson's (1984) characterization of the categories noun and verb (cf. 1.1.1), Tomasello (2003a:169ff; also see Tomasello, 1992) proposed that children identify the communicative functions (such as *reference* and *predication*) with which particular words are used within utterances, and gradually detect similarities between them with respect to the distribution of communicative roles they can play across usage events. Tomasello summarizes developmental evidence for such a learning mechanism which he terms *functionally based distributional analysis*.

Perceptual cues

With respect to the distinction between function words and content words, Shi, Morgan, and Allopenna (1998; Morgan, Shi, & Allopenna, 1996) list various phonological and acoustic cues that can potentially help the child to discover these superordinate categories. For instance, English function words, relative to content words, tend to contain fewer and simpler syllables in which vowels are typically pronounced more quickly. The authors demonstrate for a typological variety of languages (English, Turkish, and Mandarin Chinese) that while, in isolation, no single cue would be particularly useful, they together suffice for building categories that closely approximate the content–function word distinction.

Kelly (1992, 1996) reviews a collection of phonological cues that discern the categories noun and verb in English. For instance, nouns on average contain more syllables than verbs (even when inflectional suffixes are included); when they have the same number of syllables, the pronunciation of nouns generally takes longer. In particular, when comprising exactly two syllables, nouns strongly prefer to have trochaic stress (i.e., on the first syllable) whereas verbs tend to take iambic stress (i.e., on the second syllable); and in the case of noun–verb homographs, these tendencies are never reversed. Durieux and Gillis (2001) demonstrated that, as in the case of the content–function word distinction, these cues individually are only weak predictors of lexical category but together can serve to discriminate nouns from verbs quite reliably. They further showed that the results extend very well to a typologically related language such as Dutch, and to the distinctions between all major open class categories noun, verb, adjective, and adverb.

Prosodic cues may assist category acquisition in a more indirect way. Pauses, final lengthening, and change in pitch mark the boundaries between prosodic units which often correspond to syntactic units such as clauses and phrases (e.g., Fisher & Tokura, 1996; Jusczyk, 1998). Even though these correspondences are not systematic, they may help the child to break down input utterances into syntactically meaningful smaller chunks which they can then further analyze using other kinds of cues, thus employing a "divide and conquer" strategy (Jusczyk). In this way, prosodic cues can constrain the individual words' properties that are recorded by the child and thereby support category acquisition. This becomes immediately evident in cases where prosodic units consist of single words such as subject pronouns.

Formal-distributional regularities

In most general terms, the formal-distributional properties of a word pertain to the spectrum of combinations which it can occur in. Thus, while most of the former kinds of information concern properties of a word that become more or less manifest in each of its instances, its formal-distributional properties are probabilistic in nature, summarizing a set of possibilities that only show across a sufficiently representative collection of instances. Four general types of formal-distributional cues to lexical category can be identified which arise from two fundamental distinctions. One concerns the linguistic level (lexical vs. morphological) on which the distributional regularities hold whereas the other distinction concerns their accessibility — some may be extracted directly from overt speech while others involve some structural analysis. The resulting types of distributional information are presented in Table 1-1 below.

Overt distributional cues at the lexical level involve a word's tendency to occur in particular positions of an utterance (serial position) or in particular positions relative to other words (lexical co-occurrence). At the morphological level, overt distributional cues mainly involve the different suffixes a given word stem is observed to take. But in principle, also more complex morphological operations can be considered that alter the stem itself. One particularly influential proposal of overt distributional cues (lexical and morphological) was put forward by Maratsos and Chalkley (1980). It mainly focuses on a word's co-occurrences with function words (such as determiners and auxiliaries) and bound function morphemes (such as inflectional suffixes). In order to discern ambiguous words and morphemes, their semantic function is taken into account. For instance, in the case of English, the morpheme *-s* can be attached to a noun to signal possession (e.g., *the book's cover is nice*), plural (e.g., *the books are nice*), or as a

contracted form of *be* (e.g., *the book's nice*) while on verbs, it flags present tense (e.g., *she books a flight*). Co-occurrences with such function words, together with a specification of their semantic effect, are referred to as *semantic-distributional patterns*.[13] Thus, this particular type of information may be formal-distributional in its essence, but it incorporates semantics as secondary information. However, the notion of overt distributional regularities does not require reference to semantics (cf. 1.3.1).

Table 1-1: *Four types of formal-distributional information*

Accessibility	**Linguistic level of combination**	
	Lexical	Morphological
Overt	Lexical co-occurrence and serial position	Morphological variation
Structure-dependent	Syntactic slots in phrase structure	Inflectional and derivational schemas

Whereas overt distributional cues can be inferred directly from the input (possibly supported by nonlinguistic context), structure-dependent distributional cues presuppose some rudimentary knowledge of syntactic and morphosyntactic structure in terms of at least a basic scaffolding of lexical categories. At the level of lexical combinations, the child receives cues about the lexical category of a previously unclassified item when she encounters this item in familiar syntactic structures which she can identify by other words that she already classified (be it only provisionally). At the morphological level, likewise, even partial knowledge of inflectional paradigms should help the child to detect inflecting words and assign them to their appropriate category. The processes by which children might exploit structure-dependent distributional information have been subsumed under the terms *context-sensitive distributional learning* (Pinker, 1984) and the *old-rules-analyze-new-material-principle* (Braine, 1992).

But among these various sources of information, which are the relevant ones? For quite some time, this seemingly fundamental question has received a good deal of attention; and opposing theories of category acquisition made quite different

[13] In a later paper, Maratsos (1990) relabels these patterns as *small-scale combinations*.

hypotheses. To illustrate this, I briefly sketch two elaborate accounts that have been quite influential. In the first one, Maratsos and Chalkley (1980) proposed that the child approaches the task by detecting overt semantic-distributional patterns in which a particular word can appear. She then gradually discovers the lexical categories as correlations among these patterns. For instance, English words that can take the suffix *-ed* to signal occurrence in the past also tend to co-occur with *don't* to mark non-occurrence etc. In contrast, words that tend to co-occur with *were* to signal occurrence in the past tend to co-occur with *aren't* to mark non-occurrence etc. The discovery of the former set of correlated patterns guides the child to develop the verb category while the second set gives rise to the adjective category.[14]

Pinker (1984) formulated an alternative account, the *semantic bootstrapping hypothesis*, which had previously been outlined by Grimshaw (1981) and Macnamara (1982). Appealing extensively to the innateness assumption (cf. 1.2.1), it claims that the child initially identifies tentative exemplars of the innate categories by the canonical semantic associations (e.g., *physical object* for noun) which are likewise posited to be innate. Together with innate grammatical knowledge linked to the categories, these exemplars help the child to gradually acquire some language-specific grammatical rules of the language. Equipped with such rules, the child can in a second stage exploit structure-dependent distributional cues (both at the syntactic and morphosyntactic level) to classify other words and to revise the initial classifications that were based on semantics. These growing categories in turn give rise to more and better grammatical rules. Semantic cues thus play a role only at the onset of category acquisition; they merely guide the child to link the innate knowledge to some words and structures in the target language before putatively more powerful learning mechanisms can kick in. As Macnamara (1982:134) puts it, "the child climbs to grammar on a semantic ladder and then kicks the ladder away".[15]

---

[14] Of course, such correlations are not perfect. For instance, irregular verbs do not take *-ed* past tense but in general occur in the same correlated patterns as regular verbs. The authors discuss at length how the child might handle such imperfections and recover from initial overgeneralizations.

[15] Braine (1992), elaborating work by Schlesinger (1988) and himself (1987, 1988), reformulated Pinker's (1984) account in a way that can dispense with any assumptions about the innateness of lexical categories and other specifically linguistic knowledge, while retaining the two-phase structure. Braine suggests that during the first phase, children attempt to identify *predicates* and *arguments* in input utterances and to further analyze the elements in each of these constituents in terms of semantic notions such as *object*, *action*, *place*, *time*, *event*, etc. Around these, the child constructs initial proto-categories which are, during the second phase, gradually extended, merged, and molded into the adult categories by using structure-dependent distributional regularities. The relevant structural knowledge is acquired without the help of innate prespecification, but rather through the repeated analysis of input utterances into predicate and arguments (and their smaller constituents).

Thus, the crucial differences between these two accounts concern the role and kinds of semantic and distributional information, as well as the sequential order in which the child begins to utilize them. Both accounts have presented criticisms against each other. Maratsos and Chalkley (1980) acknowledge the canonical semantic associations as statistical tendencies but argue that, because of their imperfections (cf. 1.1.1), a child solely relying on semantics, even if only temporarily, would be expected to produce certain types of errors which they do not find in developmental data.[16] Their own account is intended to replace such a *semantics-first* view on category acquisition but not semantics per se. It explicitly incorporates semantic cues but focuses on the semantic functions of the patterns in which a word can appear, rather than on the word's own inherent semantics.

Pinker (1984:47-50), in turn, brought forward a whole collection of objections against their account (some of these are discussed at greater detail in 5.1). In a nutshell, the first one concerns the general issue of overgeneralization and negative evidence (cf. 1.2.1). The second states that a blind search for correlations among semantic-distributional patterns inevitably leads to a "combinatorial explosion" which is problematic in face of finite processing capacities. Pinker also warns that correlating overt distributional patterns would detect "spurious correlations" arising from different syntactic structures with parallel surface word ordering (such as *"John eats meat."* vs. *"John eats slowly."*). He further asserts that overt lexical-distributional cues "are in general linguistically irrelevant" while "relevant properties are abstract, pertaining to phrase structure configurations, syntactic categories, grammatical relations, and so on". Finally, Pinker argues that "even looking for correlations among linguistically relevant properties is unnecessarily wasteful" because many potential correlations simply do not occur in natural languages.

Given these fundamental objections, it is not surprising that neither theory is still advocated by their original proponents. Maratsos (1990) revised the original account by Maratsos and Chalkley (1980), making a full U-turn with respect to the noun category while essentially reiterating the original claims for other categories. He argues that the noun category is special in that it is in fact defined and acquired much better in terms of inherent semantics (concrete object) than in terms of correlated semantic-distributional patterns.[17] Pinker (1987) presented an extensive list of severe problems with his own

[16] This argument is challenged by Macnamara (1982:134f).

[17] Maratsos (1990) points out that all words denoting a concrete object are nouns; hence, the semantic association constitutes a reliable criterion that it is sufficient, though not necessary for nounhood. The

bootstrapping proposal (Pinker, 1984) which he subsequently abandons altogether, concluding that all proposed types of information potentially contribute to the acquisition process.

These conclusions mark a gradual shift in the field. It now appears to be widely recognized that all sources of information might be relevant at any point during category acquisition; that different cues can interact with each other in multidirectional ways; and that languages and categories might rely differently on the various sources. The focus of the field has thus shifted from generating hypotheses about the theoretical primacy of one source over another, to empirically identifying the cues that the input actually does provide, and to assessing their usefulness for acquisition. No single source of information will suffice in isolation, but the role and interplay of the proposed sources are still far from being understood. Therefore, it is a viable research strategy to start by investigating the different sources separately before considering their interactions. Identifying the particular cues within one such source in isolation would therefore be an important first step; and as was pointed out before (cf. 1.2.1), it is relevant for both nativist and empiricist perspectives.

1.3 Automated distributional models

The dissertation at hand is not intended to outline another account of category acquisition, nor does it extend the earlier theoretical debates between opposing accounts. Instead, I here focus on one particular type of information in the input — namely, overt lexical-distributional regularities — and explore its potential contribution to the acquisition of lexical categories. While earlier controversies about distributional cues largely relied on claims about the kinds of cues that the input is expected to provide *in principle*, I here present an empirical assessment of the input that a child actually gets to hear, using an automated model and other computational tools to extract and evaluate distributional cues. Furthermore, the theoretical debates tended to pick out

corresponding semantic criteria for other categories, in contrast, are neither sufficient nor necessary. For instance, a word denoting an action need not be a verb, and a verb need not denote an action (cf. 1.1.1). With respect to the distributional problems that Maratsos reports for the noun category, it should be noted that some of them arise from the fact that he includes pronouns as one of its subclasses. He argues that if they are treated as a separate category, one runs into problems as well because they share many, though not all, distributional properties with first names.

a few individual cues and discuss their effects in isolation; the empirical approach, by contrast, allows for evaluating a large spectrum of potential distributional cues and their interaction, and it introduces an important dimension into the discussion, namely that of *usage frequency*. Without such information, there is a potential danger of vastly overestimating the importance of certain cues discussed in the literature, and of underestimating — or even failing to notice — other cues that arise only in spoken language and particularly in CDS.

The distributional model and evaluation methods are introduced in chapter 3. In the current section, I first characterize the particular notion of overt lexical-distributional information to be implemented in the model (1.3.1). Subsequently, other automated approaches and their central findings are reviewed (1.3.2). The section concludes with an outline of the ways in which the dissertation is intended to go beyond these previous studies (1.3.3).

1.3.1 Highly local distributional information

The characteristics of the kind of overt distributional information that I investigate are best understood by considering how it contrasts with the most well-known notion of such information, viz., the one introduced by Maratsos and Chalkley (1980; cf. 1.2.2). First of all, Maratsos and Chalkley presuppose words to be already dissected into stems and suffixes (where applicable) before the distributional method is carried out; I make no such assumption but instead take observable word forms as input to the model.[18] In consequence, morphological cues like the ones proposed by Maratsos and Chalkley play no role in the model; only distributional information pertaining to lexical co-occurrence and serial position are considered (cf. 1.2.2). Maratsos and Chalkley focus on co-occurrences only with function words and only when these co-occurrences encode some semantic function; but co-occurrences with content words and other function words may provide useful cues on category membership, as well. Therefore, no co-occurrences are excluded by linguistic criteria on the words involved; in particular, any semantic functions of co-occurrences are ignored. However, a word's co-occurrences are restricted in the sense that they are only recorded within the immediate lexical environment around this word (by default, defined as the two nearest words to either side; cf. 3.1.1).

[18] Nevertheless, like most work in the field, I *do* assume that the input speech stream is already segmented into words and utterances (cf. 2.1.1).

This highly local notion of overt lexical-distributional information that was roughly characterized above is henceforth referred to plainly as *distributional information*, in order to keep terminology simple. But why would such general information be of any use in category acquisition? The basic idea is that words of the same category can essentially occupy the same syntactic slots in utterances, and that these categories can in part be characterized by these slots. The hypothesis made here is that this relation translates from syntactic slots to structure-independent local contexts; that is, words of the same category tend to occur in the same kinds of local lexical contexts and, conversely, words which tend to occur in the same kinds of local contexts also tend to belong to the same category.

The basic idea essentially restates the main tenet of *distributional analysis* (DA) which is a categorization method of structural linguistics (cf. p. 6). However, it is important to note that, despite some commonalities, the kind of distributional information extracted by the model differs in at least four crucial ways from that underlying DA. First, DA relies on grammaticality judgments, which, in principle, cannot be performed by a child and therefore play no role in the model.[19] Second and closely related is the fact that the grammaticality judgments used by a linguist presuppose access, in principle, to an infinite set of sentences whereas the child has to solve her categorization task with a finite (though large) set of utterances. Correspondingly, the model operates on a subsample of the finite input to one particular child. A third important difference is that the model assesses distributional regularities in terms of highly local lexical contexts (as described above) while DA considers complete sentences. Each of these three differences makes the model less powerful than DA. The only way in which it actually provides a potentially more useful kind of information is by the fact that it construes distribution in terms of observed *usage frequency* (How often does a word occur in its various contexts?) rather than *grammatical possibility* (Which contexts can a word occur in?) like in DA. As I will argue in the general discussion, usage frequency information can in some sense serve the same function for the model (and putatively also for the child) as do grammaticality tests for the linguist.

[19] DA involves grammaticality judgments of two different sorts. Two words are assigned to the same substitution class only if they can be substituted for each other in a given set of sentence frames such that (i) the relevant sentences are grammatically well-formed for both words, and (ii) in each of these sentences, the two words take on the same grammatical role — without the need to specify what exactly that role is. In Fries' terms, this second test verifies that a substitution does not alter the *structural meanings* in the sentence (Fries, 1952/1957:56,74).

1.3.2 Previous approaches

The notion of distributional information described in the previous subsection has been investigated in a number of studies, beginning with pioneering work by Kiss (1973). He implemented this notion in a spreading-activation network model and proposed further procedures to evaluate the distributional information it extracts. A number of more recent studies have essentially adopted this general paradigm but replaced Kiss' particular model by a family of *co-occurrence statistical models* that in effect extract the same information (e.g., Brill, 1993; Redington, Chater, & Finch, 1998; Mintz, Newport, & Bever, 2002).

The general paradigm consists of four distinct steps. These are merely outlined here; a detailed description for the particular model employed in this study is provided in chapter 3. The first step concerns the model proper. In essence, it extracts for each word its distributional properties from a corpus of CDS and stores these in the form of a high-dimensional vector which is called a *co-occurrence vector*. In the second step, the distributional similarity between any two words (i.e., the extent to which they have similar distributional properties) is quantified by applying some formal similarity measure to their co-occurrence vectors. In the next step, the resulting values are fed into a *hierarchical clustering algorithm* to derive a clustering tree (called a *dendrogram*) which captures the overall similarity structure among all words considered. In the final step, the structure in this dendrogram is assessed as to how well it reflects the category structure of the target language. Most commonly, this is done by cutting the tree at a particular level of similarity to derive a set of discrete classes of words (viz., the cut-off branches) which can then be evaluated — in terms of a quantitative evaluation score — against some independent and linguistically motivated approximation of the *true* categories.[20]

Within this general paradigm, a lot of variation is found across studies. In the first step, for instance, approaches differ with respect to how they define a word's distributional context (ranging from one to eight neighboring words to either side of this word) while in the remaining three steps a variety of similarity measures, clustering algorithms, and evaluation scores have been used. Moreover, studies differ as to whether they are designed to primarily evaluate the computational performance of the overall model (e.g., Brill, 1993; Hughes & Atwell, 1994) or rather the structure of the

[20] In less ambitious studies, however, researchers solely rely on the ad-hoc intuitions of a native speaker who directly inspects the tree structure — a strategy which Hughes and Atwell (1994:534) humorously disqualify as a "*looks good to me* approach".

input (e.g., Redington et al., 1998; Mintz et al., 2002). While the former group of studies is generally concerned with developing machine learning applications such as *electronic parsers* and *part-of-speech taggers*, the two studies in the latter group are explicitly concerned with category acquisition by children and constitute the main references on which the approach presented here is largely built. Throughout the dissertation, I therefore frequently compare aspects of methodology, data basis, and results with these two studies.

Finally, two other distributional approaches should be mentioned that are viable alternatives to co-occurrence statistics. The first concerns connectionist models, most prominently the *simple recurrent network* (SRN) architecture introduced by Elman (1990). SRNs can detect regularities in sequential data such as language, by attempting to predict the next item in the sequence given the previous items, and by gradually learning from mismatches between predicted and actually observed items. As far as learning is successful, these models develop similar internal representations for words that, everything else being equal, tend to make similar predictions about their likely successors. These internal representations thus arise from a complex notion of distributional regularities in the input; and it was shown in simulations on artificial miniature languages that their similarity structure closely reflects the lexical categories underlying the language (Elman, 1990, 1991). It is still an open question whether SRNs will readily scale up to a sizeable sample of a full-blown natural language. If this scaling problem can be solved, a connectionist model is clearly preferable over co-occurrence statistical models, as a tentative account of how the child might exploit the distributional regularities in the input. However, the deliberate goal of this dissertation was to leave aside such questions about the learning mechanism and to identify and assess the cues that are actually *available* in the input. And with a co-occurrence approach, these cues can be accessed more directly, and thus evaluated more systematically, than with SRNs.

The second alternative approach is unique in several ways. It is a model developed by Cartwright and Brent (1997) which implements a nonlocal notion of distributional information but is nevertheless mentioned here for three reasons. First, the model incrementally builds a category system such that at each point during development it entertains a hypothesis about the categories to acquire. Second, it simultaneously records and updates the (increasingly abstract) sentential contexts in which these categories can be used. This is an attractive feature of the model because it directly links category acquisition with some rudimentary syntax acquisition and thereby combines overt with structure-dependent distributional cues. The third reason pertains to the fact

that the model is in principle capable of dealing with categorial ambiguity — an issue that I shall return to later.

1.3.3 Goals for the current approach

Both studies by Redington et al. (1998) and Mintz et al. (2002) successfully demonstrated that overt and highly local distributional regularities in the linguistic input to children provide a considerable amount of information about lexical categories. As a starting point, this dissertation was intended to verify this general result but also to go beyond the former two studies in at least four different ways.

One goal was to apply the co-occurrence approach to a language other than English which was the target language in both previous studies (and, in fact, in all distributional approaches cited in 1.3.2). Because models exploiting lexical co-occurrence crucially rely on the ordering of word tokens, it was not clear to which extent they would also work for languages with less restricted word order, as for instance German (cf. Redington et al., p. 463).[21] Another potential challenge arises from the fact that inflectional morphology is more complex for German than for English. Thus, in German, inflecting lexemes (particularly verbs, pronouns, and determiners) generally have a greater number of different forms which individually occur less frequently but nevertheless all have to be discovered to belong to the same category. The consequences for a co-occurrence model were not clear a priori.

A second goal was to apply the co-occurrence approach to a presumably more realistic data sample. Redington et al. combined several hundred individual corpora (recorded at different places and time periods) to one large amalgam of linguistic input that is likely to contain a variety of spoken language that realistically no individual child would ever encounter. This problem can be avoided by analyzing the input to individual children separately. Mintz et al. did precisely that but in consequence used relatively small data sets (roughly between 7,000 and 20,000 utterances per child). To work with language data that are more realistic both in terms of size and variety, I took advantage

[21] It was not until very recently that researchers have begun to analyze distributional cues in the linguistic input to children for several other languages, including French and Dutch (Monaghan, Christiansen, & Chater, 2004). Previously, Martin Redington and colleagues also tested their method on corpora of written adult speech for Mandarin Chinese (Redington et al., 1995) and German (personal communication, September 27, 2004) and found it to yield a substantial amount of information regarding the system of lexical categories in these two languages. To my knowledge, however, no study has yet systematically explored the particular distributional regularities in German child-directed speech.

of a so-called *high-density* corpus in which the recordings cover a long period at an exceptionally high sampling rate (cf. section 2.1).

In evaluating the usefulness of distributional information, Redington et al. focused on the category structure as a whole, with only one of their experiments looking at the individual categories in isolation. Mintz et al. consistently performed such category-specific analyses throughout their study but only considered the categories noun and verb. A third goal was therefore to combine the advantages of both studies by considering an exhaustive set of categories (exhaustive in the sense that it covers the entire lexicon) and systematically assessing the usefulness of distributional cues for each individual category in isolation.

The fourth goal was to evaluate the raw distributional similarities between words as directly as possible, without any dispensable intermediate transformations that might distort the overall similarity structure. In the four-step evaluation paradigm that both previous studies followed (cf. 1.3.2), such a transformation is performed by the hierarchical cluster analysis (step 3). This intermediate step can be problematic because the resulting clusters largely depend on the choice of a particular clustering algorithm. Following Zavrel (1996), I therefore decided to drop this third step altogether and use an evaluation score (step 4) that operates directly on the similarity values produced by the second step. Note that because there is a considerable variety of potential similarity measures and evaluation scores, the choice of one particular measure and score can also influence the final results. But unlike the third step, steps 2 and 4 cannot be dispensed with; and as a precaution, several candidates were tested to select measures and scores that are well-suited given a number of considerations (cf. sections 3.2 and 3.3).

To anticipate possible misinterpretations, I do not mean to discredit cluster analysis as a tool for inspecting and visualizing some fundamental structure underlying complex data. In fact, even in the given situation, the clustering analysis in step 3 has the nice advantage that it allows for inferring from the input discrete classes of words which can be taken as the model's guess about the *true* categories of the target language. However, as was stressed before, this dissertation is deliberately not concerned with how categories are actually acquired; and furthermore, it is very unlikely that children would rely on formal co-occurrence statistics and clustering algorithms.[22] Thus, unlike the previous studies, the current work does not actually build any lexical categories from

[22] Correspondingly, the authors of both previous studies explicitly do not claim to model the mechanism by which children acquire categories.

the input. But as a payoff, the modified approach provides a more natural way of tracing any results about category distinctions that are reflected particularly well (or poorly) by distributional information, back to specific distributional cues that might cause these results (cf. sections 4.3 and 4.4). Ultimately, such links could help to devise more realistic learning mechanisms that build on the kinds of distributional cues the child is likely to attend to.

These four central goals pertain to the input data (language and data sampling) and the evaluation procedures (category focus and immediacy of evaluation) such that it becomes difficult to compare any results of this dissertation with those of the previous studies. Therefore, I will repeatedly refer to these differences and provide intuitions about how the results might relate to each other across studies. Having said this, it should be emphasized that I do not aim to demonstrate the superiority of one approach over another — quite on the contrary. For instance, results that are found to hold across studies, *despite* the different data bases and methodologies, can be taken as solid evidence about certain properties of distributional information.

In addition to the central goals above, three further goals were pursued that concern supplementary analyses rather than data and methodology. First, the issue of robustness of the information is addressed from several different perspectives. For instance, children might not always pay attention to their input, or they might only use cues from reasonably frequent co-occurrences. This raises the question how the distributional cues and their reliability are affected if such possibilities are simulated in the model. In section 4.2, I present several such simulations, some of which were also inspired by the two previous studies.

The second additional goal was to shed some light on certain patterns of acquisition as identified in the psycholinguistic literature (cf. sections 4.4 and 4.5). For instance, experimental evidence suggests that, for English, the noun category is acquired prior to the verb category (Tomasello & Olguin, 1993; Olguin & Tomasello, 1993). Detailed analyses of these two categories in German will reveal a fundamental disadvantage of the verb category with respect to distributional cues, and this finding might help to explain such developmental asynchrony, provided that it holds for German as well. Related findings of the two previous studies will be discussed.

Finally, in subsection 1.2.2, I referred to various theoretical objections that have been made against the general usefulness of overt distributional cues. As it turns out, only three of these potentially apply to the notion of distributional information

investigated here. As a final goal, this dissertation therefore serves as an empirical assessment of these objections (cf. discussion in section 5.1).

Chapter 2
Language material

This chapter presents language materials of two quite different types. Section 2.1 describes the corpus of CDS data from which distributional regularities were extracted. Section 2.2 provides a set of benchmark categories which was employed to assess how useful the distributional regularities in the corpus might be for acquiring these categories. Thus, for the purpose of the dissertation, these two language resources — benchmark and corpus — served as provisional answers to the questions *What is to be acquired?* and *What is the evidence?*, respectively.

2.1 Corpus

As was stated earlier, one goal was to study the distributional regularities in German CDS using a more realistic data base than previous studies. For this purpose, I utilized a so-called *high-density* corpus that was built at the Max Planck Institute (MPI) for Evolutionary Anthropology in Leipzig, Germany. It documents the linguistic interaction between the boy *Leo* — a monolingual first-born child of a middle-class family speaking standard High German — and his environment at a very high sampling rate (Behrens, 2002). The recordings started on the boy's second birthday just after he began to produce multi-word utterances. They span a period of three years with five one-hour sessions every week during the first year and five one-hour sessions every month during the second and third year of the study. A few minor deviations from this sampling pattern occurred during the first year, for instance when the boy was sick or when the recording equipment did not work properly. Moreover, 10 instead of five sessions were recorded for one month in the second year. Overall, this yielded a total recording time of 377 hours (252 hours in the first year, 65 in the second, and 60 in the third).

Longitudinal data covering such a long period at this high sampling rate offer unprecedented possibilities for linguistic and psycholinguistic research. The only drawback of using this data source is, however, that, to my knowledge, there currently exists no second corpus of German CDS that would be at least roughly comparable in size and density. For the particular purposes of this dissertation, it did not seem reasonable to compare distributional patterns in the *Leo* corpus with those in much smaller corpora. It was therefore decided to approach the general research questions that were laid out earlier in form of a case study. Thus, the presumably more representative sample of the language addressed to one particular child in turn raises the question how representative the results of the study might be with respect to the language input of German children in general. Given the currently available corpora, there seems to be a trade-off between these two levels of representativeness.

At the MPI, the recorded data were transcribed into CHAT format (MacWhinney, 2000), further complying with a set of transcription guidelines spelled out for the particular demands of German (cf. Behrens, 2002). Of this large corpus, I extracted the actual input data comprising all utterances that were produced by anyone other than Leo. These involve a total of 14 different speakers ,but the lion's share of utterances (97.8%) was contributed by the boy's mother (his primary caregiver; 63.0%), his father (21.9%), and one particular research assistant of the MPI (12.9%) who came to baby-sit the boy once or twice every week and therefore was a part of his natural environment.

Conversations entirely between adults had not been transcribed; but the corpus nevertheless contains some adult-to-adult utterances which occurred within conversational sequences that did involve the boy. Inspection of random samples of the full corpus suggests that these constitute less than 1% of all input utterances. For this reason, I make no sharp distinction between the terms *child-directed speech* and *linguistic input* when referring to the *Leo* corpus.

By using these transcribed corpus data, I implicitly make the assumption that the child first segments the input speech stream into words and utterances before exploiting lexical-distributional regularities. Some at least rudimentary word segmentation mechanism is a logical prerequisite for even considering lexical-distributional regularities, and, in fact, for any acquisition of lexical categories to take place. In general, infants become quite skilled at word segmentation during the second half of their first year and nearly reach adult performance before their second birthday (for review see Jusczyk, 1999) which is around the time that the first categories appear to become productive.

The issue of what counts as an utterance in fluent speech is not always clear to decide. In many cases, utterances are unmistakably delimited by change of turn between speakers, or by extended pauses between word sequences uttered by the same speaker. In less obvious cases, transcribers relied on prosodic patterns and the connectivity of syntactic units (clauses and phrases) to make their transcription decisions. Insofar as infants begin early to exploit such prosodic patterns (for review see Jusczyk, 1997, 1998), it is reasonable to assume that they infer utterance boundaries that roughly correspond to those in the corpus.[23] But even if the child would not always note the same word and utterance boundaries, a rough correspondence to the transcriptions should suffice for the given purposes. After all, the distributional information to be derived from this corpus is probabilistic in nature.

The extracted input utterances still included a rich set of transcription codes most of which register various linguistic and extra-linguistic aspects of utterances (e.g., speaker identification, pauses within an utterance, intonational contour, and temporal overlap between utterances) that are not captured by standard orthography. These transcription codes had to be removed or resolved before the raw input data could be evaluated electronically. The precise *decoding* manipulations are listed and discussed in subsection 2.1.1 before a quantitative summary of the decoded corpus data is given in 2.1.2. In the final subsection (2.1.3), I derive from the corpus a set of *target words* which are those words for which distributional information was investigated.

2.1.1 Decoding the transcribed data

Each utterance in the corpus comes tagged with a symbol denoting the speaker, and with a reference linking to the corresponding portion of the actual recording. These specifications were simply removed. A more substantial manipulation, however, involved a few utterances that had been transcribed as *completions* of preceding utterances (standard CHAT terminology is signaled by italics). This occurred when a speaker got interrupted by another speaker such that one utterance was transcribed as two utterances by the same speaker with an intervening utterance by a second speaker. Another type of completions occurred when one speaker finished an utterance started by a second speaker while the intonational contour was typically preserved. In the

[23] Analyzing the influence of cues from utterance boundaries will demonstrate that although some categories capitalize substantially on these cues, the general usefulness of distributional information does not hinge them (cf. 4.2.3).

decoding step, both cases of utterance completions were resolved by appending the second utterance to the first. However, there were only 23 instances to which this manipulation could be applied.

After this step, the input comprised 268,638 utterances of which 17,558 (6.5%) were removed entirely for the three following reasons. The first concerns diary entries. In addition to the recordings, Leo's parents had kept a *diary* to document his acquisition of vocabulary and other linguistic knowledge, but also of peculiar errors he produced. These diary entries had been transcribed and included in the corpus but were explicitly marked as such. While diary entries were typically utterances by the boy, adult utterances had occasionally been provided as discourse context. There were only 54 (.02%) such utterances in the input corpus which were thus removed.

As a second group, 8,779 (3.3%) of all utterances were removed because they were unsubstantial, i.e., they contained no actual words but possibly material that was either unintelligible or not transcribed in order to protect the family's privacy.[24] And the third group consists of 8725 (3.2%) utterances that were removed for containing partly unintelligible material or phonological fragments. *Phonological fragments* arise when a speaker starts to utter a word but only produces a first sound, while *unintelligible material* can involve parts of words, single words, or groups of words. It is not clear whether the material was unintelligible in the actual situation as well or only in the recording. In any case, these utterances were removed because analyzing co-occurrence relations between unintelligible items would not be meaningful — even less so as it is unclear whether they represent a word or multiple words. Moreover, it seems reasonable to assume that utterances of this third group are randomly distributed such that their removal should not render the data sample less representative of Leo's full input.[25]

The transcription codes in the remaining 251,080 utterances were handled as follows. Several types of CHAT codes mark unintentional repetition of parts of an utterance (*multiple attempts, repetitions*). A total of 11,208 such cases were found in the data, and for each of them, only one complete copy of the repeated material was retained. The rationale here was that presumably, unintended repetitions are marked not only by their redundancy but also by prosodic patterns. Since children have been shown

[24] The test for substantiality was made only after all transcription codes had been resolved.

[25] In fact, rather the opposite is the case — for material that was unintelligible only on the recording but not in the actual situation, it is the inclusion and not the removal of such utterances that would provide the model with an unrealistically bad sample of the boy's real input.

to exploit a wide range of prosodic cues before their first birthday (for review see Jusczyk, 1997), it seems a plausible assumption that they can filter out such repetitions.

A similar argument holds for self-corrections and reformulations (in CHAT coded as *retracements*). In such cases (4,384 overall), the corrected or rephrased parts were removed and only the version that was uttered last was retained. Self-corrections typically correlate with a sudden rise in intonation and other prosodic markers which highlight the corrected material as not being meant for interpretation.

At the level of individual words, speakers occasionally show sloppy or nonstandard articulation. This can lead to sounds being altered (CHAT: *assimilations*) or omitted (*shortenings*). In the corpus there were 60,989 instances of word shortening which most often involved the form *is* (for *ist*; English: *is*) and *nich* (for *nicht*; English: *not*). Word assimilations occurred 2,902 times and most frequently concerned *mer* (for *wir*; occasionally also for *man*; English: *we* or one_{pronoun}, respectively) and *nix* (for *nichts*; English: *nothing*). In all these cases, the deviating forms were replaced by their standard versions. The rationale here was that in most cases, the deviations only involved minor phonological operations such that the child could potentially identify the deviating and standard forms as phonological variants of each other. For particular cases like *mer* for *wir*, one can arguably find other relations by which the child could link them; especially because the replacements involved only a few forms that occurred fairly frequently in the corpus such that the child has abundant evidence for detecting them.[26]

The transcriptions contained a number of additional codes which, unlike some of those discussed above, could be removed without altering anything about the words that had actually been uttered by the speaker. Hence, these markers were simply stripped off. As a final manipulation, all nonfinal punctuation (i.e., commas) and capitalization was removed entirely, that is, all upper case letters were converted to lower case because this distinction is not marked in the phonological input either. After these decoding steps, only plain words and utterance-final punctuation (.?!) remained in the corpus.

Most of these decoding procedures should be entirely uncontroversial because they only involve the removal of tags that are useful for the investigator but that do not correspond to information explicitly available to the child. However, some manipulations — those that were described above in more detail — did alter the

[26] It should be noted that 14.3% of the occurrences of *mer* and 40.7% of the occurrences of *nix* had not been transcribed as assimilations. These occurrences were therefore retained as such by the automatic decoding procedures. Thus, due to these minor transcription inconsistencies, the decoding of *mer* and *nix* only involved the shifting of some occurrences to other word forms without entirely removing these assimilations as distinct forms from the analysis.

particular words that the child has actually heard. Although I believe that each of these manipulations reflects plausible assumptions about the child's filtering capacities well before he develops productive lexical categories (around and after his second birthday), it was important to address the concern that these manipulations may ameliorate the outcome of this study. For this reason, an additional line of analysis was conducted using a more conservative decoding scheme in which utterance completions were ignored, unintelligible material was retained, repetitions, multiple attempts, and retracements were retained as such, and shortenings and assimilations were left uncorrected. Relative to this conservative decoding scheme, the distributional consequences of the standard decoding turn out to be marginal (cf. 4.2.1).

2.1.2 Descriptive statistics

The standard decoding scheme described above yielded a CDS corpus (henceforth simply referred to as *the corpus*) comprising a total of 251,080 utterances, 1,282,051 word tokens, and 30,232 different word types. Note that the term *word type* is used here (and from now on) to denote types of syntactic word forms rather than types of lexemes; e.g., different inflected forms of the same verb are treated as different word types. The motivation for this nonstandard definition was that word forms rather than lexemes are the observable entities in the boy's input (cf. 1.3.1). In consequence, word forms in the corpus were retained as such and not transformed to their base forms.

The global frequencies of these word form types displayed the typical *Zipf distribution*, with relatively few words occurring extremely often and the bulk of words being extremely rare. As an illustration, the single most frequent word form, *das* (English: *the*$_{\text{neut.}}$, *that*$_{\text{neut.}}$) occurred 38,518 times overall whereas 14,916 (49.3%) of all word forms were found only once in the entire corpus. To visualize the skewed distribution despite these extreme differences in *base frequency* (i.e., frequency of occurrence), Figure 2-1 plots the cumulative frequencies, as a proportion of the number of all word tokens. The steep initial rise of this chart reflects that the bulk of word tokens belong to a small minority of (high-frequency) word types. For instance, the 100 most frequent word types (0.3% of all word types) together account for 61% of all word tokens in the corpus; and the 1,491 (4.9%) most frequent word types account for 90% of all word tokens.

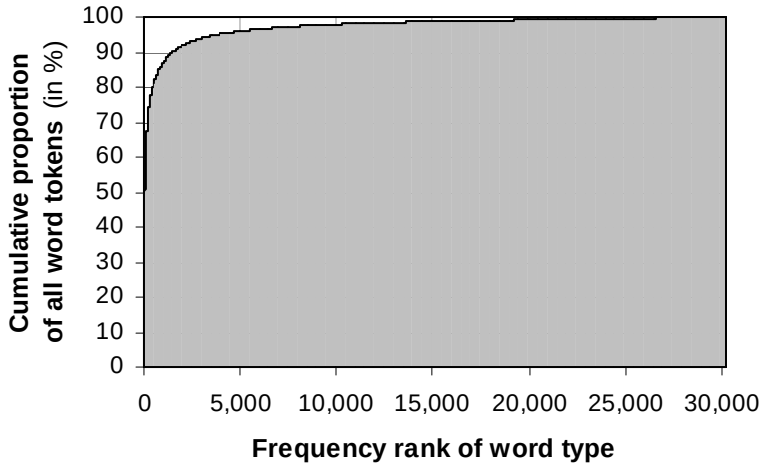

Figure 2-1: *Frequency distribution of all word types*

Frequencies are presented cumulatively and relative to the total number of word tokens, i.e., as the proportion of word tokens accounted for by the *r* most frequent word types (*r* = frequency rank).

Moving from word types to utterances: Of the 251,080 utterances in the corpus, 147,637 (58.8%) terminate on a period character, 88,614 (35.3%) on a question mark, and 7,005 (2.8%) on an exclamation mark. The remaining 7,824 (3.1%) utterances carry no final punctuation.[27] On average, the input utterances comprise 5.1 words. The distribution of utterance lengths is summarized in Figure 2-2. By far most common are one-word utterances (21.2%), and 77.9% of all utterances contain seven words or less.

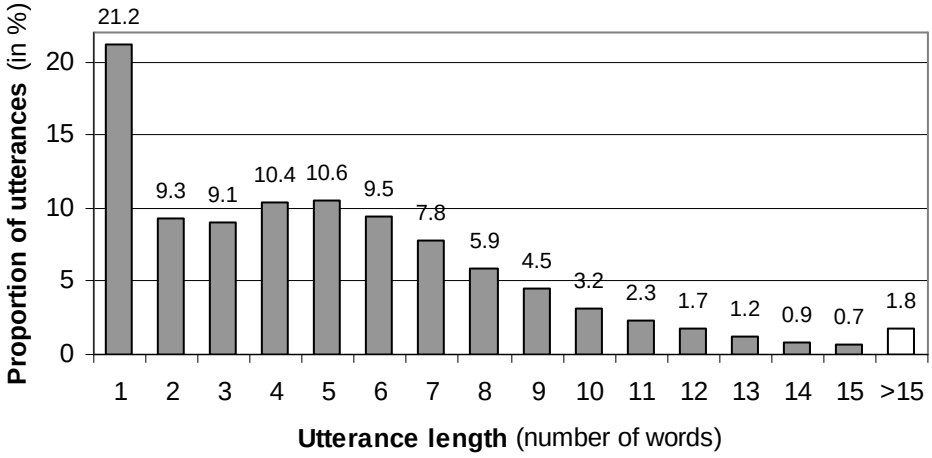

Figure 2-2: *Distribution of utterance lengths*

Utterances with more than 15 words are grouped together.

[27] As it is unclear to which extent the child detects such distinctions of utterance termination and arrives at similar decisions as the transcribers, I explored the distributional effects of punctuation (cf. 4.2.3).

But this distribution is not constant over time. As the child gets older, not only his own language productions get longer, on average, but also the utterances that he gets to hear. This presumably reflects a gradual increase of the proportion of complex messages addressed to him by his caregivers. Figure 2-3 shows how the *mean length of utterance* (MLU; here defined as the average number of words) changes over time for the input data and for the child's own utterances.[28] The steep ascent for the child's utterances during most of his third year of life documents his development from the one-word stage well into multi-word speech — during this same period, the average length of input utterances remains fairly constant. However, in the following two years, the increase in MLU slows down considerably for the boy's utterances whereas a trend towards longer utterances can be observed for the input data.

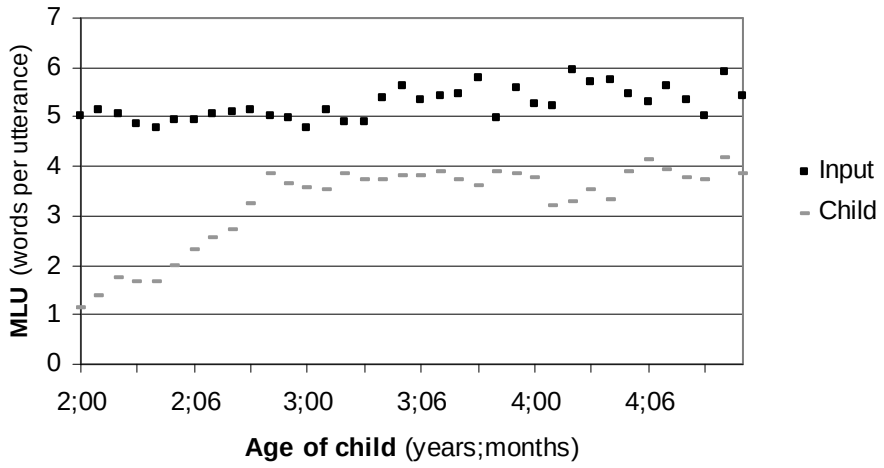

Figure 2-3: *Mean length of input utterances as a function of the child's age*

For comparison, the corresponding curve for the boy's own productions is included. The greater variation of MLU values in the period after 3;00 are due to the lower sampling rate.

2.1.3 Target words

To approach the main research questions that were laid out earlier, I analyzed and compared the distributional properties of word types in the corpus. However, this cannot be reasonably done for word types that occur very infrequently in the corpus because the contexts which they happen to occur in for these few instances may not reliably represent the full spectrum of their distributional properties. For this reason, the

[28] The child's productions were extracted from the original corpus by applying the same decoding scheme as for the input data.

analyses of distributional information were deliberately restricted to those word types that occur at least 100 times in the corpus. The word types selected by this criterion will be referred to as *target words*. The corpus contains 1,017 such target words; and while these represent only 3.4% of all word types, they together account for 87.1% of all word tokens in the corpus.[29] Their frequency distribution displays the before-mentioned Zipf pattern (Figure 2-4). Base frequencies range from 100 (by definition) to 38,518 tokens, with average frequency 1,098.3 and median frequency 237.0.

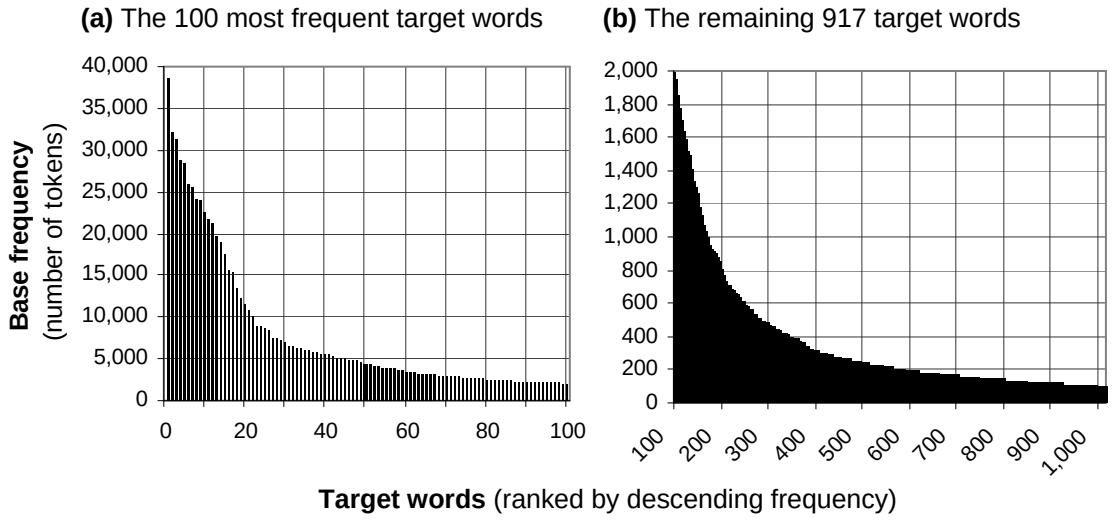

Figure 2-4: *Frequency distribution of target words*

To accommodate their wide range, the base frequencies of target words are displayed separately for highly frequent (a) and less frequent target words (b).

2.2 Benchmark categories

To be able to assess how much the distributional information present in the *Leo* corpus could have taught the child about the lexical categories of the German language, it was necessary to specify before-hand what the right categories are that Leo was supposed to acquire. To this end, a *benchmark* or *gold standard* of word classifications was built against which the model was evaluated. It is absolutely critical that the benchmark

[29] As a comparison: Redington et al. (1998) used for their standard analysis the 1,000 most frequent word types while Mintz et al. (2002) used the 200 most frequent word types. Minimal frequencies of occurrence for these word types were not reported. The former study worked with a corpus of roughly 2.5 million word tokens and the latter with corpora of between 20,000 and 75,000 word tokens.

classifications were *not* available to the model at any time. Just as crucial is the converse requirement, namely that the benchmark is based on linguistic evidence independent of the model.

In the current section, I present and justify the particular categories that were distinguished at the top level (subsection 2.2.1); I then describe how these categories were in practice assigned to target words (2.2.2) and quantitatively summarize the composition of the resulting categories (2.2.3). The section concludes with a few comments, linking back to representational issues that were discussed in 1.1.3.

2.2.1 Building the benchmark category system

Section 1.1 offered but an incomplete picture of past and ongoing debates among grammarians and psycholinguists concerning the status (Which role do they play in linguistic theory and the mind?) and nature (What do they look like?) of lexical categories, as well as the questions which particular categories are to be distinguished and by what criteria they are to be defined. Given these unsettled issues, it seemed reasonable to resort to traditional category distinctions, with all their shortcomings — including defining criteria that are partly inconsistent. The benchmark described below is therefore not intended to be a precise representation of the categories that are used by adults and acquired by children. But it can be construed as an approximation of these categories and employed as a heuristic for evaluating distributional information.

The benchmark distinguishes the following 11 categories (abbreviating category labels in parentheses): the four *open classes* of verbs (V), nouns (N), adjectives (ADJ), and adverbs (ADV); six *closed classes* of interrogative words (INTG), pronouns (PRON), determiners (DET), prepositions (PREP), conjunctions (CONJ), and particles (PTCL); and finally the category of interjections (INTJ) which has a special status outside the open–closed class distinction.[30] The remainder of this subsection documents the external resources and further considerations on which the decision for this category system was based.

[30] Interjections are often treated as closed class words because they are typically short (one or two syllables in German, cf. Nübling, 2004) and carry little, if any, referential meaning, and many of them occur very frequently — characteristic properties of closed classes. On the other hand, they also display properties characteristic of open classes in that new interjections are created more readily than, for instance, new prepositions or conjunctions. Furthermore, their overall frequency distribution in fact resembles more closely that of open classes, in particular that of adverbs (cf. Table 2-1, p. 50).

While contemporary grammarians generally recognize the four open classes V, N, ADJ, ADV as independent categories within the German language, there is considerable variation as to which closed classes are distinguished at the top level. The distinctions used here are based on two standard books of German grammar, *Duden: Grammatik der deutschen Gegenwartssprache* (1998; henceforth DUDEN) as well as Hentschel and Weydt (1994).

Reconciling between these two sources involved regrouping a few subclasses. Hentschel and Weydt group determiners and pronouns together as one category, reflecting the observation that possessives, demonstratives, and numerals can function either as a pronoun or as a determiner. In contrast, DUDEN distinguishes pronouns and determiners as independent categories but treats numerals as special adjectives.[31] As a compromise, I chose to keep determiners and pronouns separate and to classify most numerals, possessives, and demonstratives as determiners (e.g., *zwei*, *viele*, *mein*, *dieser*; English: *two*, *many*, *my*, *this*$_{\text{masc.sg.}}$) and in some cases as pronouns (e.g., *eins*, *dies*; English: *one*, *this*$_{\text{uninflected}}$), depending on the predominant usage of these items in the corpus (cf. subsection 2.2.2). In general, the forms treated as determiners were inflected whereas those assigned to the pronoun category were not.

Regrouping was also required for prepositions and conjunctions. Whereas Hentschel and Weydt treat them as subclasses of the particle category, I followed DUDEN where they are granted the status of separate categories.

Another issue concerns interjections, a term that I use with a wider scope than common, to capture prototypical interjections, onomatopoeia, conversational particles, and response particles, as well as greetings and other conventionalized exclamations. With the exception of response particles, these words fall into what Nübling (2004) calls the *interjectional spectrum*. In her thorough analysis, she identified a set of (predominantly) functional dimensions along which the words within this spectrum differ systematically — giving rise to various subclasses of interjectional words — but found their formal properties to display some substantial commonalities. Most strikingly, interjectional words are syntactically autonomous in the sense that they, in general, do not interact syntactically with other elements in a sentence. Very often, these words occur as one-word utterances all by themselves, or right before a complete sentence without being integrated in its syntactic structure (e.g., *naja* in "*Naja*, *das ist*

[31] This contrast illustrates the subclass problem that was described in 1.1.1.

nicht neu.”; English: *“Well, this is not new.”*).[32] Because they share this status as *sentence equivalents* with response particles, I decided to join these two groups of words in one single category which I shall refer to as *interjections*, for the sake of simplicity and because interjection words constitute the bulk of its members.[33] Note that by treating these words as a category of their own right, I went beyond both grammatical sources (DUDEN and Hentschel & Weydt) which classified them as particular kinds of particles.

A second category was introduced in the benchmark, namely that of interrogative words. This class comprises a relatively small set of words that DUDEN as well as Hentschel and Weydt treat as subclasses of two distinct categories, namely, as interrogative or modal adverbs (e.g., *wie, wo*; English: *how, where*) and interrogative pronouns (e.g., *wer, welcher*; English: *who, which*$_{\text{masc.sg.nom./fem.sg.dat.}}$).[34] The main motivation for joining these words to form a small category of their own was to acknowledge their salient syntactic role as lexical markers of questions when they are used in their prototypical function.

Some of the reconciliations and further modifications listed above reduced the scope of the particle category. However, it was extended by the addition of three subclasses that neither DUDEN nor Hentschel and Weydt discuss as such. The first concerns *verbal particles* which occur as separable prefixes of *particle verbs* which are quite common in German. Such verbal particles are attached to the verb when it occurs as a nonfinite form (e.g., *zusammen* in *zusammenarbeiten*; English: *to collaborate*) but are used as isolated words for all finite verb forms (e.g., *wir arbeiten zusammen*; English: *we collaborate*). Because their existence as isolated word forms depends on inflection, verbal particles are commonly not treated as lexemes and in consequence are

[32] The only exception is the small subclass of adverbial interjections which are typically embedded in syntactic structure (e.g., *schwupps* in *„Und schwupps!, war alles weg.*“; English: “And before you knew it everything was gone.“) but nevertheless also routinely occur outside of syntax.

[33] To pick up Nübling’s (2004) interjection subclasses, the classification heuristic that is described in the next subsection assigned to this category a total of 36 prototypical, or primary, interjections (e.g., *aeh, aua*; English: *um, outsch*), four secondary interjections (e.g., *oh+gott*; English: *oh god*), 16 conative interjections and greetings (e.g., *he, vorsicht, hallo*; English: *ey, watch out, hello*), four discourse markers (e.g., *aeh*; English: *uhm*), six onomatopoeia and adverbial interjections (e.g., *bumm(s)*; English: *bang*/*boom*/*boing*) and eight response particles (e.g., *ja, nein, bitteschoen*; English: *yes, no, you are welcome*).

[34] As a consequence of the categorial ambiguity between determiners and pronouns, matters are in fact even more complex. Of the subclass of interrogative pronouns, another class of interrogative determiners could be segregated, comprising words like *welcher* (English: *which*$_{\text{masc.sg.nom./fem.sg.dat.}}$) that can be used either with and without a noun or pronoun — in contrast to words such as *wer* (English: *who*) that prescribe a strict pronominal use, i.e., to occur without a noun.

not classified in DUDEN, nor by Hentschel and Weydt.[35] However, since target words are word form types rather than lexemes, verbal particles had to be classified in the benchmark. Thus, they were treated as particles.

Closely related is the second addition to the particle category, namely the relatively small subclass of *copula particles* which can serve as predicate adjectives (e.g., *los* in *"Was ist los?"*; English: *"What's up?"*) but not as adjectival attributes of nouns (cf. Engel, 1996). The third addition concerns the small subclass of question particles which can be described as lexicalized tags for tag questions (e.g., *oder* in *"Das macht Spaß, oder?"*; English: *"This is fun, isn't it?"*). These words occur quite frequently in spoken German and often take dialect-specific forms (such as *gell* or *ne*).

Finally, it should be noted that, in accordance with DUDEN and Hentschel and Weydt, proper names were classified as nouns, and similarly, auxiliaries as verbs, rather than as a special category by themselves. Inasmuch as auxiliaries form a closed class, however, this blurs the distinction between closed and open classes.

2.2.2 Classification heuristic

After having established a set of benchmark categories, mapping the 1,017 target words onto these categories involved two separate steps. First, target words were assigned to all categories that they can instantiate. However, these *complete classifications* assign many words to more than one category. This categorial ambiguity was removed in the second step to derive a single category membership for each target word which will be referred to as its *benchmark classification*. While by default the distributional analyses were based on these benchmark classifications, the complete classifications were useful to study the distributional effects of categorial ambiguity (cf. 4.1.2).

To carry out the first step, complete classifications were collected from several resources. The before-mentioned references, DUDEN as well as Hentschel and Weydt, were used again, mainly to assign words to the six closed classes distinguished in the benchmark. Both books provide extensive member lists for each category (and their various subclasses) which were used as a primary resource for these categories.[36] A few other target words were added to the different closed classes by applying the categories' defining criteria listed in these two books.

---

[35] More precisely, the verbal particles as such do not exist as independent lexemes. But they generally have homonyms belonging to other categories that have the status of lexemes.

[36] The two particular subclasses of copula particles and verbal particles were classified on the basis of Engel (1996) and WSD, respectively. WSD is introduced in the next paragraph.

Membership of the four open classes and the category of interjections was decided using two electronically available dictionaries — *Duden: Deutsches Universal-wörterbuch* (2003) which focuses on standard usage of standard words, and the online database *Wortschatz Deutsch* (Quasthoff, 1998; henceforth WSD) which has a good coverage of interjections and nonstandard usages of standard words. However, WSD classifications for the inflecting open classes V, N, and ADJ had to be used with caution. Most of these had been generated automatically by a heuristic that exploits inflectional paradigms; and because German inflectional morphology is not free of ambiguities, this resulted in WSD listing incorrect classifications for a small portion of word forms.[37] Such erroneous cases were detected using my own intuitions.

By these procedures, all 1,017 target words were covered and assigned to one or multiple categories. Overall, the degree of categorial ambiguity is considerable as a total of 384 (37.8%) target words were found to belong to more than one category, with an average of 2.4 categories per ambiguous word. The most extreme cases — *gleich* (ADV, PTCL, ADJ, CONJ, PREP, V) and *wie* (INTG, PTCL, PREP, CONJ, ADV, PRON) — can instantiate as many as six of the 11 benchmark categories.[38]

The issue of categorial ambiguity of word forms is an important one and poses a challenge to evaluating distributional and other cues to lexical categories. One popular solution is to simply assume entirely discrete categories, and to consider for each ambiguous target word only the single category that predominates its usage in some given language sample. For instance, Redington et al. (1998) determined the predominant category of their target words from an external lexical database. However, in different language registers, different categorial possibilities might predominate in the usage of a particular word; and, as Redington et al. (1998:439) suggest, it might be more appropriate to decide about predominant category membership based on usage frequencies in the corpus under investigation.[39] It is a plausible assumption that a word's more frequent usages in the child's input are likely to have more influence on what he learns about that word.

Therefore, I chose to resort to the *Leo* corpus to determine for each of the 384 ambiguous word forms its predominantly used category which was taken as the word's

---

[37] I wish to thank Uwe Quasthoff for meeting with me to explain various details about the WSD project (October 10, 2002).

[38] Some of the category assignments for these two items are at least disputable. However, they play no role in the analyses because only the unquestioned predominant category was used.

[39] Mintz et al. (2002) only classified nouns and verbs, and in ambiguous cases they also picked the predominant usage in the respective corpus.

benchmark classification. This is the second classification step mentioned above. In practice, for ambiguous word forms, I selected and inspected random samples of their occurrences. In cases where these did not yield a clearly predominant category, larger samples were scanned to make reliable decisions.

Selecting categories in this *the-winner-takes-it-all* fashion occasionally lead to surprising classifications. The German word *das*, for example, — mentioned earlier as the most frequent word form in the corpus — would typically be characterized as a neuter definite article by native speakers. However, in the corpus, *das* is used much more frequently in its function as a pronoun (in English corresponding to *that* or simply to the neuter personal pronoun *it*). Thus, according to the described heuristic, *das* was analyzed as a pronoun.

A more intricate case of categorial ambiguity involved *clitics*, that is, morphemes that have syntactic properties of a word but are phonologically bound to the following (*proclitic*) or preceding word (*enclitic*). For instance, frequently occurring enclitics in English are the possessive marker *'s* (as in *"The emperor's new clothes …"*) and contracted forms of auxiliary verbs (as in *"That's great."*, *"Where've you been?"*, or *"They'd be surprised."*). In German, enclitics are common, too, and often result in blended forms of a finite verb form with a subject pronoun (e.g., *haste* for *hast du*; English: *have*$_{\text{2nd.sg.}}$ *you*$_{\text{sg.nom.}}$), or of a preposition with a determiner form (e.g., *beim* for *bei dem*; English: *at/by the*$_{\text{masc.+neut.:sg.dat.}}$).[40] In the corpus, these had generally been transcribed as one single word, and some of these can also be found among the set of target words, with nine merged verb–pronoun forms (*haste, isser, isses, kannste, meinste, musste, siehste, weisste, willste*) and six preposition–determiner blends (*am, beim, im, vom, zum, zur*).

While categorial ambiguity has been discussed thus far as a property of word types manifested across different occurrences, these blended forms obviously instantiate two categories simultaneously, that is, in every single occurrence. Thus, in these cases, benchmark classifications could not be decided by the predominant-category criterion. One possible solution would be to introduce a new category for each type of blend; but instead, the blended forms were assigned to the larger category involved (i.e., verb–pronoun blends were treated as verbs, and preposition–determiner blends as

[40] While many preposition–determiner blends have become standard in written German, verb–pronoun blends tend to occur mainly in spoken German.

determiners). In the latter case, this also acknowledges the fact that preposition–determiner forms display case morphology inherited from determiners.

2.2.3 Descriptive statistics

The heuristic outlined in the preceding subsection mapped the 1,017 target words to the 11 benchmark categories such that each word belongs to exactly one category. The detailed composition of the resulting categories can be found in Appendix A, while a quantitative description is provided here (Table 2-1).

Table 2-1: *Composition of the 11 benchmark categories*

	Category	Target words [a] (unambiguous [b])	Underlying lexemes [c]	Tokens accounted for [a]	Average frequency [d]	Median frequency [d]
	INTJ	77 (59)	77	109,493	1,422.0	249.0
Open classes	V	288 (235)	132	212,944	739.4	244.0
	N	268 (227)	244	85,290	318.2	178.5
	ADJ	96 (38)	63	29,868	311.1	210.5
	ADV	94 (36)	94	135,435	1,440.8	304.0
Closed classes	INTG	17 (0)	12	39,168	2,304.0	266.0
	PRON	35 (21)	20	158,194	4,519.8	1,072.0
	DET	61 (10)	44	120,715	1,978.9	508.0
	PREP	15 (1)	15	37,106	2,473.7	1,152.0
	CONJ	13 (2)	13	55,385	4,260.4	1,550.0
	PTCL	53 (4)	51	133,358	2,516.2	911.0
Overall		**1,017 (633)**	**765**	**1,116,956**	**1,098.3**	**237.0**

[a] A category's number of target words and the number of tokens accounted for by these target words are closely related to, but not identical with, the category's actual type and token frequencies, respectively. The reasons for the non-identity are that low-frequency word forms are excluded and categorial ambiguity is not taken into account.

[b] Category members that do not simultaneously belong to any other category.

[c] The number of different lexemes underlying the (possibly inflected or contracted) target words of a category. Due to categorial ambiguity and polysemy, the actual values might be higher than those listed here.

[d] Computed across the individual frequencies of the category members.

The nonsyntactic category of interjections comprises 7.6% of all target word types, together accounting for 9.8% of all target word tokens in the corpus. The four open classes together comprise 73.4% of all target words, accounting for 41.5% of all target word tokens. The six closed classes only comprise 19.1% of all target words but

because of their high average frequencies together account for 48.7% of all word tokens.[41] In terms of average frequency, these six closed classes indeed range clearly above the four open classes and interjections, with adverbs and interjections being somewhere in the middle. However, in terms of median frequency, the distinction between open and closed classes becomes slightly blurred, in that interrogative words fall into the range characteristic of open classes. This reflects in part that the category of interrogative words comprises a few highly frequent items which raise average frequency but a majority of members with relatively low frequency. A second important explanation is that median frequency is not a very robust measure for describing the frequency distribution of such a small category.

The table also specifies the number of different lexemes underlying the target words of the various categories. The ratio between target words and lexemes can be taken as a very rough measure for the categories' relative tendency to inflect.[42] The most target words per underlying lexeme are found for the verb category (2.18), followed by pronouns (1.75), adjectives (1.52), interrogative words (1.42), determiners (1.39), and nouns (1.10). Of particular interest in the distributional analyses were the two largest categories verb and noun which comprise roughly the same number of target words (28.3% and 26.4% of all target words, respectively). But nouns are almost twice as many in terms of underlying lexemes whereas verb target words on average occur about twice as frequently.

As can be seen from Table 2-1, categorial ambiguity concerns the different categories at varying degrees. To study the distributional effects of this ambiguity, it is useful to have a more differentiated picture of the particular combinations of categories to which ambiguous words most frequently belong. To this end, Table 2-2 below lists for each benchmark category the percentage of its member target words that have secondary memberships in any of the other categories (according to their complete classifications, cf. p. 47).

Most prominently, 15 (88.2%) of the 17 target words that were classified as interrogative words can also be used as pronouns (in relative clauses). The high mutual ambiguity values between pronouns and determiners (37.1% and 65.6%) reflect what has been said earlier about possessives, demonstratives, and numerals (cf. 2.2.1).

---

[41] Note that the low-frequency word forms that were not selected as target words display a distribution across categories very different from the one shown here for target words. Among these low-frequency items, one mainly finds forms of nouns, adjectives, and verbs.

[42] Contraction also affects this ratio but is a marginal factor here because overall, only 6 target words are contracted forms, while 302 are inflected.

The broadest spectrum of ambiguities can be observed for the category of particles. This reflects not only the fact that this category is a very heterogeneous collection of words but also that two large classes of words are borrowed from other categories. This involves the so-called *Abtönungspartikeln* (English: *toning particles*) which originated as words of other categories — mostly adverbs and conjunctions — that took on additional communicative functions (cf. Hentschel & Weydt:280). The second class involves verbal particles that were discussed earlier. They typically are full words borrowed from potentially any other category, most frequently adverbs and prepositions.

Table 2-2: *Categorial ambiguity by benchmark category*

Benchm. category	**Ambig. (in %)** [a]	**Secondary membership in category … (in %)** [b]										
		INTJ	V	N	ADJ	ADV	INTG	PRON	DET	PREP	CONJ	PTCL
INTJ	23.4	—	6.5	10.4	5.2	3.9						6.5
V	18.4	1.0	—	12.2	2.4			2.8	0.3			
N	15.3	1.5	12.3	—	2.2			0.4	0.4			0.4
ADJ	60.4	4.2	24.0	26.0	—	7.3		2.1		1.0		12.5
ADV	61.7		5.3	3.2	18.1	—		2.1		2.1	8.5	50.0
INTG	100.0					23.5	—	88.2	35.3	5.9	11.8	11.8
PRON	40.0			2.9		8.6		—	37.1			2.9
DET	83.6		9.8	1.6	1.6	3.3		65.6	—	9.8		6.6
PREP	93.3					46.7				—	13.3	80.0
CONJ	84.6		15.4	7.7		30.8		7.7		23.1	—	38.5
PTCL	92.5	5.7	17.0	7.5	47.2	75.5		3.8	7.5	7.5	15.1	—

[a] Proportion of ambiguous members of the benchmark category (derived from Table 2-1).

[b] Proportion of benchmark category members that can also be used in particular other categories. Missing values are 0.0%.

Verbs and nouns show the lowest overall degree of ambiguity; and the majority of ambiguous cases involves the respective other category, i.e., most ambiguous verbs can also be used as nouns, and vice versa. Also remarkable is the observation that roughly one in four adjectives can also instantiate the verb category, and a similar proportion can also be used as nouns.

It should be stressed that many of these *secondary* category memberships play no role at all in how the target words are used in the given corpus of CDS; and some secondary memberships correspond to antiquated usages of the respective words that

have become very rare in general, even in adult-to-adult speech and written language. Thus, Table 2-2 represents the worst case of categorial ambiguity and is an overestimation of the ambiguity relevant in the corpus.

It is also important to note that the phenomenon of categorial ambiguity poses a challenge to both the model and the child. Their situation, however, is not entirely analogous. The child has to deal with phonological input whereas the model operates on orthographic transcriptions of this input. While this has no effect on the general problem of ambiguity as such, it does alter the degree of ambiguity for individual words. For instance, the German homophones *Meer* (English: *sea*) and *mehr* (*more*) would be perceptually identical for a child, but constitute two separate words for a model using orthographic representations. On the other hand, the German homographs *Weg* (*alley*, *path*) and *weg* (*away*) are pronounced differently ([ve:k] vs. [vɛk], respectively) and therefore distinguishable for a child but not for the model since capitalization was removed from the transcripts.[43]

2.2.4 Discussion

In terms of the representational distinctions that were discussed in 1.1.3, I here proceeded by coding the benchmark system as discrete categories, and by specifying category membership as an explicit property of word types. This coding raised several classification problems with respect to categorial ambiguity; and the solution taken here simply bypasses the problem and is by no means the ultimate answer to the challenge of representing lexical categories. But despite these shortcomings, I believe the benchmark categories and individual benchmark classifications are reasonably appropriate in the context of this dissertation. By selecting the category with which a word predominantly occurs in the corpus, the classifications emphasize the words' usage that is likely to be relevant for what the boy learns about them.

Nevertheless, the benchmark categories should not be interpreted literally as a proposal about the *real* categories underlying adult language processing. They are employed as a mere heuristic to evaluate how useful distributional regularities in the

[43] These particular examples actually appeared among the target words. While *Meer* corresponds to the target word *meer* that was unambiguously classified as a noun, *mehr* can be a particle, adjective, or the inflected form of a rare verb. In contrast, the unambiguous noun *Weg* and the word *weg* which can be either particle or adverb were joined to the same target word *weg* which thus can be a particle, adverb, or noun. Reflecting their prevalent usage in the corpus, *mehr* and *weg* were both classified as a particle in the benchmark.

corpus might be to acquire the *real* categories. The benchmark categories may differ from these real categories, but most likely not in arbitrary ways as is suggested by the substantial commonalities that exist even among competing views on categories (cf. 1.1.1), and by the various psycholinguistic traces of lexical categories (cf. 2.1.2).

It will be useful to keep the benchmark's weaknesses in mind when distributional results are discussed. For instance, Table 2-2 can assist in controlling for effects of ambiguity between pairs of categories to determine how well the unambiguous core categories could be learnt (cf. 4.1.2).

Chapter 3
Computational methods

The preceding chapter described a corpus of linguistic input and 11 benchmark categories that were assigned to a set of target words occurring in this corpus. The central question of this study is how much the highly local distributional properties of the target words could have taught the boy Leo about the lexical category of these words. In the current chapter, I formally introduce the computational tools that were used to investigate this and further questions. These comprise a co-occurrence model that extracts the intended distributional properties from the input corpus (section 3.1), a measure to compute for each pair of words how similar their distributional properties are (3.2), and several evaluation scores to quantify how well these similarities reflect the benchmark category system (3.3). These three sections correspond to steps 1, 2, and 4 in the general four-step paradigm that was described in 1.3.2. Step 3 of the paradigm is by-passed here, for the reasons laid out in 1.3.3. Along the way, I occasionally present some preliminary results of applying the computational tools to the *Leo* corpus. These are not intended to anticipate the actual analyses in chapter 4, but rather to motivate the subsequent steps of formal investigation.

As I point out in the first two sections, some settings of the co-occurrence model and the choice of a particular similarity measure are in part fine-tuned to the corpus data. Since both the model and the measure are used to explore the same data, this may seem to constitute a methodological circularity. Indeed, this fine-tuning of the method to the data would be problematic if the goal of this study were to demonstrate that one particular method is in principle superior to other accounts, or that the given corpus data provide more information than do other data.[44] But instead, the study aims at uncovering the relevant information that is contained in just the given data; and it is

[44] Studies in the field of machine learning that do compare the performance of different methods typically avoid the danger of circularity by partitioning the data into a training set and a test set such that the performance of a learning mechanism can be evaluated on different data than those on which it has learned.

justified to use methods that do this the best way. The methods may be fine-tuned to this information but, crucially, they cannot ameliorate it.

The situation is somewhat different, however, for the evaluation scores to be described in the third section. Fine-tuning these scores to the input data would not be legitimate but could result in severe artifacts. Therefore, to control for possible artifacts, these scores were tested on random baselines. Because some of these baselines are derived from the corpus, they may in themselves appear to constitute an instance of fine-tuning. But this is not at all the case, for two reasons. First, in these baselines, the information under investigation is entirely obscured from the original corpus; and second, the baselines are not used to maximize an evaluation score but instead to test whether a score reflects that, by hypothesis, the baseline does not provide any cue to lexical categories.

3.1 Co-occurrence model

So far, the notion of distributional information to be investigated has been characterized only informally (cf. 1.3.1). In this section, it is operationalized by an implemented model of lexical co-occurrence. The model first derives from the corpus a co-occurrence vector for each target word (subsection 3.1.1) which, roughly speaking, summarizes the kinds of local contexts the word occurs in. This vector is then transformed into a standardized form that is interpreted as representing the word's distributional properties (3.1.2). To provide an intuition of the structure in the resulting vector space, a two-dimensional projection of this space is shown which constitutes a first, purely visual, verification that the co-occurrence vectors do contain *some* cues to lexical category (3.1.3).

3.1.1 Computing co-occurrence vectors from the corpus

Although lexical co-occurrence statistics are a fairly simple technique — all they do is to count —, they comprise a variety of formal models that can be applied to a wide scope of research questions; and even when restricted to the question at hand, they provide several parameters which have to be set by the modeler. The particular model to be used here essentially adopts the settings of the standard analyses by Redington et al.

(1998). Therefore, I first describe the commonalities with their approach and then sketch how my model differs.

To illustrate the general procedure, assume for a moment that we are dealing with an English corpus and that the word *some* is one of the L = 1,017 target words. Suppose further that one of its occurrences in this corpus is the utterance *"look leo i got some of your crayons ."* then its two neighbors to the left (*i* and *got*) and its two neighbors to the right (*of* and *your*) would serve as the local context of this particular token of *some* (Figure 3-1).

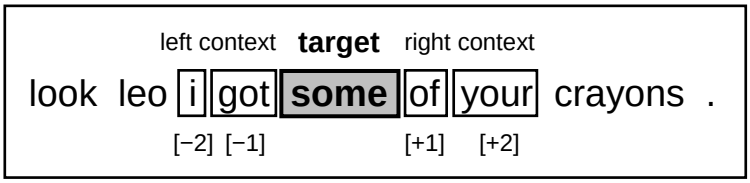

***Figure 3-1**: Context window around a target word token*

The context window consists of the two left and two right context positions, immediately adjacent to the target word.

To establish some fundamental terminology, this local context comprises precisely the words occupying the four relative *context positions*: two to the left of the target word token, [−2], one to its left, [−1], one to its right, [+1], and two to its right, [+2]. The combination of these four context positions forms a *context window*.[45] After having recorded the local context of this token of *some*, the context window slides to the next occurrence of this word to determine its local context, and so on. Having scanned all instances of *some* in the corpus in this fashion, the set of local contexts that were encountered is summarized as a long list of numbers, that is, a vector which will be called the *co-occurrence vector* of the target word *some*.

More specifically, the model starts by building four individual vectors $v^{[-2]}$, $v^{[-1]}$, $v^{[+1]}$, and $v^{[+2]}$, one for each of the four relative context positions; and their dimensions correspond to a selected set of words that *some* could co-occur with (these will be called *context words*). The purpose of the individual vector $v^{[-2]}$, for instance, is to count for each context word c_i how often *some* is preceded by c_i in the position two to its left. In

[45] Larger and smaller context windows were also considered but, in concordance with the study by Redington et al. (1998), I found the context window [−2, −1, +1, +2] to work best (for related results see 4.3.1). By contrast, Mintz and colleagues (2002) found no window size effect on the usefulness of information about the noun category, whereas for verbs, the relatively wide context window [−8, −7, …, +7, +8] generally yielded the best results.

practice, this is achieved in the following way: The model scans all occurrences of *some* in the corpus, and in turn records for each of them its local context by incrementing the values in the four vector elements that correspond to the particular context words found in the four context positions relative to *some* (Figure 3-2).

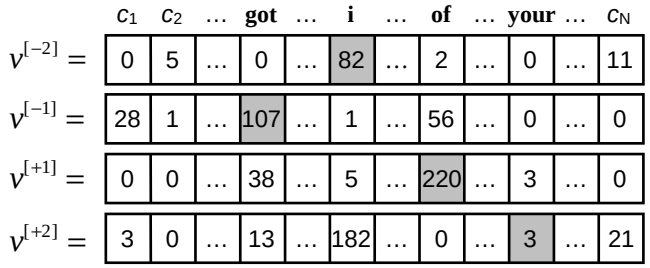

Figure 3-2: *Updating co-occurrence counts*

Hypothetical co-occurrence counts for the target word *some* before the particular instance *"look leo i got some of your crayons ."* is encountered. To record the local context of this particular token, the counts in the corresponding vector elements (shaded) are incremented by 1.

After all occurrences of *some* have been processed in this way, the four resulting vectors are concatenated to one long vector

$$v = \left(v^{[-2]}, v^{[-1]}, v^{[+1]}, v^{[+2]}\right)$$

which summarizes the co-occurrence profile of *some* across all four context positions (Figure 3-3). This long vector is the *co-occurrence vector* of the target word *some*, and it consists of $4N$ dimensions (where N denotes the number of context words), one for each context word in each of the four context positions. In the same fashion, a co-occurrence vector is also derived for each of the other target words.

Note that the method only records co-occurrences between words selected as target words (the *target lexicon*) and words selected as context words (the *context lexicon*). Co-occurrences of other word pairs are ignored. Thus, in the above example, if the word *got* happened to be excluded from the context lexicon, no co-occurrence of this token of *some* would be recorded in the context position [−1]. Potentially, the context lexicon could comprise all 30,232 word types found in the corpus. By default, however, it will be identical to the target lexicon such that we have $L = N = 1{,}017$. The other word types were excluded from the context lexicon because, due to their relatively low base frequency, co-occurrences with them would be largely influenced by random factors and provide rather unreliable cues (cf. related considerations in 2.1.3).

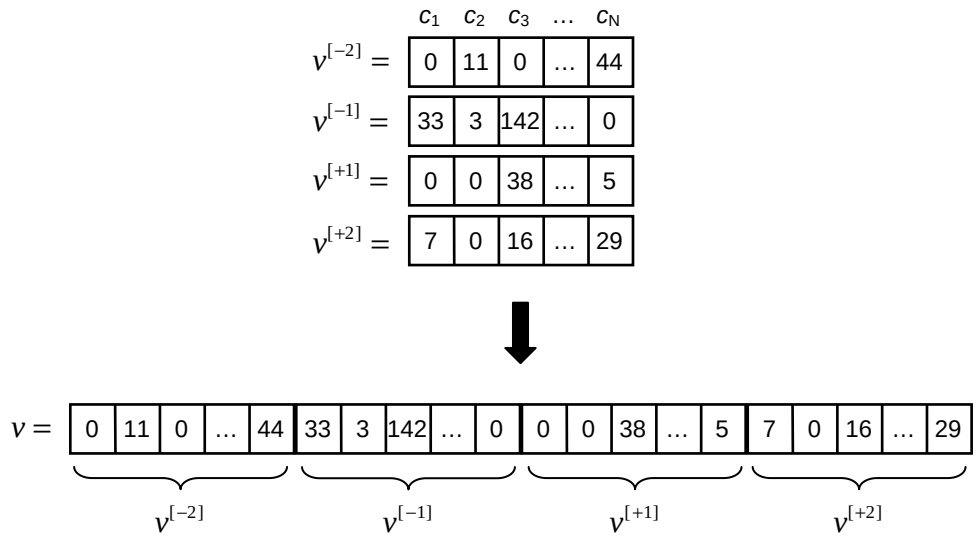

Figure 3-3: *Deriving co-occurrence vectors*

Hypothetical co-occurrence counts between a particular target word and all context words c_1, c_2, …, c_N, in the four context positions [−2], [−1], [+1], and [+2], represented by four separate vectors $v^{[-2]}$, $v^{[-1]}$, $v^{[+1]}$, and $v^{[+2]}$ (top), and concatenated to one large co-occurrence vector v (bottom).

This general method for constructing co-occurrence vectors captures the commonalities between the models used in Redington et al. and in the current study. The differences between them mainly concern the handling of utterance boundaries. In their standard analysis, Redington et al. ignore utterance boundaries entirely such that the entire input data in effect become one extremely long sequence of words. In particular, they also record co-occurrences across utterance boundaries when the context window for a target word token does not entirely lie inside an utterance. Because Redington and colleagues also operate with the context window [−2, −1, +1, +2], such cross-utterance co-occurrences can only arise for target word tokens in the first, second, last but one, or final position of an utterance.

Words in adjacent utterances presumably impose only weak, if any, categorial constraints on each other such that cross-utterance co-occurrences essentially add statistical noise to any distributional cues to lexical category.[46] Removing these co-occurrences should therefore clean up the distributional information; and in this sense, utterance boundaries can serve as an implicit cue to lexical category. Alternatively, utterance boundaries can be represented explicitly by introducing a new

[46] It should be noted here that we will later encounter evidence that co-occurrences across utterance boundaries can indeed provide some weak cues about lexical category (cf. footnote 103, p. 120).

symbol as a virtual context word and recording co-occurrences of target words with this *utterance boundary marker* (if it appears within their context window). Redington et al. explored the effects of both ways of representing utterance boundaries.[47] In my own standard analysis, I employ the explicit representation and go one step further by distinguishing between several kinds of utterance boundary symbols.[48]

More specifically, I use a set of four different utterance boundary markers which comprise one *pre-utterance marker* (_<_) to represent the beginning of utterances, and three *post-utterance markers* corresponding to questions (_*?*_), exclamations (_*!*_) and declarations (_._), as transcribed in the corpus. Thus, for instance, in the sample utterance *"look leo i got some of your crayons ."*, the word *leo* co-occurs with the pre-utterance marker _<_ in the context position [−2] and with *look* in the position [−1]. The word *look*, in contrast, co-occurs with this marker in the context position [−1] and with no item at all in context position [−2]. Similarly, the word *your* co-occurs with the post-utterance marker _._ in the context position [+2] while *crayons* co-occurs with this marker in the position [+1] and with no item at all in context position [+2]. Thus, in addition to genuine lexical co-occurrence relations between words, the model extracts information about the target words' tendency to occur in a particular serial position within an utterance (first, second, last but one, or last), and — for the latter two positions — more differential information about their tendency to occupy these positions depending on the kind of utterance termination.[49] With the inclusion of the single pre-utterance marker and the three post-utterance markers, the individual vectors $v^{[-2]}$ and $v^{[-1]}$ both arrive at $N + 1 = 1{,}018$ dimensions while $v^{[+1]}$ and $v^{[+2]}$ have $N + 3 = 1{,}020$ dimensions each. In consequence, the concatenated co-occurrence vectors v all have a total of n=4,076 dimensions.

To further extend terminology, these dimensions will occasionally be referred to as *context dimensions*. By writing the co-occurrence vectors of all target words below each other, one obtains a large *co-occurrence matrix* with 1,017 rows (one for each target word) and 4,076 columns (one for each context dimension). Each of the 4,145,292

[47] Mintz et al. (2002) excluded cross-utterance co-occurrences from all their analyses and did not experiment with explicit cues from utterance boundaries.

[48] The usefulness of the cues arising from this extension was assessed and compared with the corresponding explorations by Redington et al. (cf. 4.2.3).

[49] This modification therefore corresponds to the assumption that before acquiring lexical categories, children are able to distinguish between these three different kinds of utterance endings, on the basis of salient prosodic and intonational cues. The transcribers of the *Leo* corpus identified and classified utterances by a similar heuristic (the resulting distribution of utterance final punctuation in the corpus was given in subsection 2.1.2).

individual matrix cells thus indicates how often a particular target word was found to co-occur with a particular context word (or utterance boundary marker) in a particular context position.[50]

Table 3-1: *Distribution of co-occurrence values across matrix cells*

Co-occurrence value [a]	**Proportion of matrix cells** (in %) [b]	**Co-occurrences accounted for** (in %) [c]
0	90.209	0.0
1	4.876	5.6
2	1.562	3.6
3	0.773	2.7
4	0.471	2.2
5	0.330	1.9
6 – 10	0.747	6.5
11 – 20	0.461	7.7
21 – 50	0.330	11.8
51 – 100	0.124	10.0
101 – 500	0.099	22.4
501 – 1,000	0.011	8.4
1,001 – 5,000	0.006	11.9
5,001 – 18,942	0.001	5.5

[a] Greater values are pooled into intervals.

[b] The percentage of matrix cells that carry the given co-occurrence value (or a value in the given range).

[c] The sum of all co-occurrences in the respective matrix cells, as a percentage of all co-occurrences recorded.

Before proceeding with the second step of the model, it is useful to get a rough impression of the co-occurrence values obtained for the *Leo* corpus. In total, the model recorded 3,620,473 individual co-occurrences from the corpus. This total number, however, is far from being evenly distributed across the matrix. The lion's share is concentrated in a small portion of matrix cells; e.g., 1.0% of all cells have values greater than 10 and together account for as many as 77.7% of all recorded co-occurrences (Table 3-1). The single highest co-occurrence count by itself contributes 18,942 (or 0.5%) of all co-occurrences. By contrast, the bulk of matrix cells contain very small values; in fact, a large proportion (90.2%) of them is 0.[51]

[50] Note that this matrix is not symmetric although it contains a lot of redundancy.

[51] The large number of zero entries is not surprising. Recall that half of the target words occur 237 or fewer times in the corpus (cf. 2.1.3). Thus, in a particular context position, they cannot co-occur with

These brief descriptive statistics are not part of the model but rather provide a first intuition of the information that it extracts from this particular input corpus. The observation that the recorded co-occurrences are not evenly distributed across all matrix cells is important; for it is precisely this skewed distribution that can provide cues to lexical category. A perfectly homogeneous matrix, by contrast, in which all cells carry the exact same value, would not provide any information.

3.1.2 Standardizing co-occurrence vectors

A target word's co-occurrence vector *v* summarizes the local contexts that it is found to occur in. If the tendency to occur in these contexts relates at all to any intrinsic property of the word itself (as is hypothesized here to be the case), one would expect this tendency to be fairly robust across the corpus. That is, suppose one randomly selects half of all instances of a particular target word in the corpus and derives a second co-occurrence vector *v'* for it that summarizes the local contexts of just these tokens. The prediction would then be that each co-occurrence count in *v'* is roughly half of the corresponding value in *v*. But if this is the case, one would wish to express formally that both vectors, although they are clearly not identical, represent essentially the same distributional properties.

These considerations portray the issue in an oversimplified manner — as, for instance, the co-occurrence effects of so-called *rare events* are not taken into account — but they point to the important fact that a word's *base frequency* (i.e., frequency of occurrence) crucially influences its co-occurrence vector. To filter out this immediate influence and extract the *pure* distributional properties, each co-occurrence vector

$$v = (v_1, v_2, \ldots, v_n)$$

is rescaled to a vector $\overline{v}$ defined by equation (1).

$$\overline{v} = \frac{1}{\sum_{i=1}^{n} v_i} \cdot v = \left(\frac{v_1}{\sum_{i=1}^{n} v_i}, \frac{v_2}{\sum_{i=1}^{n} v_i}, \ldots, \frac{v_n}{\sum_{i=1}^{n} v_i} \right) \quad (1)$$

The resulting *standardized co-occurrence vector* (henceforth *SCO vector*) $\overline{v}$ has *unit mass* (i.e., its elements $\overline{v}_i$ are all nonnegative and add up to 1) and therefore formally

more than 237 different context words. In subsection 4.2.3, one way of reducing the proportion of zero co-occurrences will be evaluated for its distributional consequences.

constitutes a probability distribution. One can think of its dimensions as representing *relative co-occurrence frequencies* rather than absolute co-occurrence counts.[52] This rescaling allows for comparing the distributional properties of different target words irrespective of their frequency of occurrence.

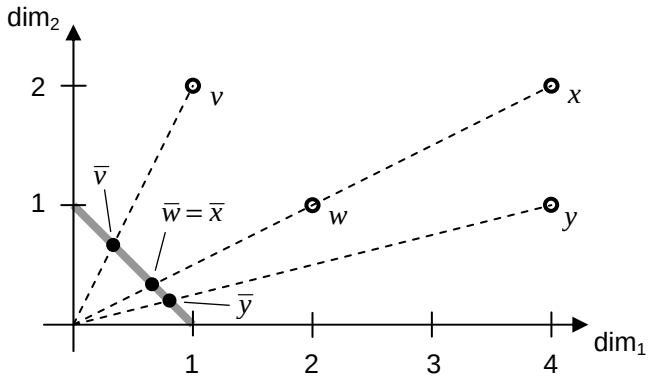

Figure 3-4: *Standardizing co-occurrence vectors*

Two-dimensional illustration of co-occurrence vectors *v*, *w*, *x*, *y* and their corresponding SCO vectors $\overline{v}, \overline{w}, \overline{x}, \overline{y}$. Geometrically, $\overline{v}$ is derived from *v* as the intersection between the dashed line (connecting *v* with the origin) and the hyper plane of points with unit mass (gray diagonal). The different co-occurrence vectors *w* and *x* yield the same SCO vector $\overline{w} = \overline{x}$.

The way how a word's distributional properties and base frequency interact in its co-occurrence vector can be illustrated geometrically as in Figure 3-4, for the case of two (rather than 4,076) dimensions. In this simple visualization, the gray diagonal line represents the set of possible SCO vectors. Rescaling a co-occurrence vector *v* to its corresponding SCO vector $\overline{v}$ is equivalent to projecting it onto this gray diagonal, along the radial line (dashed) connecting the origin of the coordinate system with *v*. The length of this radial line (i.e., the distance of *v* from the origin) is roughly proportional to the word's base frequency; the resulting SCO vector represents its distributional properties. Thus, different co-occurrence vectors that lie on the same radial line (like *w* and *x* in the figure), project onto the same SCO vector and therefore have the same

[52] Note, however, that for two reasons the entries of the obtained SCO vectors are not exactly identical with the true relative co-occurrence frequencies. First, the relative co-occurrences with selected context words and utterance boundaries get overestimated because co-occurrences with nonselected words are ignored. And second, relative co-occurrence frequencies would have to be computed separately for each of the four context positions (as in 4.3.2). Standardizing the long concatenated vectors thus systematically underestimates relative co-occurrence frequencies by factor 4.

distributional properties. Vectors on different dashed lines (like *v*, *w*, and *y*) project onto different SCO vectors and thus have different distributional properties.

Let me sum up the crucial facts about the SCO vectors which have been derived from the corpus. First, they have 4,076 context dimensions, each of which is reserved for a particular context word in a particular context position. Second, a target word's SCO vector represents the distributional properties of that target word; and I therefore treat these two concepts as identical. Third, the co-occurrence approach only exploits overt word ordering but it receives no explicit information about the underlying syntactic structure. Finally, the co-occurrence model does not have access to the benchmark category system that was introduced earlier.

3.1.3 Visual inspection

The central hypothesis motivating this study (cf. 1.3.1) predicts that words of the same category have similar SCO vectors, and, conversely, that vectors with the similar SCO vectors tend to belong to the same category. This claim is to be tested against the null hypothesis that the similarities between SCO vectors are completely random and provide no cues to lexical categories. To make sense of both the hypothesis and the null hypothesis, it is necessary to first specify what it means for SCO vectors to be similar. A formal definition will be given in the next section; for the purposes of the current subsection, an intuitive notion will do. It simply interprets vectors geometrically, as points in space (as was already done in Figure 3-4) and defines their degree of similarity by their proximity in space; viz., the closer two vectors are, the more similar they are. With this intuitive notion, the basic hypothesis translates to the prediction that words of the same category occupy the same regions in space whereas words of different categories are located in different regions.

As a first informal test of this prediction, it is worth to subjectively take a glance at the SCO vector space. However, because it is impossible to plot vectors with several thousand dimensions, I applied *Principal Components Analysis* (PCA) to determine the most relevant *principal components* (i.e., rotated coordinate axes) in the SCO vector space.[53] Figure 3-5 plots all 1,017 SCO vectors relative to the second and the third principal component which together explain 16.4% of the overall variation (and, thus,

[53] In a nutshell, PCA rotates the entire vector space to identify the coordinate axes along which the SCO vectors show the greatest variation. Due to the rotation, these axes are typically not identical with any of the original context dimensions; instead they are weighted combinations of these dimensions.

potential information) in the set of SCO vectors. The first principal component — potentially the most informative one — is not displayed here because it mainly discriminates interjections from all other categories but does not show much structure within these other categories. For the purpose of verification, vectors are color-marked for their benchmark category, though only for five of the 11 categories, to avoid overcrowding the graph.

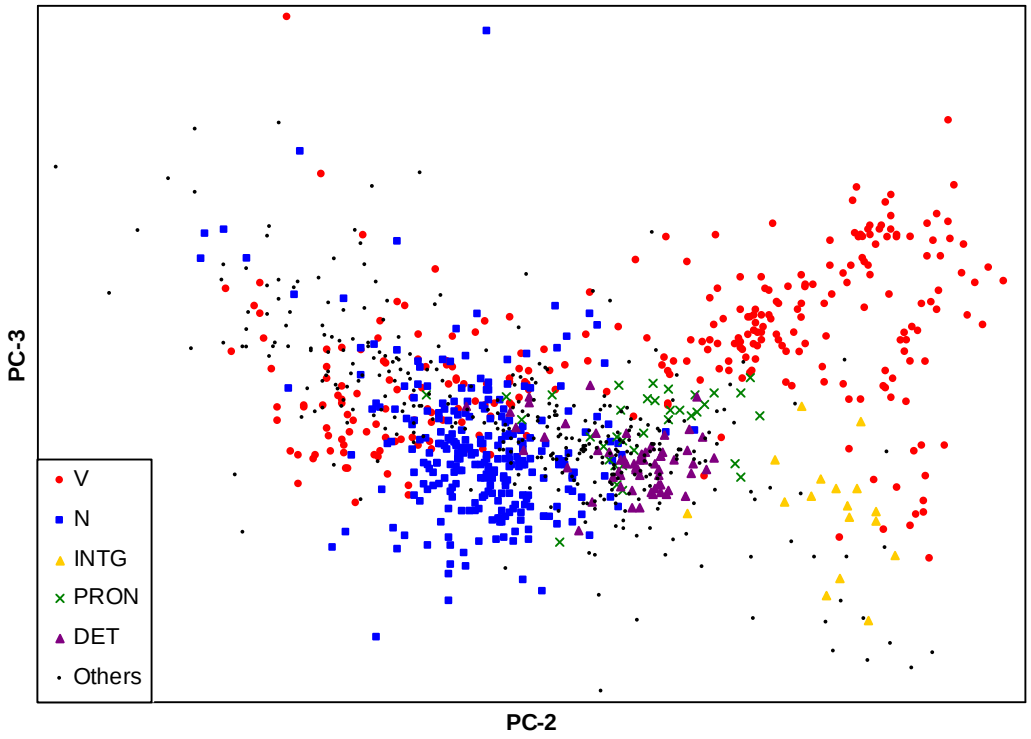

***Figure 3-5**: Two-dimensional projection of SCO vector space*

The coordinates PC-2 and PC-3 are the second and third principal component of the full SCO vector space. For five selected benchmark categories, their corresponding vectors are color-marked.

With regard to the prediction, the plot clearly does reflect some category structure. Vectors tend to form clusters that roughly correspond to the benchmark categories. However, these clusters are neither compact nor well-separated from each other; instead some of them overlap substantially. Taken together, this visual inspection confirms the prediction only in part. However, the graph merely captures two out of 4,076 dimensions and only 16.4% of the full information immanent in the SCO vectors. Thus, in the real high-dimensional space the different categories might form much more coherent clusters in more distinct regions of that space. The formal methods to be

described in the following two sections will therefore serve to assess this full space and evaluate how well the SCO vectors of one category cluster together.

3.2 Measures of similarity between vectors

In this section, the intuitive notion of similarity between SCO vectors is formalized. There are a large number of mathematical measures that can serve this purpose. Each of these *similarity measures* assigns to any two vectors a value which quantifies how similar they are.[54] I considered and tested five such measures — borrowed from geometry, statistics, and information theory — which are described below (3.2.1). I then present an empirical test of these measures to determine which of them are most sensitive to the relevant structure among SCO vectors (3.2.2).

In line with Redington et al. (1998) and Mintz et al. (2002), all similarity measures were applied directly to the concatenated SCO vectors. In pilot work, I also explored the alternative of treating each context position separately (as done by Brill, 1993, and Schütze, 1995). To this end, SCO vectors were rescaled to unit mass in each context position, and similarity values between any two vectors were first computed independently for each position and then averaged to a single global value. However, empirical results favored computing similarity values directly from the concatenated vectors.

3.2.1 Candidate measures

Probably the most intuitive way to formally measure similarity between vectors is to pick up the earlier space metaphor and link similarity to geometric distance. One such measure was included in the tests, namely, the L_1 metric (*Manhattan distance*). Its formal definition (and that of the four other measures) is given in Table 3-2 (p. 69). When applied to SCO vectors, L_1 distances range from 0 (for identical vectors) to 2 (for

[54] Note that some of these measures in fact quantify *dissimilarity*, in the sense that they produce greater values for less similar vectors. Others are genuine measures of similarity; i.e., they produce greater values for more similar vectors. But each of the former group of measures can be transformed into one of the latter, and vice versa. Therefore, and to keep terminology simple, I refer to both groups as *similarity measures*.

orthogonal vectors).[55] As a second geometric measure of similarity between any two vectors, I chose the *cosine* of the angle between these vectors. When applied to SCO vectors, cosine values range from 0.0 (for orthogonal vectors) to 1.0 (for identical vectors).

If SCO vectors are interpreted as the outcomes of different random variables on the same data set — with the dimensions corresponding to cases — their similarity can be measured in terms of the statistical correlation between any two such random variables. Two such measures were considered, viz., the *linear correlation coefficient* and *Spearman's rank correlation coefficient*. Both measures produce values between −1.0 (for maximally different SCO vectors) and +1.0 (for identical SCO vectors in the case of linear correlation; and for vectors with identical rank orders in the case of rank correlation).[56]

The fifth measure was derived from *relative entropy* (also referred to as *Kullback-Leibler divergence*) which is a standard measure in information theory. By applying it to vectors, one interprets these vectors as probability distributions over the same event space, with the dimensions corresponding to alternative events (recall that SCO vectors have unit mass). However, relative entropy itself has two properties that disqualify it as a candidate measure of similarity between vectors. First, it is nonsymmetric (the relative entropy between *v* and *w* might not be the same as that between *w* and *v*).[57] And second, it is not defined for vectors containing any zero entries.[58] Especially this latter property is problematic since roughly 90% of all values across SCO vectors *are* in fact 0 (cf. Table 3-1, p. 61). There are several ways of deriving from relative entropy a measure which is both symmetric and well-defined on any pair of SCO vectors. I chose one such proposal, the *Jensen-Shannon divergence* (cf. Dagan, Lee, & Pereira, 1999)

---

[55] Note that SCO vectors have unit length with respect to L_1; that is, their L_1 distance from the origin is 1.

[56] In practice — because negative similarity values may be counterintuitive — this range of values [−1;1] was mapped onto the interval [0;1], by virtue of the linear transformation $x \mapsto (1-x)/2$. But this step can be neglected as it does not alter the notions of similarity to which these measures are sensitive.

[57] While an intuitive notion of similarity is symmetric, Lee (1999) suggests that in the given context, operating with nonsymmetric similarity measures might in fact provide an advantage. That is, to the extent that lexical categories are definable as substitution classes (as done in distributional analysis, cf. p. 6), a target word *t* might be a better substitute for target word *s* (across sentential contexts) than vice versa. Lee's proposal is not picked up in the current study because using just *any* asymmetric measure would not be advisable; preferably, the particular type of asymmetry that it implements would be linguistically meaningful.

[58] A solution often taken in such cases — and also in the context of category acquisition (e.g., Brill, 1993) — is a technique called *smoothing* which slightly raises probability estimates for rare (and particularly unseen) events such that each zero frequency is replaced by a small positive number. However, I decided not to apply any smoothing because, in the given corpus, zero frequencies may be meaningful cues to lexical category, and furthermore, the child has to deal with such zero frequencies (i.e., possible co-occurrences that he does not observe in his input) as well.

and extended it by incorporating an idea underlying Lee's (1999) *skew divergence*.[59] The resulting measure will be referred to as *generalized Jensen-Shannon divergence*; its values range from 0 (for identical vectors) to infinity.

The generalized Jensen-Shannon divergence is probably applied here for the first time; but in general, measures derived from relative entropy are fairly common in the field (e.g., Dagan, Lee, & Pereira, 1999; Lee, 1999; Brill, 1993). The four other measures all have been previously used in related work. The cosine appears to be the default similarity measure for various kinds of linguistically motivated vectors (e.g., Zavrel, 1996; Schütze, 1995, 1998; Landauer & Dumais, 1997). Mintz et al. (2002) used a nonlinear transformation of the cosine, viz., the angle between vectors. By contrast, Redington et al. (1998) worked with the rank correlation measure.[60]

All five measures implement related notions of similarity between vectors but emphasize different aspects of similarity. For instance, relative to the other measures, cosine and linear correlation focus on the vectors' similarity in dimensions with greater values (thus, on the relatively high co-occurrence frequencies). By contrast, L_1 distance and the generalized Jensen-Shannon divergence are also quite sensitive to similarity in dimensions with medium and low values, with the latter being particularly sensitive to dimensions with values close to 0. The rank correlation coefficient is special in several ways. By switching from relative frequencies to rank order, it abandons a lot of information contained in the SCO vectors and is therefore, in theory, the least sensitive of all five measures considered. However, this property need not be a disadvantage but

---

[59] Elsewhere, Dagan, Lee, and Pereira (1997) refer to the Jensen-Shannon divergence as *total divergence to the average* while Manning and Schütze (1999:304) call it *information radius*.

[60] Another very popular geometric measure is the L_2 metric (*Euclidean distance*). I did not include it here because it is closely related to the cosine and the linear correlation coefficient. If SCO vectors are rescaled to vectors $\tilde{v}, \tilde{w}$ with unit length (in terms of the Euclidean norm, cf. Table 3-2), L_2 is a strictly monotonic function of the cosine

$$L_2(\tilde{v}, \tilde{w}) = \sqrt{2\left(1 - \cos(\tilde{v}, \tilde{w})\right)} \ .$$

Thus, on these rescaled SCO vectors, Euclidean distance and cosine are sensitive to precisely the same similarity structure, and it would be no gain to consider them both. On the other hand, if each SCO vector is first translated by the mean of its vector elements (such that the sum of vector elements becomes 0) before the resulting vector is rescaled to unit length like above, one obtains for any two such normalized vectors $\breve{v}, \breve{w}$ the equation

$$L_2(\breve{v}, \breve{w}) = \sqrt{2\left(1 - \rho_{\text{linear}}(\breve{v}, \breve{w})\right)} \ .$$

On these transformed SCO vectors, Euclidean distance and linear correlation are therefore equivalent. It should be noted that, although the relations of Euclidean distance with cosine and linear correlation appear formally analogous, the direct relation between cosine and linear correlation is rather complex, due to their different standardization functions. In theory, both measures are sensitive to very different notions of similarity; and it is only due to the statistical properties of realistic SCO vectors that the performance of both measures turns out to be consistently the same (cf. the next subsection).

Table 3-2: *Similarity measures under consideration*

Measure	Computation for SCO vectors [a]
Manhattan distance	$L_1(v,w) = \sum_{i=1}^{n} \lvert v_i - w_i \rvert$
Cosine [b]	$\cos(v,w) = \cos(\tilde{v},\tilde{w}) = \sum_{i=1}^{n} \tilde{v}_i \tilde{w}_i$ with each SCO vector v being rescaled to $\tilde{v} = \frac{1}{\lVert v \rVert_2} v$ where $\lVert v \rVert_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$ is the Euclidean norm.
Linear correlation coefficient [b]	$\rho_{\text{linear}}(v,w) = \rho_{\text{linear}}\left(Z^{(v)}, Z^{(w)}\right) = \frac{1}{n}\sum_{i=1}^{n} Z_i^{(v)} Z_i^{(w)}$ with each SCO vector v being transformed to *Z scores* $Z_i^{(v)} = \frac{v_i - \mu_v}{\sigma_v}$ where $\mu_v = \frac{1}{n}\sum_{i=1}^{n} v_i$ is the mean of its vector elements and $\sigma_v = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \mu_v)^2}$ their standard deviation.
Spearman's rank correlation coefficient [b]	$\rho_{\text{rank}}(v,w) = \rho_{\text{rank}}\left(R_i^{(v)}, R_i^{(w)}\right) = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n}\left(R_i^{(v)} - R_i^{(w)}\right)^2$ with each SCO vector v being transformed to the rank order $R^{(v)} = \left(R_1^{(v)}, R_2^{(v)}, \ldots, R_n^{(v)}\right)$ of its vector elements; e.g., $R_i^{(v)}$ is 2 if v_i is the second greatest of all n vector elements; when multiple vector elements are identical, their rank is defined as the mean of ranks they occupy; thus the vector $v = (.5, .2, .1, .1, .1, 0, 0)$ yields the rank order $R^{(v)} = (1, 2, 4, 4, 4, 6.5, 6.5)$.
Generalized Jensen-Shannon divergence [c]	$J_\alpha(v,w) = \frac{1}{2}\left[D\left(v \,\|\, \alpha w + (1-\alpha)v\right) + D\left(w \,\|\, \alpha v + (1-\alpha)w\right)\right]$ where $D(v \,\|\, w) = \sum_{i=1}^{n} v_i \log_2 \frac{v_i}{w_i}$ is the relative entropy, and with $0 < \alpha < 1$.

[a] Formulated for SCO vectors $v = (v_1, v_2, \ldots, v_n)$ and $w = (w_1, w_2, \ldots, w_n)$.

[b] This measure is unaffected by any rescaling of vectors — it can therefore be computed either from SCO vectors or directly from the raw co-occurrence vectors.

[c] By default, the parameter α was set to .99; but other values were tested as well, without substantial effects. The value α=.5 recovers the original Jensen-Shannon divergence.

might in fact serve to yield more robust results. Furthermore, as Finch and colleagues (Finch, 1993:94; Redington et al., 1998:437) point out, rank correlation is attractive in that its computational power does not rely on any assumptions about the statistical properties of the language data. I shall return to these issues below.

3.2.2 Testing the candidate measures

Because of these differences between measures, it was not clear a priori which of them is most sensitive to just those aspects of similarity that may provide useful information about lexical categories. In pilot work, I therefore identified the measure that yields the best results on the given input corpus. To this end, I applied the full evaluation paradigm (co-occurrence model, similarity measure, and evaluation score), varying the settings of the co-occurrence model and also using several different evaluation scores, in order to isolate the influence of the similarity measure.

Leaving aside rank correlation for a moment, the pattern of performance was essentially always the same across this variation. The most sensitive to the relevant aspects of similarity are L_1 distance and generalized Jensen-Shannon divergence, with a slight but consistent advantage for L_1. Substantially less sensitive is a second group consisting of cosine and the linear correlation coefficient with no advantage for either measure.

The rank correlation measure does not fit nicely into this stationary pattern. Whether or not it performs better than another measure turns out to depend largely on the frequency distribution of the context lexicon. This dependence is in turn found across a variety of model settings and evaluation scores — a representative example is given in Figure 3-6. Recall that, by default, context words were selected as those word types that occurred at least 100 times in the corpus. For this default lexicon, rank correlation shows by far the lowest sensitivity of all five measures.[61] However, when the frequency threshold is gradually raised — resulting in smaller context lexica — the performance of rank correlation catches up with the other measures. In fact, as the context lexicon is reduced to the 50 (or fewer) most frequent words, rank correlation performs even better than any of them.

This is a very interesting pattern because, presumably, the information provided by co-occurrences with context words is reduced as the number of context words

[61] The absolute sensitivity values are not important here, only differences between measures and relative changes across the different context lexica.

decreases.[62] Correspondingly, the sensitivity values of the other measures gradually declines or remains roughly unchanged as the context lexicon is reduced from 1,017 to 50 words. Only when the frequency threshold is raised still further such that the context lexicon approaches the 10 most frequent words, the extracted information gets reduced to an extent that performance drops substantially for all measures, including rank correlation, while their order of performance remains the same as for the 50 most frequent words.

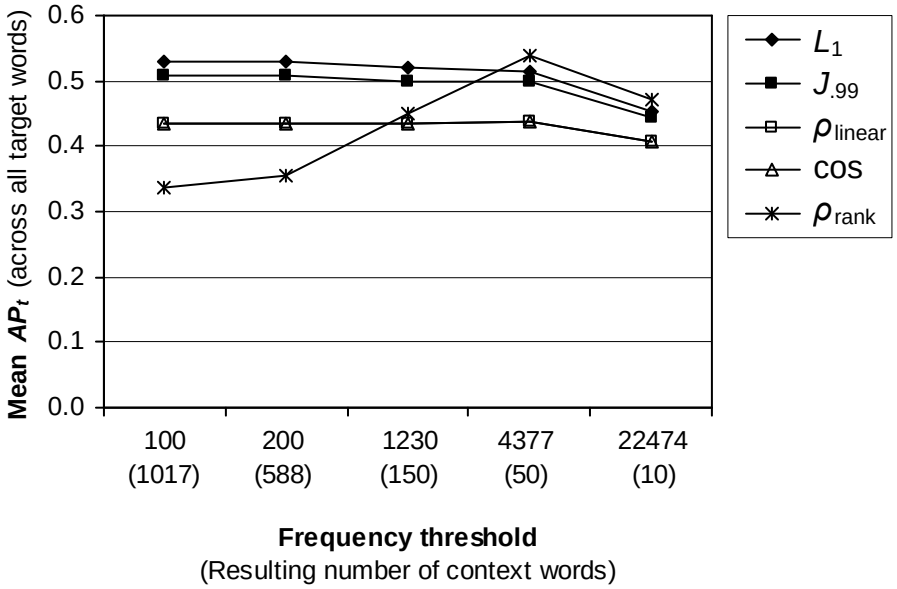

***Figure 3-6**: Sensitivity of similarity measures*

The threshold for base frequency by which words are selected for the context lexicon is gradually raised from 100 to 22,474. All other settings of the co-occurrence method are held constant and as described in section 3.1. The evaluation score is derived from *Average Precision* (cf. section 3.3). Linear interpolations between data points are only meant to highlight rough tendencies and need not be good approximations of the measures' sensitivities on intermediate context lexica.

This peculiar performance pattern of the rank correlation measure can be explained directly from its formal properties. By transforming SCO vectors to rank orders, it forces the distribution of continuous vector elements onto a discrete sequence of consecutive integers (i.e., the ranks), only compromising this strict ranking when

[62] Of course, whether smaller context lexica really provide less information about lexical categories is an empirical question which I address in 4.2.3. The important observation made here is that rank correlation is the only measure suggesting that smaller context lexica can provide substantially more (or better) information than larger ones.

exactly identical values are encountered. Thus, where the distribution of vector elements of a given SCO vector is gappy, this transformation compresses the gaps onto single rank increments; on the other hand, when there are many vector elements with roughly the same value, small differences are inflated. Given the skewed distribution of SCO vector elements that can be inferred from Table 3-1 (p. 61), the compression of gaps mainly applies to the vector elements with the highest values while the inflation of small differences concerns the vector elements with values around 0.

For instance, consider the three hypothetical SCO vectors *u*, *v*, and *w* that are defined below.

$$\left.\begin{array}{l} u = (.50, .45, .03, .01, .01, 0, 0, 0) \\ v = (.85, .10, .03, .01, .01, 0, 0, 0) \end{array}\right\} \quad R^{(u)} = R^{(v)} = (1, 2, 3, 4.5, 4.5, 7, 7, 7)$$

$$w = (.85, .10, 0, 0, 0, .03, .01, .01) \quad \rightarrow \quad R^{(w)} = (1, 2, 7, 7, 7, 3, 4.5, 4.5)$$

Although *u* and *v* display extreme absolute differences in the first two dimensions, they map onto the same rank order and are consequently judged as identical by rank correlation. By contrast, *v* and *w* are identical in the first two dimensions (which together account for 95% of the probability mass) and display only marginal differences in the other six dimensions (which together account for only 5% of the probability mass in either vector) but yield fairly different rank orders and are therefore treated as not very similar by the measure.

The compression of gaps may mask some possibly important information present in the SCO vectors (e.g., the huge difference between vectors *u* and *v* in the example); but it also renders the rank correlation measure potentially more robust with respect to statistical noise in the high-frequency range. Inflating small differences among the many vector elements around 0, however, makes the measure highly susceptible to statistical noise in the low-frequency range — the very range where noise plays the biggest role anyway.

It follows from these considerations that the rank correlation measure performs poorly when the proportion of SCO vector elements around 0 is very high; and it performs better when this proportion is reduced. This is precisely the effect that we observed in Figure 3-6 because raising the frequency threshold for the context lexicon necessarily reduces the relevant proportion for all SCO vectors. For instance, when moving from the 1,017 to the 50 most frequent word types, the overall proportion of vector elements with the value 0 drops from 90.2% to 50.0% while the proportion of

elements with values greater than 10 is boosted from 1.0% to 12.1%.[63] The fact that rank correlation outperforms all other measures at the 50-word level indicates that at this point, its robustness in the high-frequency range outweighs its weakness in the low-frequency range.

These insights might account for the fact that across different studies on lexical categorization, both L_1 distance and rank correlation have been found to outperform each other (e.g., Finch, 1993; Hughes & Atwell, 1994). Finch interprets his differential findings to imply that L_1 works better for artificial languages while rank correlation is more suitable for naturalistic data. My own results suggest a more general conclusion. The crucial factor deciding about the superior performance of either L_1 distance or rank correlation is the overall proportion of co-occurrence values around 0. And this proportion may be influenced both by the choice of the data sample (e.g., artificial vs. natural languages; spoken vs. written language) and by the method (e.g., raising the frequency distribution of the context lexicon or of the target lexicon). In consequence, there is no universal answer and each study has to assess which of the two (or possibly other) measures is most suitable for the data and method at hand.

For the purposes of this dissertation, I decided to use L_1 distance as the default measure. Rank correlation may yield the single highest sensitivity value for any context lexicon (cf. Figure 3-6 above); but L_1 is more consistent across all context lexica. This property is important since one major goal of this study is to explore the character of the distributional regularities in the input, and being restricted to co-occurrences with the 50 most frequent word types might obstruct this goal.

One intuitive property of L_1 that sets it apart from the other measures concerns the fact that it computes distance by driving along the perpendicular grid of the coordinate system, like a cab in Manhattan (hence the name *Manhattan distance*; cf. Figure 3-7 below). Thus, L_1 emphasizes the partial independence of the individual dimensions — i.e., of the observable context words in particular context dimensions — weighting each by its direct distance, rather than overemphasizing the short or the long distances. This desirable property might contribute to the consistently high performance of L_1.

[63] Note that I switch here between the levels of SCO vectors and their underlying co-occurrence vectors, of which only the latter contain absolute frequencies as their elements. However, this switch is justified here as it makes no difference whether rank correlation is computed from SCO vectors or directly from their underlying co-occurrence vectors (cf. Table 3-2 on p. 69):

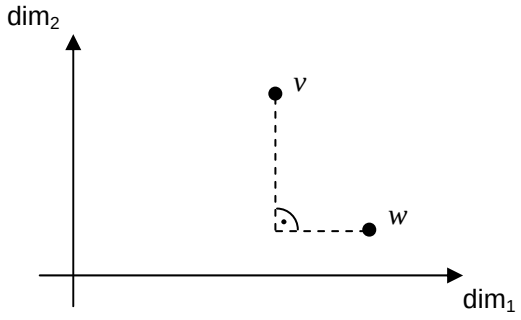

***Figure 3-7**: Computing the L_1 distance between vectors*

Two-dimensional illustration of two vectors *v* and *w* and their L_1 distance (total length of the dashed path).

In addition to its geometric visualization, L_1 has a statistical interpretation as well. For probability distributions *v* and *w*, one can easily prove the following equation

$$\frac{1}{2} L_1(v,w) = 1 - M(v,w) \qquad \text{where} \quad M(v,w) = \sum_{i=1}^{n} \min(v_i, w_i) \ . \tag{2}$$

Applied to SCO vectors, the sum $M(v,w)$ specifies the expected degree to which the target words corresponding to *v* and *w* co-occur with the same context words in the same context positions, given some random sample of their tokens (cf. Manning & Schütze, 1999:305).

3.3 Evaluation scores

This section introduces the evaluation scores that were used for assessing formally to which extent the similarity structure among the SCO vectors correlates with the benchmark category system. The task of selecting a particular evaluation score is a tricky one because candidate scores may produce artifacts that make the correlation between similarity structure and benchmark categories appear better than it is. But what counts as an artifact essentially depends on one's null hypothesis against which the actual information is to be compared. Beyond the possibility of artifacts, the applicability of a particular evaluation score is constrained by its computational properties. For the scores to be used here, I therefore controlled for artifacts and report

the relevant computational properties that have to be kept in mind when interpreting and comparing any results in terms of these scores.

In the given situation, the general challenge that any evaluation score has to meet is that the SCO vector space accommodates a notion of continuous and implicit categories whereas the benchmark categories are discrete and explicitly coded (cf. 1.1.3 and 2.2.4). The standard approach to overcome this formal discrepancy between the two structures has been to represent the similarity structure between SCO vectors as a clustering tree and to derive discrete and explicit classes of words by cutting this tree at a particular similarity level (cf. 1.3.2). These can be interpreted as hypotheses about the benchmark categories, and the task becomes that of comparing two structures of the same type. However, for the reasons laid out in 1.3.3, I chose to avoid the detour via clustering trees, and instead to evaluate the similarity structure more directly. To this end, I considered and tested three alternative evaluation paradigms and decided on one of them that transforms the vector space into a set of *rank lists* of target words. Within this general paradigm, several particular scores can be defined. I tested several of them and finally decided on a set of three complementary scores.

To illustrate how the general paradigm and the three scores evaluate the continuous SCO vectors against the discrete benchmark categories, it will be useful to distinguish four general types of constellations in the SCO vector space (subsection 3.3.1). One of these types concerns the case that the SCO vectors contain no information at all about a particular category; and some important considerations about how this case can be properly defined are given in subsection 3.3.2. I then formally introduce the rank list evaluation paradigm and the three selected scores and describe in terms of the four types of vector constellations how these scores were employed in the actual analyses (3.3.3 and 3.3.4). The final subsection briefly discusses several alternative scores from all three evaluation paradigms which were also considered.

3.3.1 Category scenarios to be distinguished

From the perspective of a particular benchmark category Γ, the similarity structure in the SCO vector space is maximally informative about this category if all its members (more precisely, the SCO vectors of its members) clump together in a single compact cluster such that they are closer to each other than to any other words (the *nonmembers*). Such a constellation will be referred to as *Clump Scenario* (Figure 3-8a below). At the opposite extreme, the similarities between SCO vectors are not informative at all about

Γ; there is no pattern whatsoever in the distribution of members and nonmembers that would reflect the substantiality of Γ as a category. Such a constellation depicts the null hypothesis which is marked by complete randomness and will therefore be called *Random Scenario* (Figure 3-8c).

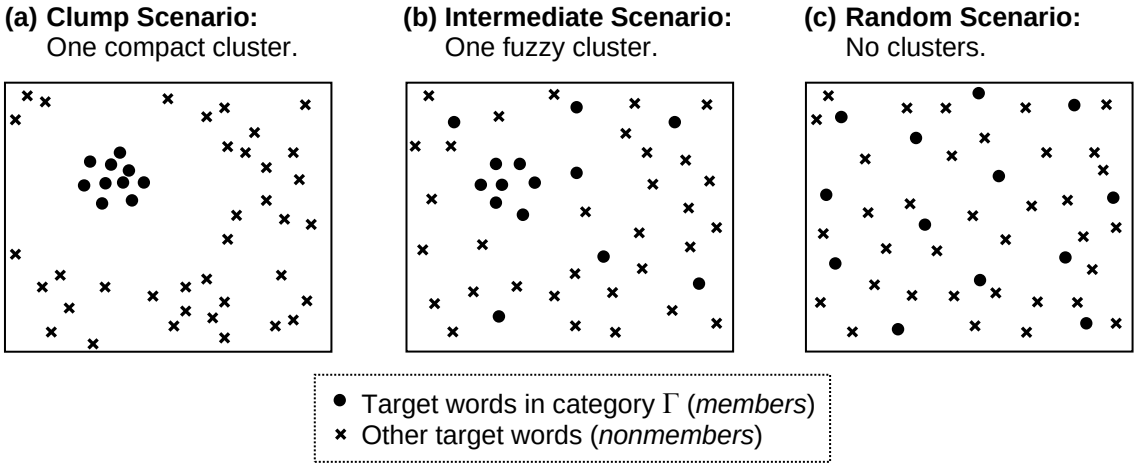

***Figure 3-8**: Category scenarios of decreasing informativeness*

Two-dimensional illustrations of possible constellations in which the SCO vector structure completely reflects the given category Γ (a), partly reflects it (b), or does not reflect it at all (c). These sketches appeal to an intuitive notion of geometric distance corresponding to Euclidean distance. Although similarities between SCO vectors are in fact computed using L_1 distance, this notion is good enough for conveying the qualitative differences between these scenarios.

A suitable learning algorithm should be able to discover a category from the Clump Scenario, and any plausible learning algorithm will fail to identify a category from the Random Scenario. In most realistic cases, however, categories are somewhere between these two extremes, forming an *Intermediate Scenario* as illustrated in Figure 3-8b. The category members tend to form a core cluster (as for the Clump Scenario) which becomes more fuzzy and random towards its edges (as for the Random Scenario). Thus, the Intermediate Scenario in fact represents an entire spectrum of possible constellations, each of which can be thought of as a snapshot that is taken during the process of morphing a Clump Scenario into a Random Scenario: The well-defined edges of the Clump Scenario are already faded but its core is not yet fully dissolved into the unstructured Random Scenario.

However, there is another — qualitatively very different — way in which category constellations can form a blend between Clump and Random Scenario. In such a constellation which I will call *Hybrid Scenario*, the category members show a strong

tendency to group together but they do so in several isolated clusters that are scattered throughout the vector space as illustrated in Figure 3-9. Thus, while an Intermediate Scenario can be viewed as a deficient version of a Clump Scenario, the Hybrid Scenario represents multiple instances of a complete Clump Scenario — it still has the sharp edges of the Clump Scenario and randomness is only found *between* the different instances, rather than *within* them.

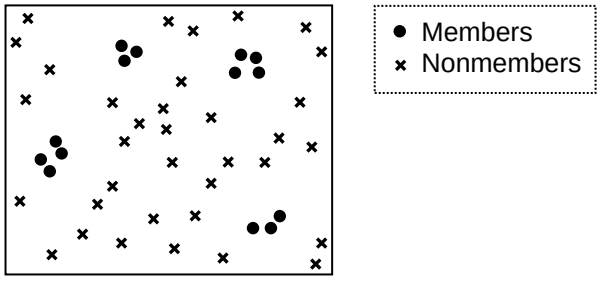

***Figure 3-9**: Hybrid Scenario: Multiple clusters*

Two-dimensional illustration of possible constellations in which the SCO vector structure reflects coherent subclasses of a particular category but does not reflect the category in its entirety.

It is important to distinguish the Hybrid Scenario from an Intermediate Scenario because both suggest very different conclusions. In either scenario, the distributional information captured by the SCO vectors is only moderately useful for acquiring the given category in its entirety. However, while an Intermediate Scenario appears to reflect a genuine insufficiency of the analyzed information, a Hybrid Scenario rather suggests that the information is highly useful but that the benchmark category system may be overly coarse — to the extent that the isolated clusters are linguistically relevant subcategories. To discern these two category scenarios, I used a pair of evaluation scores that is specialized to this very task (cf. 3.3.4).

Of course, all four category scenarios are just prototypical constellations and the question to which of these a particular constellation corresponds is in fact a matter of degree. The purpose of these scenarios is to organize the evaluation of the similarity structure between the SCO vectors, along three complementary questions; namely, (i) whether the similarity structure is at all more informative about a particular benchmark category Γ than would be a Random Scenario (*presence of any information*); and (ii) if so, how useful this information is relative to a Clump Scenario (*usefulness of the*

information); and (iii) if less useful than a Clump Scenario, whether the constellation is best described as an Intermediate Scenario or looks more like a Hybrid Scenario (*consistency of the information*).

It has proven difficult to devise a single evaluation score that is equally sensitive to all three questions. For this reason, I chose to apply a set of three scores. The primary score, Distributional Usefulness, can be applied to answer the two first questions but not the third (cf. 3.3.3). This third question in turn can be addressed by applying the two other scores, Global Coherence and Local Coherence, in combination — but neither of them is well-suited for the first two questions (cf. 3.3.4). Before formally introducing these scores, I first discuss general issues which are relevant for any score that is applied to the first question.

3.3.2 Appropriate random baselines

In order to be able to investigate whether the similarity structure contains any information about a particular lexical category, it is necessary to conversely specify the circumstances under which this structure is considered to provide no information at all about this category. In other words, one needs to formulate the null hypothesis in precise technical terms, by characterizing the distribution of *random baseline* values — i.e., of values that the selected evaluation score would assign to instantiations of the Random Scenario. The values observed for the real SCO vector space can then be compared against this distribution, in order to test the null hypothesis and possibly reject it.

Random baselines can be generated in several ways. First, they can be derived from the original data (either from the input corpus, the SCO vectors, or the similarity values) by removing or obscuring the relevant structure in them that is hypothesized to provide the crucial cues to lexical categories while other properties of the original data are retained. Applying the remaining steps of the method to these randomized data results in a similarity structure which can be assessed in terms of the selected evaluation score; and averaging the values obtained for multiple independent randomizations then yields a random baseline. A second way for constructing a random baseline is to generate random data (corpus, SCO vectors, or similarity values) from scratch, without using the original data at all and to further proceed as above. In a third approach, random baselines are generated directly within the evaluation paradigm by constructing random

constellations of the format on which the particular paradigm operates.[64] And finally, abstracting from this, one can consider all possible constellations within the paradigm and theoretically derive from them the expected value of the chosen evaluation score.

Any random baseline potentially captures a distinct null hypothesis. When derived empirically from the original data, the null hypothesis is that the particular structure removed from the original data is not informative about lexical categories. When a baseline is constructed without the use of the original data, the null hypothesis is that the particular probability distribution by which it is constructed is not less informative about lexical categories than are the actual data. While in principle each null hypothesis would be equally valid, only some are actually reasonable to make in the given situation. To find an appropriate baseline, I considered several candidates which are discussed here.

Maybe the most intuitive way for deriving random baselines from the original data is to scramble the word tokens in the input corpus (across utterances) such that word order regularities are removed while the distribution of utterance lengths and the frequency distribution of word types are both preserved (e.g., Mintz et al., 2002:403). However, my own explorations indicate that the SCO vectors derived from such a randomized corpus in fact contain implicit information about the target words' base frequency: As a statistical artifact of the randomization procedure, target words that occur more frequently in the corpus are more likely to have similar SCO vectors than are less frequent target words.[65] Because benchmark categories differ substantially with respect to the base frequency distribution of their members, this artifact produces a statistical advantage of categories with a relatively high proportion of high-frequency members.[66] Scrambling word tokens thus removes one kind of information from the

[64] For instance, within in their dendrogram-based paradigm, Redington et al. (1998:442) construct random baselines by randomly grouping target words into discrete classes that are comparable in size and number to the discrete classes derived from the actual input data.

[65] Since the frequency distribution of all word types is preserved by the scrambling procedure, the probability of a particular word token to co-occur with particular other words is proportional to their base frequencies. Thus, as the SCO vector of a particular target word is gradually built up across the word's tokens in the scrambled corpus, it approaches a generic vector which reflects the skewed base frequency distribution of context words. If given enough evidence (that is, enough tokens), all SCO vectors would converge on this same generic vector. But because the target words' own base frequencies differ considerably, the SCO vectors of the more frequent target words are more likely to be more similar to the generic vector than are those of the less frequent ones. Given the properties of the L_1 metric, it follows that the more frequent target words are also more likely to be similar to each other than they are to less frequent words (and than are less frequent words with each other).

[66] Applying the scrambling procedure to the *Leo* data, this prediction was confirmed empirically. When the resulting SCO vectors were evaluated against the benchmark categories using the evaluation score *Distributional Usefulness* (cf. 3.3.3), the proportion of category members that are among the 150 most frequent target words is an extremely good predictor of the scores achieved by the categories. In

SCO vectors (lexical co-occurrence and serial position) but in return introduces another one (base frequency). This information is indisputably present in the input and thus available to the child. But it is not represented in the SCO vectors that the given co-occurrence model derives from the input; and it is not intended to be assessed here. The unique null hypothesis captured by scrambling word tokens might be valid for other kinds of models that do incorporate base frequency information, but it is inappropriate for the current purposes.[67]

Another straightforward way to derive random baselines is to randomly generate SCO vectors that are orthogonal to each other (such that any two vectors have their positive vector elements in different dimensions). Thus, all vectors have the same (maximal) L_1 distance from each other such that there is no similarity structure that could possibly provide any (random) cue to lexical category. The outcome of such a simple structure is entirely predictable by the properties of the evaluation score — and it might well serve to explore these properties — but there is nothing random about this structure that would constitute a reasonable baseline. Whereas scrambling word tokens was found to generate artifacts, baselines from orthogonal vectors turn out to be completely pointless. Essentially the same considerations also apply to random baselines that are derived from the actual SCO vectors by generating random permutations of SCO matrix cells — either across the entire matrix or separately within each SCO vector.[68] Given the large proportion of cells containing the value 0, most pairs of SCO vectors would become orthogonal.

For the real SCO vectors, the overall distribution of L_1 distances displays a very wide range of values, with most values occurring for only a few different target word

particular, the proportion for N and ADJ is below 2% and both categories receive substantially less information than would be expected by mere chance (in terms of the null hypothesis implemented by Distributional Usefulness); V and INTJ both have a proportion around 15% and perform roughly at chance whereas the seven remaining categories all have a proportion above 22% and consistently perform better than chance.

[67] To complicate things even further, if one assumes the distributional regularities inherent in the *Leo* corpus to be, in principle, very informative about the benchmark categories, then the resulting SCO vectors of more frequent target words are more likely to reflect the relevant distributional properties, putting categories with more high-frequency members at a statistical advantage. In this sense, the SCO vectors derived from the real data would indeed implicitly exploit some base frequency information. But here, base frequency can only support an existing cue but not provide a cue by itself (as for a scrambled corpus). Even more importantly, this would happen only under the assumption that there *is* some useful information in the target words' actual co-occurrence properties captured by their SCO vectors — and this assumption is precisely what any appropriate null hypothesis is intended to deny. Base frequency information should therefore not be represented in random baselines.

[68] In analogy to the co-occurrence matrix, the term *SCO matrix* refers to the complete set of SCO vectors when arranged beneath each other.

pairs. And the lesson here is that an appropriate random baseline should roughly mimic this dispersed distribution but completely erase the information about lexical categories that may be contained in it. This can be achieved, for instance, when random permutations of SCO matrix cells are generated independently within each column (i.e., dimension).[69] A closely related alternative is to retain the original SCO matrix but to randomize the mapping between SCO vectors and target words.[70] This preserves the original similarity structure but obscures any inherent cue to categories. Abstracting from this idea, a third appropriate type of random baseline can be constructed directly from the set of observed similarity values, by randomizing the mapping between these values and the possible pairs of target words. In practice, it turns out that all three approaches yield very similar random baseline values and therefore capture essentially the same null hypothesis.

In my own analyses, these three kinds of empirical baselines all played a role. My default evaluation score, Distributional Usefulness, is theoretically balanced for randomness such that its expected value is 0. The baselines were used to empirically determine the range of values around 0 that are ascribed to randomness.

3.3.3 Distributional Usefulness

The primary evaluation score, *Distributional Usefulness*, is derived from *Average Precision*, a measure widely used in signal detection theory and information retrieval. Below, I first derive from Average Precision a preliminary score, *Distributional Learnability* which is then further transformed into Distributional Usefulness. To this end, let Λ denote the target lexicon comprising all $L = 1{,}017$ target words. Let Γ be a benchmark category with C target words, and let t be one of them. As a first step, the $L-1$ remaining target words are ranked by their L_1 distance from t in the SCO vector space such that the target word closest to t occupies the first rank, and the most distant target word correspondingly the last rank (Figure 3-10 below).

[69] Alternatively, one can generate such within-column permutations from the raw co-occurrence matrix, rather than from the SCO matrix. In either case, the resulting row vectors have to be rescaled to unit mass.

[70] In effect, this second approach is a special case of the first one, only with the additional constraint that all columns are permuted simultaneously rather than independently.

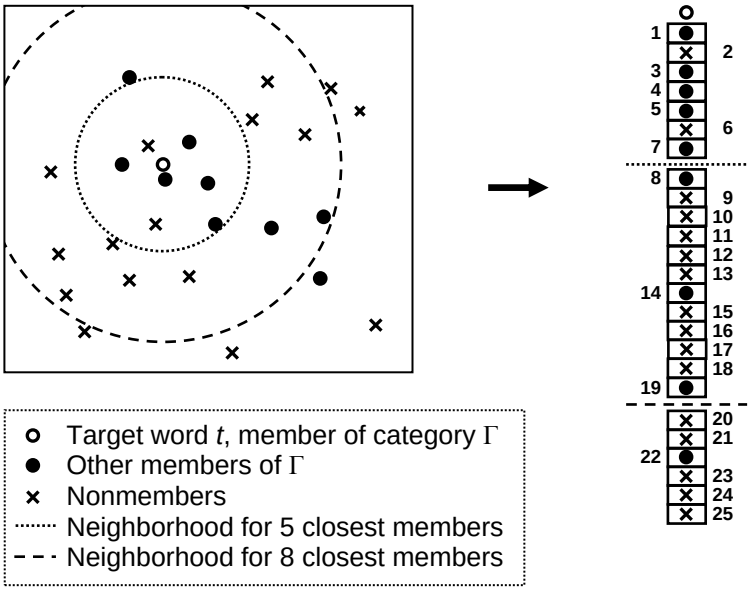

***Figure 3-10**: Transforming the SCO vector space into a rank list*

A hypothetical situation in the SCO vector space for a target word *t* and its benchmark category Γ (left), and the derived rank list (right). Ranks occupied by members of Γ are placed to the left of the list, ranks of nonmembers to its right. The dotted line marks the neighborhood around *t* that is just large enough to contain the 5 closest others members of Γ. The dashed line marks the corresponding neighborhood for the 8 closest members. In terms of L_1 distance, the local neighborhoods in the SCO vector space are not circular. They are depicted here as circular only to appeal to an intuitive notion of distance in space.

Next, one identifies in this list those ranks that are occupied by the $C-1$ other members of Γ (i.e., members other than *t* itself) and writes them as

$$\text{rank}_t(1) \;<\; \text{rank}_t(2) \;<\; \ldots \;<\; \text{rank}_t(C-1)\ . \tag{3}$$

With this notation, the rank list's *Precision at the i-th member* is computed as

$$P_t(i) = \frac{i}{\text{rank}_t(i)}\ . \tag{4}$$

To interpret this ratio with respect to the SCO vector space: It specifies the proportion of category members among all target words in the local neighborhood around *t* that is just large enough to contain the *i* members closest to *t*. As an illustration, two examples of such neighborhoods are marked in Figure 3-10.

Averaging these values across all members in the rank list, as in equation (5), yields the *Average Precision* AP_t for the rank list.

$$AP_t = \frac{1}{C-1}\sum_{i=1}^{C-1} P_t(i) \tag{5}$$

Intuitively, AP_t measures the density of the $C-1$ other category members around t, emphasizing their density in the more immediate neighborhoods of t. Thus, for instance, moving a category member from rank 101 down to rank 150 would reduce AP_t more than moving a member from rank 901 down to rank 950. In this way, Average Precision is sensitive to the *existence* of outliers — i.e., category members that are distributionally relatively dissimilar from t — but less sensitive to the *degree* of their being an outlier. This property is intuitively plausible in the given context because for a category with, say, 80 members, an outlier on rank 901 is hardly any less problematic than an outlier on rank 950: Neither outlier would be discovered to belong to the same category as t. By contrast, an outlier on rank 150 is clearly more problematic than a member on rank 101 (which may not even be considered an outlier). In this sense, the emphasis of Average Precision on the local situation, without ignoring more distant regions altogether, is quite desirable for present purposes.

The quantity AP_t evaluates the category Γ from the perspective of the single member t. To obtain a preliminary score that is independent of the choice of a particular category member, the individual AP_t values are averaged across all members t of Γ:

$$DL_\Gamma = \frac{1}{C}\sum_{t\in\Gamma} AP_t \ . \tag{6}$$

This mean value will be called the *Distributional Learnability* of Γ. It directly inherits from Average Precision the property of measuring the density of category members, with an emphasis on the members' immediate neighborhoods. It can be thought of as quantifying how well the overall category Γ could be acquired by an actual learning mechanism that only exploits distributional patterns captured by the SCO vectors (hence the suggestive label).[71]

[71] This is precisely the score that Zavrel (1996) uses within a closely related approach targeted at machine learning applications. He, however, simply calls it *P-value* or *Average Precision* like the score it is derived from.

Distributional Learnability always yields positive values, and its maximal value 1.0 is equivalent to the Clump Scenario. If the Random Scenario is equated with the assumption that rank lists are generated randomly (by a uniform probability distribution across all possible rank lists), the expected value of Distributional Learnability is

$$\mu_{DL_\Gamma} = \frac{1}{(C-1)\binom{L-1}{C-1}} \sum_{i=1}^{C-1} \sum_{r=i}^{L-C+i} \binom{r-1}{i-1} \binom{L-1-r}{C-1-i} \frac{i}{r} \tag{7}$$

which follows from

$$\mu_{DL_\Gamma} = \frac{1}{C} \sum_{t \in \Gamma} \mu_{AP_t} = \mu_{AP_t} \tag{8}$$

together with the formal derivation given in Appendix B. This value is always positive, and, in most realistic cases, it can be reasonably approximated by the *relative category size* C/L. This implicates that there is a residual Distributional Learnability even when distributional information is not useful at all, and that it is greater for larger categories. This residual and its dependency on the category size may appear counterintuitive at first. But it only reflects the fact that for a larger category, its members are likely to occupy the space at a higher density. In the extreme, the largest possible category comprises all L target words such that the density of its members is trivially maximal (i.e., has the value 1.0).

In a certain sense, Distributional Learnability favors larger categories in general and not only in the case of random constellations. To demonstrate that this general property is meaningful for the given purposes, let me illustrate it in terms of Average Precision AP_t for a particular category member t. Suppose that the other category members are located around t at a medium density above the expected value (cf. Figure 3-11, left). If these members are *densified* by uniformly inserting N new members in adjacent rank slots, relative category size will increase, and so will Average Precision (Figure 3-11, right). Conversely, if this densified category is again *diluted* by randomly removing N members, the resulting constellation is likely to look very much like the original constellation. In this sense, both constellations are equivalent, disregarding relative category size. But the densified constellation corresponds to a more substantial cluster in the actual SCO vector space, with relatively fewer nonmembers intruding; and Average Precision — as well as Distributional Learnability — acknowledges this by assigning a greater value.

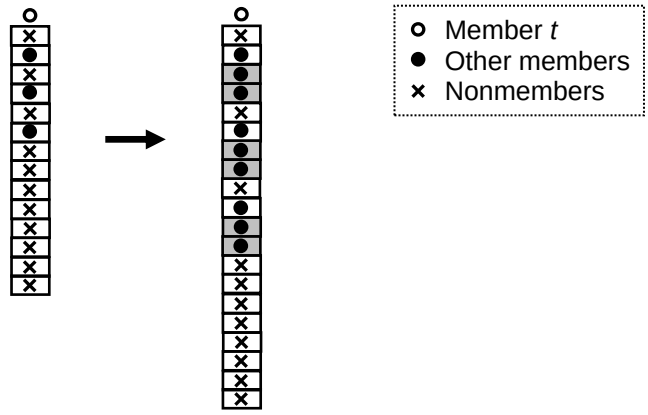

***Figure 3-11**: Influence of relative category size on Average Precision*

A hypothetical rank-list constellation around a category member *t* (left) and a modified constellation derived from it by introducing two extra copies (shaded) for each category member other than *t* itself, resulting in a greater relative category size (right). In the depicted example, relative category size rises from 27% to 48% and Average Precision AP_t increases from .50 (left) to .69 (right).

The general bias towards larger categories is thus a desirable property of Distributional Learnability; but the positive residual is not. It is impractical when the SCO vector constellation for a particular category is compared against a Random Scenario. To solve this, I derived from Distributional Learnability another score, *Distributional Usefulness* DU_Γ by virtue of the simple transformation

$$DU_\Gamma = \frac{DL_\Gamma - \mu_{DL_\Gamma}}{1 - \mu_{DL_\Gamma}} \ . \tag{9}$$

This transformation computes how much better than chance the Distributional Learnability of Γ is (the numerator), as a proportion of how much better than chance it could maximally be (the denominator). Distributional Usefulness DU_Γ can therefore be interpreted as quantifying the extent to which distributional information potentially facilitates the discovery of the category Γ, beyond residual category size effects. It still measures the density of category members, only corrected for the residual category size effect. The score has three crucial properties; namely, (i) it is standardized for the Random Scenario (expected value 0.0); (ii) it is standardized for the Clump Scenario (value 1.0); and (iii) it inherits from Distributional Learnability the emphasis on local neighborhoods and consequently also a general bias towards larger categories. There are other conceivable transformations with these three properties, but the one used here is

the simplest one and should be interpreted as a provisional solution. Distributional Usefulness was therefore used as the default evaluation score.[72]

How do the properties of Distributional Usefulness relate to the three types of questions formulated in subsection 3.3.1? The first question asks whether the similarity structure in the SCO vector space is more informative about a particular category than would be a Random Scenario. Knowing that Distributional Usefulness has the expected value of 0.0 (given the null hypothesis of random rank lists) does not suffice because it is not clear a priori to which degree actual random baselines would vary around this value. To determine this variation empirically, I generated for each of the 11 benchmark categories 10 independent random baselines (from random rank lists, with one such list for each baseline and around each target word).[73] Of the 10×11 resulting Distributional Usefulness values, the highest one was .012.

However, each individual rank list was generated independently whereas in reality, the rank lists derived from SCO vectors co-vary. Therefore, Distributional Usefulness might display much greater variation on more realistic random baselines. To test this as well, I considered the three types of appropriate corpus-derived baselines described in 3.3.2 (within-column permutations of SCO matrix cells; randomized mapping between SCO vectors and target words; randomized mapping between similarity values and target words) and derived for each of them 10 independent random baselines, with one value for each category.[74] Across the 3×10×11 resulting Distributional Usefulness values, the highest one was .017. Interpreting these results rather conservatively, I will conclude for any of the 11 benchmark categories that an observed Distributional Usefulness value above .05 is not merely due to chance. This simple threshold criterion for rejecting the null hypothesis suffices entirely. In particular, it would be neither necessary nor informative to instead operate with tests of statistical significance because, in practice, Distributional Usefulness values observed for the real data are substantially above the conservative threshold and would thus always lead to extremely high levels of significance, whichever statistical test is applied.[75]

---

[72] Note that, in theory, Distributional Usefulness can also become negative when category members are distributed less densely than expected for a Random Scenario. In practice, negative values occur very rarely, and when they do, they are just below 0 and thus essentially reflect the Random Scenario.

[73] These baselines correspond precisely to the null hypothesis that was used to compute the expected value (cf. Appendix B).

[74] For each of these three kinds of random baseline, the empirical mean value of Distributional Usefulness was very close to 0.0 (deviating less than .0001). This confirms that these baselines capture essentially the same null hypothesis also underlying the baseline defined by random rank lists.

[75] Redington et al. (1998) and Mintz et al. (2002) work with evaluation scores that are not standardized for their respective null hypotheses. Instead, they report separate random baselines for each of their

The second question asks how useful the information in the SCO vector space would be for a particular category, relative to a Clump Scenario. To get a sense of the Distributional Usefulness scale with respect to this question, I explored the possible SCO constellations achieving a particular value on this scale. This pilot work suggested that categories with Distributional Usefulness greater than 0.6 come fairly close to a Clump Scenario such that in these cases, distributional information can be considered *very useful* for the given category, and *extremely useful* for values above 0.8. No crucial inferences will be based on these ranges; they merely serve as reference points on the Distributional Usefulness scale.[76]

The third type of question asks whether the SCO vector constellation for a particular category resembles an Intermediate Scenario or rather a Hybrid Scenario. However, while Distributional Usefulness is the first choice for the former two questions, it is not suitable for deciding this third question. Because the score emphasizes the target words' local neighborhoods and is little sensitive to the degree of outliers, constellations in the Hybrid Scenario achieve medium Distributional Usefulness values as do constellations in the Intermediate Scenario.[77] Therefore, some additional evaluation methodology is required to reliably distinguish these two kinds of SCO constellations.

3.3.4 Global Coherence and Local Coherence

For this purpose, I developed a simple score called *Global Coherence*. Because of certain weaknesses, it was supplemented by a secondary score, *Local Coherence*. Roughly speaking, the former score evaluates a category's global configuration whereas

---

analyses. In interpreting the difference between the observed and random baseline values, however, the authors implicitly standardize their scores as well. Zavrel (1996), using the same score that I termed Distributional Learnability, provides no random baselines at all. Being aware of the problem, he proposes to standardize the score against its minimal value, rather than against the expected value.

[76] To give a rough interpretation, for the smaller benchmark categories, a Distributional Usefulness value above .6 is achieved, for instance, by a constellation of SCO vectors in which, on average, each category member is closer to at least 60% of the other category members than to any nonmember; and correspondingly at least 80% for Distributional Usefulness above .8. For the largest categories noun and verb, the corresponding percentages would be 70% (for Distributional Usefulness greater than .6) and 85% (for values above .8).

[77] In a Hybrid Scenario, a typical category member *t* contributes a medium AP_t value because it has some other members in its immediate neighborhood but is fairly distant from a large portion of members in the other clusters. In an Intermediate Scenario, by contrast, category members *t* within the core cluster have more intruders but also more members in their local neighborhoods and thus tend to produce slightly greater AP_t values. But these are compensated for by the low AP_t values contributed by the isolated outlier members *t* that are fairly distant from the core cluster and have hardly any other members nearby.

the latter exclusively looks at the members' local neighborhoods. By contrast, Distributional Usefulness is a blend of both aspects, measuring the global configuration while emphasizing the local neighborhoods. Like Distributional Usefulness, both coherence scores are derived from rank lists of target words.

With the notations used in the preceding subsection, Global Coherence is defined as follows. First, for each given category member *t*, one calculates the average R_t of all ranks occupied by the $C-1$ other category members as in equation (10).

$$R_t = \frac{1}{C-1} \sum_{i=1}^{C-1} \text{rank}_t(i) \tag{10}$$

This average rank is always positive and assumes its minimal value $\min_{R_t} = C/2$ for the Clump Scenario, i.e., when the category members occupy the ranks 1, 2, …, $C-1$. Given the null hypothesis of random rank lists, the expected average rank of category members is simply the rank list's center $\mu_{R_t} = L/2$. With this, R_t is transformed to the ratio

$$S_t = \frac{\mu_{R_t} - R_t}{\mu_{R_t} - \min_{R_t}} = \frac{L - 2R_t}{L - C} \ . \tag{11}$$

This value S_t is standardized in the sense that its possible values range from −1.0 to +1.0, and given the null hypothesis of random rank lists, its expected value is 0.0. It quantifies how much the average rank of the $C-1$ category members is above or below the rank list's center, relative to how far above this center it could maximally be.

The *Global Coherence* GC_Γ of a particular category Γ is now defined as the average of S_t values, computed across the rank lists of all members *t* of Γ:

$$GC_\Gamma = \frac{1}{C} \sum_{t \in \Gamma} S_t \ . \tag{12}$$

Global Coherence values range from −1.0 to +1.0. A category Γ with Global Coherence $GC_\Gamma = 1$ corresponds to the Clump Scenario. The Random Scenario, by contrast, yields a Global Coherence value around 0.[78] The score shares these properties with Distributional Usefulness. However, both scores behave quite differently on the spectrum of possible SCO vector constellations between a Random Scenario and a Clump Scenario. The crucial differences arise from the fact that Global Coherence is a

[78] Negative values have mainly theoretical status and are not relevant in the study.

purely global score while Distributional Usefulness is a global score with an emphasis on local neighborhoods. This has two important consequences.

First, Global Coherence is highly sensitive both to the existence of outlier members and to their degree of being an outlier whereas Distributional Usefulness is mainly affected by the number of outliers and much less by their particular degree. To pick up the earlier example (cf. 3.3.3): Moving a category member from rank 101 down to rank 150 affects AP_t (and thus Distributional Usefulness) more than moving a member from rank 901 down to rank 950. By contrast, the decrease in S_t (and thus in Global Coherence) is the same in both cases.

Second, Global Coherence does not have the bias of Distributional Usefulness towards larger categories. When a category is densified (or diluted) as was illustrated earlier (Figure 3-11), Global Coherence will remain constant while Distributional Usefulness increases (or decreases, respectively). But as a direct consequence, Global Coherence in turn strongly favors *smaller* categories — though in a qualitatively very different way. In a nutshell, for a relatively small category, distributional information need not be very useful at all to yield fairly high Global Coherence values.

Both types of category size effects are intrinsically connected such that when one is controlled for, the other one necessarily pops up. Further, under the requirement that a candidate score is standardized both for the Clump Scenario (maximal value 1.0) and the Random Scenario (expected value 0.0), Global Coherence is a natural way of controlling for the category size effect on Distributional Usefulness — and this latter score is in turn a natural way of controlling for the category size effect on the former score. In this sense, Distributional Usefulness and Global Coherence can be regarded as complementary scores.

However, unlike the category size effect on Distributional Usefulness, the strong bias of Global Coherence towards smaller categories is not desirable for general purposes. In particular, it disqualifies Global Coherence with respect to the question of how useful the similarity structure in the SCO vector space would be for a particular category, relative to a Clump Scenario. And for the same reason, it makes this score clumsy with respect to the question of how much better than chance a given SCO vector constellation would be.[79]

[79] The expected value of Global Coherence, given the null hypothesis of random rank lists, is 0.0. But random baselines vary considerably around this value, and this variation largely depends on category size.

The strength of Global Coherence lies therefore in its first distinct property. Because it is sensitive to the number and degree of outlier members, it can in principle distinguish a Hybrid Scenario from an Intermediate Scenario. Everything else being equal, Global Coherence is always lower for a prototypical Hybrid Scenario than for a prototypical Intermediate Scenario because in the former scenario, a large proportion of category members are distant outliers to each other. However, to apply Global Coherence as a tool for discriminating the two types of category constellations, its bias towards smaller categories is problematic because it entails that neither scenario can be associated with a category-independent range of Global Coherence values. Ideally, one would systematically explore such ranges for each possible category size; but I decided against this general solution because I only applied this score under very specific conditions, namely, to directly compare two categories of roughly the same size (cf. 4.4). If, under these conditions, one makes sure that for each category, its members have a local tendency to cluster with other members, any substantial difference in Global Coherence between the two categories would indicate that the more coherent category leans more towards an Intermediate Scenario while the other one is better described as a Hybrid Scenario.

To test for the tendency of local clusters, Global Coherence is supplemented by a secondary score, *Local Coherence*. This is a very simple score, constructed by selecting for each member t of a particular category Γ a local neighborhood around t that is just large enough to comprise the 10 closest nonmembers. If M_t denotes the number of other members of Γ in this local neighborhood, then the *Local Coherence* LC_Γ of category Γ is computed by averaging these M_t values across all members t of Γ as in equation (13).

$$LC_\Gamma = \frac{1}{C}\sum_{t\in\Gamma} M_t \quad . \tag{13}$$

The range of possible LC_Γ values (from 0 to $C-1$) and their probability distribution depend largely on the category size C. Standardizing LC_Γ to control for this dependency would typically involve introducing nonlocal aspects of coherence which would run counter to the very purpose of this score. But as long as both coherence scores are only applied to categories of roughly equal sizes, this dependency is not further problematic.

3.3.5 Alternative evaluation scores

In addition to the evaluation scores described above, I originally considered several alternatives that fall into three different evaluation paradigms. The first one is the rank list paradigm on which all the above scores are based. Using rank lists translates the multidimensional problem into a one-dimensional one and makes it assessable by a large body of scores designed for ranked items. All alternative scores of this kind that were considered either produced severe artifacts on appropriate random baselines, were overly critical towards slight deviations from the Clump Scenario, or turned out to be simple transformations of Global Coherence (more precisely, of the quantity S_t from which Global Coherence is derived). In particular, such a transformation was found for *Normalized Recall* NR_t — a measure which, like Average Precision, is used in information retrieval (e.g., Belew, 2001) and signal detection theory (e.g., McNicol, 1972).[80] Its relation to S_t is a simple linear one: $S_t = 2\,NR_t - 1$. Another measure that was taken from information retrieval research is *Expected Search Length* ESL_t (Belew) which turns out to be precisely identical with the average rank R_t of category members. From equation (11), it follows that if category size C and lexicon size L are held constant, S_t becomes a linear transformation of ESL_t with $S_t = (L - 2\,ESL_t) / (L - C)$.

A second evaluation paradigm computes for each benchmark category Γ the *centroid* of all SCO vectors belonging to this category, that is, the vector

$$v_\Gamma = \frac{1}{C} \sum_{v \in \{\text{SCO vectors of } \Gamma\}} v \ . \tag{14}$$

Various scores can be devised that essentially evaluate how representative these centroids are of the corresponding category members and how unrepresentative of the nonmembers.[81] For instance, one can compute for each category the proportion of members that are closer (in terms of L_1 distance) to their category centroid than to the centroid of any other category. Other scores within this paradigm apply the rank list paradigm to each category centroid, by ranking all target words by their L_1 distance from the centroid. In this way, any of the scores defined within the rank list paradigm can be applied to the centroid approach as well. However, systematic tests revealed that, given any of the appropriate null hypotheses discussed earlier (cf. 3.3.2), all these

[80] McNicol (1972) refers to this measure only by the technical term $P(A)$.
[81] I wish to thank Rik Belew for suggesting this possibility.

centroid-based scores generate severe artifacts as they fail to reflect the absence of information for various kinds of random baselines.

The third evaluation paradigm was inspired by Burgess and Lund (1997). Taking the set of L_1 distances between any two members of a given category as one group and the set of L_1 distances between members and nonmembers as a second group, one can compute an ANOVA (which in this case is a simple *t*-test) for the average distance values in the two groups. Obviously, any appropriate null hypothesis formulated for SCO vectors — or for the resulting distribution of L_1 distance values — translates into the null hypothesis that the two groups have identical sample means. And although the statistical assumptions underlying ANOVA are partly violated by the two groups of L_1 distances — e.g., the distance values are not entirely independent of each other —, this method appears to work well, in practice, to test whether a constellation of SCO vectors constitutes a Random Scenario for a particular category. But because it dissociates distance values from the pairs of target words they were observed for, crucial aspects of the topographic structure in the SCO vector space are not evaluated by ANOVA — very different constellations may produce the same overall distribution of distance values. Therefore, ANOVA can in principle only help to decide whether the SCO vectors provide *any* information on a particular category; but it cannot evaluate how useful the information might be, nor can it distinguish between a Hybrid and an Intermediate Scenario.

Chapter 4
Distributional information in the input

In this chapter, the distributional information present in the *Leo* corpus is investigated from a number of complementary perspectives. The first section asks how useful distributional regularities in the input might be for discerning the different lexical categories, and which particular categories are distributionally most similar to each other. Section 4.2 explores various ways in which a distributional learner might fail to fully exploit all available data, and the degree to which this might alter the extracted distributional information and affect its usefulness for category acquisition. The particular distributional properties that characterize each category are determined and compared in section 4.3. Section 4.4 presents a detailed contrastive analysis of the distributional situation for the two largest categories, noun and verb, and seeks to account for the fundamental differences between them in terms of grammatical regularities and speakers' usage preferences. The chapter concludes with assessing the distributional consequences when some specific empirical evidence about language development is taken into consideration.

4.1 Usefulness of the information

In this section, the formal evaluation scheme introduced in the previous chapter is applied to assess how informative the distributional regularities in the *Leo* corpus are with respect to the 11 benchmark categories. First, each category is compared against all other categories simultaneously (subsection 4.1.1) before a more fine-grained analysis looks at the distinction between any two categories separately (4.1.2). Both subsections specify possible nondistributional factors that could partly account for the obtained results.

4.1.1 Default analysis

In subsection 3.1.3, we observed for a two-dimensional projection of the SCO vector space that the benchmark categories roughly correspond to certain regions in the projected space but that this correspondence is not as good as it could be. The Distributional Usefulness measure defined in 3.3.3 yields more objective judgments on the goodness of this correspondence, and moreover, it evaluates all of the information in the full high-dimensional vector space rather than just in two dimensions. Figure 4-1 shows the resulting Distributional Usefulness score for each of the 11 benchmark categories.

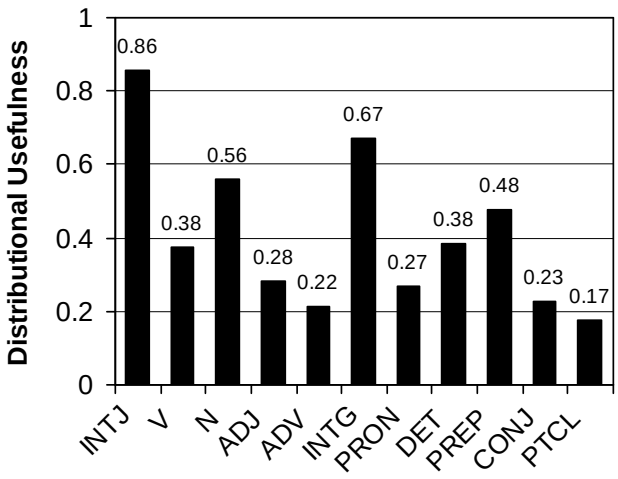

Figure 4-1: *Default evaluation of distributional information*

Distributional Usefulness score for each benchmark category.

All Distributional Usefulness values are substantially above the threshold .05, indicating that none of the categories constitutes a Random Scenario in the SCO vector space — that is, all categories receive some useful information from co-occurrence patterns (see p. 86 for the reasoning behind this threshold criterion). But just how useful this information is, varies considerably across categories. The category of **interjections** benefits by far the most from distributional cues and looks very much like what was described as a Clump Scenario. The next highest Distributional Usefulness values are found for the categories of **interrogative words**, **nouns**, and **prepositions**. **Verbs** and **determiners** achieve medium scores while the five remaining categories, **adjective**, **pronoun**, **conjunction**, **adverb**, and **particle** yield rather low Distributional Usefulness although they, too, are still clearly better than chance.

In sum, this default analysis of the full SCO vector space confirms the observations from the visual inspection in subsection 3.1.3. All benchmark categories benefit from distributional information, but this information alone does not suffice to correctly reflect the entire category structure in the SCO vector space. These are important findings which confirm the general results of the earlier studies for the German corpus at hand.

Going beyond these results, the current study is concerned with the distributional causes for the sizable variation of Distributional Usefulness levels across categories. However, before these causes can be uncovered, it is crucial to be sure that this variation indeed mainly reflects differences in the predictive power of the categories' distributional properties and not some other factors.

For instance, categories differ considerably with regard to the number and frequency distribution of their members (cf. Table 2-1 on p. 50). Thus, a priori, it was a realistic possibility that the observed Distributional Usefulness pattern chiefly arises from differences in category size and frequency distribution. To rule out this possibility, the influence of these two factors on Distributional Usefulness was assessed. A third potential factor will be considered in the next subsection.

It turned out that there is virtually no correlation between category size and the Distributional Usefulness values observed in the default analysis, as visual inspection and formal tests for monotonic correlation (Spearman's $\rho = .091$, $p = .790$, $N = 11$) revealed. Thus, category size plays hardly any role in explaining the Distributional Usefulness pattern.

This complete absence of a category size effect may be surprising at first. Recall from subsection 3.3.3 that Distributional Usefulness was derived as a transformation from Distributional Learnability that removes the category size effects for random vector constellations but preserves the plausible bias towards larger categories for better-than-random constellations. Thus, when Distributional Usefulness is positive — as is the case for each of the 11 benchmark categories — one would actually expect to see category size effects such that, everything else being equal, a larger category achieves a greater Distributional Usefulness.

To directly study this predicted effect and to reconcile it with the noncorrelation between category size and observed Distributional Usefulness, an experiment was conducted in which all categories were forced to have the same size. The largest possible size was 13 because this is the size of the smallest category (conjunction). Category sizes were reduced by randomly selecting 13 members of each category. To control for random effects of this selection, 20 independent runs of category reduction

were carried out. The means of the resulting Distributional Usefulness scores are shown in Figure 4-2.

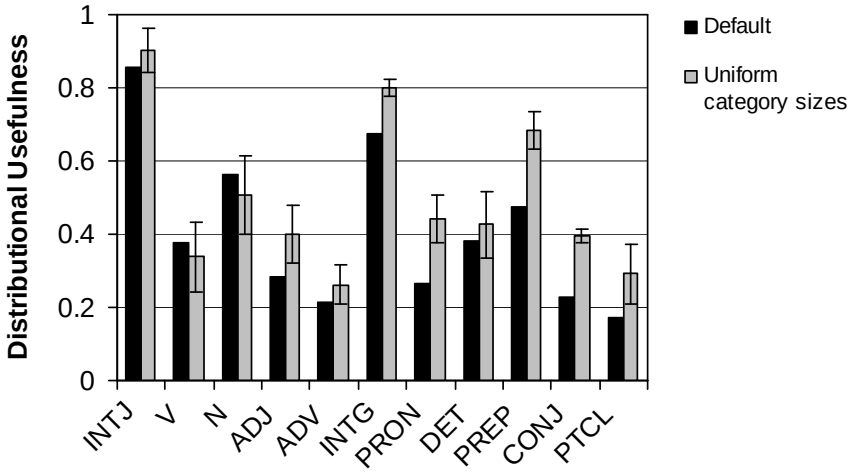

Figure 4-2: *Category size effects on Distributional Usefulness*

Effects on Distributional Usefulness when all categories are reduced to 13 members each. Subsets of category members were selected randomly in 20 independent runs. The graph shows the average Distributional Usefulness values across all 20 runs (whiskers represent standard deviation) in comparison to the default analysis (copied from Figure 4-1).

Relative to the default analysis, Distributional Usefulness drops only for the two largest categories, noun and verb, which are also the only categories whose relative category size C/L decreases upon the experimental manipulation.[82] For all other categories, Distributional Usefulness increases, and this increase is the greatest for the five smallest categories of conjunctions, prepositions, interrogative words, pronouns, and particles which are necessarily the categories whose relative category sizes are raised the most. And indeed, there is a significant positive correlation between the change in relative category size and the change in Distributional Usefulness (Spearman's $\rho = .845$, $p = .001$).[83] This confirms the existence and direction of the expected effect of relative category size.

But despite this effect, the categories' rank order in terms of Distributional Usefulness remains essentially unaltered by the experimental manipulation — especially when the two largest categories, noun and verb, are ignored. Moreover, the higher score of the noun category relative to the verb category is preserved. Thus,

[82] Strictly speaking, the relative category size of the next largest categories, adjective and adverb, decreases as well, but this decrease is marginal (from 9.4% and 9.2%, respectively, to 9.1% each).

[83] Visual inspection suggests that this relation is nonlinear.

although relative category size influences Distributional Usefulness, it is not the main factor explaining the scores we observed in the default analysis. It is simply not the case that the higher Distributional Usefulness values are assigned to the larger categories. This finding is consistent with the noncorrelation that was reported earlier. Whatever factors are causing this Distributional Usefulness pattern, category size effects only operate on top of them without substantially altering this pattern.

Let us turn to the second potential source of differences in Distributional Usefulness, viz., the categories' frequency distributions. It was found that a category's Distributional Usefulness value observed in the default analysis does not correlate significantly with the category members' average frequency (Spearman's $\rho = -.382$, $p = .247$), nor with their median frequency (Spearman's $\rho = -.427$, $p = .190$).

However, it would be premature to infer from these figures that frequency distribution has no significant influence on Distributional Usefulness — for the benchmark categories have unique frequency distributions that are not well summarized or compared by their average or median.

Therefore, an experiment was conducted in which all categories were forced to have precisely the same frequency distribution, simply by restricting each target word to the same number of tokens. The largest possible number of tokens was 100 because this is the frequency threshold by which target words were selected. Target word tokens were selected randomly in 20 independent runs. The means of the resulting Distributional Usefulness scores are shown in Figure 4-3 below.

Distributional Usefulness drops substantially for six categories while it increases, though slightly, for the five other categories (interjection, verb, noun, adjective, and interrogative word).[84] These five categories are also characterized as those whose median frequency in the default analysis is below 300 (cf. Table 2-1 on p. 50). This suggests that median frequency may be a good predictor of the change in Distributional

[84] Intuitively, one would expect Distributional Usefulness to decline for all categories because the experimental manipulation reduces the statistical basis for almost all target words. This general intuition will be confirmed in a related experiment which reduces the target words' base frequencies at fixed rates (subsection 4.2.2). In the current experiment, however, the base frequencies of highly frequent target words are reduced at higher rates than are those of less frequent words. Thus, although the statistical basis becomes smaller for all categories, it drops most substantially for the categories with the greatest proportion of highly frequent target words. These categories tend to become less compact in the SCO vector space which explains their drop in Distributional Usefulness. At the same time, it is possible that some of their members move so far that they actually move out of regions occupied by other categories. This would be to the advantage of categories with a relatively low proportion of highly frequent members. The fact that Distributional Usefulness was found to increase for these low-frequency categories suggests that this positive effect exists and that it even overcompensates the negative effect from their reduced statistical basis.

Usefulness (upon the experimental manipulation); and as it turns out, there is a highly significant linear correlation between these two variables (Pearson's $r = -.887$, $p < .001$).

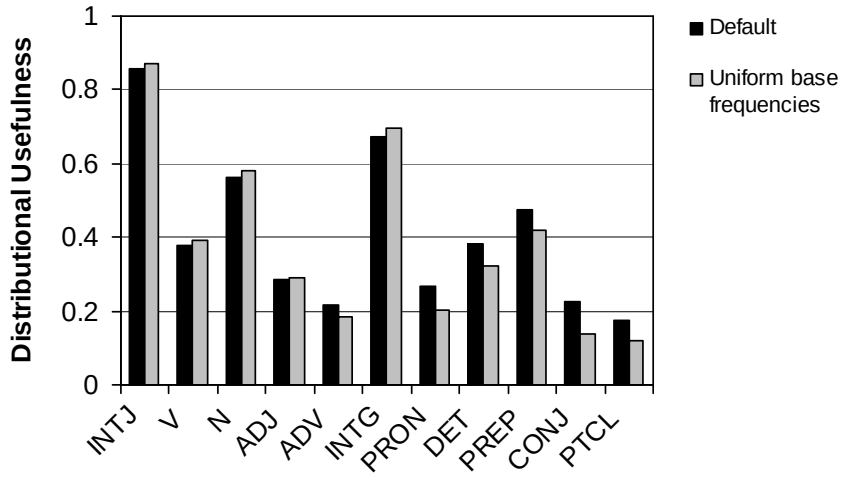

Figure 4-3: *Effects of frequency distribution on Distributional Usefulness*

Effects on Distributional Usefulness when all target words are restricted to 100 occurrences each. Target word tokens were selected randomly in 20 independent runs. The graph shows the average Distributional Usefulness values across all 20 runs in comparison to the default analysis (copied from Figure 4-1). Standard deviation figures (ranging from .003 to .024) are too small to be displayed.

But despite this influence of the frequency distribution, the overall Distributional Usefulness pattern does not change very much. Hence, we have essentially the same situation as was observed earlier for category size effects. A category's frequency distribution influences its Distributional Usefulness score, but this effect contributes very little to explaining the scores observed in the default analysis.

In sum, neither the categories' different sizes nor their different frequency distributions can account for the observed Distributional Usefulness pattern. On the other hand, both factors do influence Distributional Usefulness; and since this study is interested in the categories' intrinsic distributional characteristics it would seem reasonable to systematically control for such nondistributional influences, in ways similar to the two experiments reported here. But for the following two reasons, I decided against this option. First, the categories' variation with respect to category size and frequency distribution are linguistic facts that the child has to deal with as well. And second, both effects are plausible. The fact that Distributional Usefulness favors larger categories is a desirable property of this score (as was argued in subsection 3.3.3), and

removing this bias would inevitably introduce a different, undesirable type of category size effect (cf. 3.3.4). Likewise, it is an intuitively plausible phenomenon that distributional information tends to become less reliable as the statistical basis of this information is reduced (cf. related explorations in subsection 4.2.2).

4.1.2 Distributional confusability of categories

The general results of the default analysis indicate that, except maybe for interjections, each category partly overlaps with other categories in the SCO vector space. But it may do so with different categories at varying degrees. That is, based on distributional information, a particular category may be easy to distinguish from some categories but readily confused with other categories. Such differences in *distributional confusability* are likely to provide important insights into the distributional structure of the input data.

To investigate the patterns of distributional confusability, I determined how well any category Γ_1 separates in the SCO vector space from any second category Γ_2. A quantitative measure for this degree of separation was obtained by temporarily restricting the target lexicon to the members of Γ_1 and Γ_2 and computing Distributional Usefulness for Γ_1 within this sublexicon. This value will be called the *separation* of Γ_1 from Γ_2. The lower the separation value, the greater the probability that a purely distributional learner will confuse Γ_1 with Γ_2, i.e., that such a learner will not properly acquire the distinction between these two categories. Note that separation values need not be symmetric — it may be the case that Γ_1 separates well from Γ_2 but not vice versa, as illustrated in Figure 4-4.

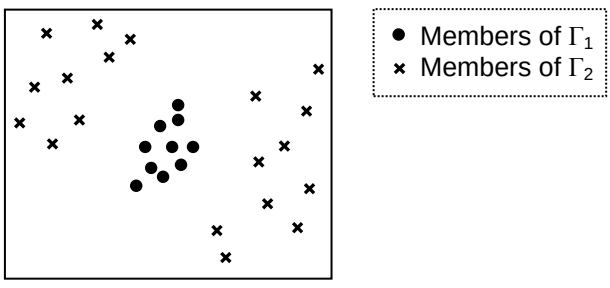

***Figure 4-4**: Asymmetric separation between two categories*

A hypothetical situation in which one category (Γ_1) separates very well from another category (Γ_2) but not vice versa.

The resulting separation values are listed in Table 4-1, along with each category's Distributional Usefulness score on the entire lexicon, for comparison (copied from Figure 4-1 on p. 94). **Interjections** — the category with the highest overall Distributional Usefulness value — separate almost uniformly well from each of the other categories. A second category with similarly good separation across the board is that of **interrogative words**; they separate least well from conjunctions and interjections.

At the other extreme, **verbs** are rather easily confused with about any other category, but most likely so with adverbs, conjunctions, and particles. Furthermore, any category (other than conjunction) appears to be distributionally more problematic for acquiring the verb category than vice versa — the separation values in the verb row of Table 4-1 are clearly lower than the corresponding values in the verb column. This suggests that the topographic relation between the verb category and any other category can be portrayed as that between categories Γ_2 (verbs) and Γ_1 (the other category) in Figure 4-4 above.[85] Verbs thus appear to be surrounding the majority of non-verbs in the SCO vector space — a finding which I shall return to in subsection 4.4.1.

Conjunctions display the same asymmetric confusability pattern: They are easily confused with almost any other category while none of these other categories (except for verbs) is easily confused with conjunctions. Translated to the SCO vector space, this indicates that conjunctions, like verbs, are located around most other categories.

By contrast, **nouns** separate very well from most categories, with the exception of adjectives and adverbs. Their topographic relationship with adjectives is clearly asymmetric, with **adjectives** surrounding a good portion of the noun category. Additionally, adjectives also do not separate very well from adverbs, particles, and determiners but in these cases, confusability is approximately symmetric which implies that adjectives overlap with each of these categories, rather than surrounding them. An asymmetric relation of adjectives, however, can be observed for interjections and prepositions; but these asymmetries hold at fairly high separation values implying that adjectives surround only small portions of these other two categories.

Adverbs in turn are most confusable with particles, pronouns, and adjectives. Likewise, **pronouns** separate worst from particles and adverbs, and **particles** from

[85] Of course, the topographic constellation depicted in the figure is an idealization that applies best when the separation value of the other category from the verb category is close to 1.0. But to the extent that the observed separation values are below 1.0 (Table 4-1), one must conclude that most of these other categories are not only surrounded by verbs but also partly occupy the same regions as verbs.

adverbs and pronouns. Adverbs, pronouns, and particles thus form a group of mutually confusable categories — with roughly symmetric topographic relations. This means that all three categories substantially overlap in space, rather than surrounding each other. Most striking is the extremely low mutual separation between adverbs and particles — the lowest value observed for any pair of categories. These two categories are virtually indistinguishable based on distributional information. Another observation that can be made for all three categories (adverb, particle, and pronoun) is their clearly asymmetric relationship with prepositions and interjections. To some extent, each of the former three categories appears to be surrounding each of the latter two. Pronouns additionally also surround some portion of the noun category.

Table 4-1: *Pairwise separation between categories*

Category Γ_1	**DU_{Γ_1}** [a]	**Category Γ_2** [b]										
		INTJ	V	N	ADJ	ADV	INTG	PRON	DET	PREP	CONJ	PTCL
INTJ	.86	—	.93	.95	.91	.94	.93	.96	.98	.98	.92	.91
V	.38	.50	—	.47	.46	.33	.52	.40	.59	.49	.36	.36
N	.56	.86	.70	—	.51	.56	.91	.73	.80	.90	.89	.70
ADJ	.28	.69	.57	.34	—	.39	.83	.62	.45	.68	.80	.40
ADV	.22	.75	.47	.50	.41	—	.82	.39	.57	.52	.64	.09
INTG	.67	.80	.84	.92	.92	.87	—	.89	.91	.96	.75	.90
PRON	.27	.75	.51	.51	.65	.33	.74	—	.44	.66	.62	.34
DET	.38	.86	.75	.61	.53	.49	.84	.50	—	.53	.78	.44
PREP	.48	.90	.80	.88	.83	.64	.91	.84	.66	—	.68	.57
CONJ	.23	.45	.35	.54	.66	.39	.37	.48	.60	.51	—	.38
PTCL	.17	.65	.51	.59	.41	.10	.83	.37	.48	.43	.60	—

[a] Overall Distributional Usefulness score for separating category Γ_1 from all other categories simultaneously (copied from Figure 4-1).

[b] Table cells specify Distributional Usefulness of category Γ_1 when the target lexicon is restricted to members of Γ_1 and Γ_2, thus quantifying how useful distributional information is to distinguish Γ_1 from Γ_2.

Prepositions separate exceedingly well from most categories, and in general much better than these other categories in turn separate from prepositions. The fact that their overall Distributional Usefulness (.48) is not higher arises from their partial overlap with particles, and to a lesser degree also with adverbs, determiners, and conjunctions.

Finally, **determiners** also show some exceedingly high separation values but at the same time overlap with several categories, most of all with particles, adverbs, and pronouns. Their separation from adjectives and prepositions is also not very high but, at

least in the latter case, the topographic relation is clearly asymmetric, with determiners surrounding prepositions. Additionally, determiners also surround some portion of the noun category.

These confusability relations will later be linked to the categories' distributional properties (sections 4.3 and 4.4). However, in analogy to the control experiments in the previous subsection, it is important to rule out the possibility that the observed confusability pattern mainly reflects some other, nondistributional factors.

One potential factor that comes to mind is *categorial ambiguity* which is quite likely to influence the confusability relations between categories to some degree. For instance, a word that is sometimes used as a conjunction and at other times as a preposition (e.g., *bis*; English: *until*) will derive its distributional evidence from all its instances and thus share some distributional properties with members of both these categories — to the extent that distributional properties generally *do* reflect lexical category membership. But the coding scheme assigned every target word to only one benchmark category (*bis* was classified as a preposition); and in consequence, ambiguous items like *bis* are bound to increase the distributional confusability between their categories (i.e., to decrease the corresponding separation values).[86]

The degree of categorial ambiguity was documented for each category in Table 2-2 (p. 52) which is reprinted here for convenience. For the current purposes it is important to read the table both by its rows and by its columns: For instance, for the confusability relation between prepositions and conjunctions, the 13.3% of all prepositions that also have a secondary membership in the conjunction category are potentially as problematic as are the 23.1% of all conjunctions with a secondary membership in the preposition category. Therefore, I tentatively interpret the average of these two percentages (i.e., 18.2%) as quantifying the overall degree of ambiguity between the two categories (analogously for any other pair of categories).[87]

[86] This reasoning illustrates that, whereas the effects of category size and frequency distribution (cf. 4.1.1) operate in some sense *orthogonal* to distributional information, the predicted effects of categorial ambiguity are actually mediated by this information. Indeed, the mere fact that there are such effects of categorial ambiguity confirms once again the general usefulness of distributional information. One possible avenue for further investigation not pursued here would therefore be to define particular classes of ambiguous words that can all be used in the same set of major categories and to characterize these classes by their distributional properties. Although these properties are all inherited from the major categories involved, they are likely to be unique in their combination.

[87] An alternative way to integrate both percentages into one single number would be a weighted sum where the weights reflect the category sizes. However, the simple average that was used here suffices for current purposes.

Table 2-2 ***(copy):*** *Categorial ambiguity by benchmark category*

Benchm. category	Ambig. (in %) [a]	Secondary membership in category … (in %) [b]										
		INTJ	V	N	ADJ	ADV	INTG	PRON	DET	PREP	CONJ	PTCL
INTJ	23.4	—	6.5	10.4	5.2	3.9						6.5
V	18.4	1.0	—	12.2	2.4			2.8	0.3			
N	15.3	1.5	12.3	—	2.2			0.4	0.4			0.4
ADJ	60.4	4.2	24.0	26.0	—	7.3		2.1		1.0		12.5
ADV	61.7		5.3	3.2	18.1	—		2.1		2.1	8.5	50.0
INTG	100.0					23.5	—	88.2	35.3	5.9	11.8	11.8
PRON	40.0			2.9		8.6		—	37.1			2.9
DET	83.6		9.8	1.6	1.6	3.3		65.6	—	9.8		6.6
PREP	93.3					46.7				—	13.3	80.0
CONJ	84.6		15.4	7.7		30.8		7.7		23.1	—	38.5
PTCL	92.5	5.7	17.0	7.5	47.2	75.5		3.8	7.5	7.5	15.1	—

[a] Proportion of ambiguous members of the benchmark category (derived from Table 2-1).

[b] Proportion of benchmark category members that can also be used in particular other categories. Missing values are 0.0%.

It turned out that these average pairwise ambiguity figures correlate significantly with the pairwise separation values reported in Table 4-1 (Spearman's $\rho = -0.426$, $p < .0001$, $N = 110$). This confirms the expected statistical effect of categorial ambiguity, namely that the separation of one category from another tends to be worse when the degree of ambiguity between both categories is greater. But this effect is not very strong which suggests that most of the variation across the separation values cannot be explained by differences in pairwise ambiguity figures. At the level of overall Distributional Usefulness scores (i.e., the default analysis), the influence of categorial ambiguity vanishes altogether. Mirroring the frequency analyses of subsection 4.1.1, no significant correlation was found between a category's overall degree of ambiguity (reported in the second column of the Table 2-2) and its Distributional Usefulness score in the default analysis (Spearman's $\rho = .155$, $p = .650$). Thus, like category size and frequency distribution, categorial ambiguity plays essentially no role in explaining the Distributional Usefulness pattern.

To study the influence of categorial ambiguity on distributional confusability more directly, I recomputed Table 4-1, this time restricting the pairwise separation between any two categories Γ_1 and Γ_2 to those members of either category that have no homonym in the other category. For instance, the modified separation between the categories preposition and conjunction is computed between those 86.7% of all

prepositions that can *not* be used as conjunctions and the 76.9% of all conjunctions that can *not* be used as prepositions.

The resulting separation values are shown in Table 4-2. In comparison to Table 4-1, separation increases for the majority of category pairs (56 such pairs) and remains unchanged for most of the remaining pairs (40) while it decreases for only a few exceptions (10). Furthermore, in all cases of decreasing separation, and in many cases of increasing separation, the change is fairly small. For some category pairs, however, their separation value increases substantially — most remarkably, the separation between pronouns and determiners rises from .44 to .77, and, conversely, the separation between determiners and pronouns climbs from .50 to .82.

Table 4-2: *Pairwise separation between categories, excluding relevant ambiguous members*

Category Γ_1	**Category Γ_2**										
	INTJ	V	N	ADJ	ADV	INTG	PRON	DET	PREP	CONJ	PTCL
INTJ	—	.96	.97	.94	.94	.93	.96	.98	.98	.92	.94
V	.52	—	.46	.48	.33	.52	.40	.61	.49	.35	.33
N	.89	.70	—	.58	.56	.91	.74	.80	.90	.90	.74
ADJ	.72	.55	.41	—	.42	.83	.64	.45	.68	.80	.42
ADV	.75	.45	.50	.45	—	.85	.40	.58	.54	.69	.16
INTG	.80	.84	.92	.92	.87	—		.99	.97	.78	.91
PRON	.75	.52	.51	.65	.35		—	.77	.66	.62	.35
DET	.86	.73	.62	.53	.48	.94	.82	—	.56	.78	.53
PREP	.90	.80	.88	.83	.62	.93	.84	.73	—	.94	
CONJ	.45	.35	.54	.66	.50	.47	.43	.60	.69	—	.51
PTCL	.73	.53	.59	.48	.21	.86	.40	.53		.78	—

Note. Table cells specify in terms of Distributional Usefulness how well the members of category Γ_1 that are not simultaneously secondary members of Γ_2 separate from those members of category Γ_2 that are not simultaneously secondary members of Γ_1. Empty cells signal cases where at least one of the two categories has less than five remaining members such that computing a separation value would not be meaningful.

In sum, the distributional separation between any two categories tends to increase when the ambiguous members between them are excluded from the analysis. But this increase is generally not as impressive as one might have expected, and the overall pattern of distributional confusability between categories remains essentially the same. However, in the above modification, categories still contain ambiguous members: For the particular example of pronouns and determiners, the excluded words were those pronouns that can be used as determiners, and those determiners that can be used as

pronouns; but any remaining pronouns or determiners that can also be used as, say, verbs, were still included in computing this modified separation value. Intuitively, the categorial ambiguity directly pertaining to determiners and pronouns should have the strongest effect on their confusability; but other ambiguous items can also influence confusability because the members of a particular category that are not categorially ambiguous at all should have more homogeneous distributional properties and thus form more compact clusters in the SCO vector space than the full set of category members.

In a second modification, I therefore excluded *all* 384 categorially ambiguous target words from the analysis to study confusability between the entirely unambiguous *core* categories. This manipulation retains 59 unambiguous interjections, 235 verbs, 227 nouns, 38 adjectives, 36 adverbs, 21 pronouns, and 10 determiners (cf. Table 2-1 on p. 50). Each of the remaining four categories — interrogative words, prepositions, conjunctions, and pronouns — has less than five unambiguous members which is too unsubstantial for reasonably computing any Distributional Usefulness scores for them. Therefore, this new confusability analysis was confined to the other seven categories.[88]

The resulting separation values are displayed in Table 4-3 below. Relative to Table 4-2, the separation value increases even further for 28 (67%) of all category pairs included while it decreases for the other 14 pairs. Most notably, the category of interjections displays nearly maximal separation from any other category. The 59 unambiguous interjections thus essentially instantiate a prototypical Clump Scenario — in consequence, the overall Distributional Usefulness score for the interjections among the 633 unambiguous target words is as high as .96. Also remarkable is the observation that most separation values involving the pronoun category leap up substantially. This indicates that the 21 unambiguous pronouns separate very well from most of the other unambiguous categories, and vice versa — in fact, the overall Distributional Usefulness for the pronouns among all unambiguous target words is .51 (whereas it was just .27 for the full pronoun category relative to the entire target lexicon, cf. Figure 4-1). By contrast, the majority of separation values involving the determiner category drop relative to Table 4-2 — and with the exception of the separation between determiners and pronouns, also relative to Table 4-1. This counterintuitive result suggests that the 10 unambiguous determiners are too few or too unrepresentative to form a compact cluster

[88] The unambiguous members of the excluded categories — this involves a total of 7 such target words — were nevertheless included in the reduced target lexicon; and they can thus influence the overall Distributional Usefulness values of the remaining categories.

in the SCO region associated with the full determiner category. Most other separation values change very little.

Table 4-3: *Pairwise separation between categories, excluding all ambiguous members*

Category Γ_1	Category Γ_2										
	INTJ	V	N	ADJ	ADV	INTG	PRON	DET	PREP	CONJ	PTCL
INTJ	—	.98	.98	.98	.98		.99	.99			
V	.57	—	.47	.53	.36		.44	.49			
N	.90	.71	—	.63	.59		.86	.69			
ADJ	.72	.48	.29	—	.37		.75	.26			
ADV	.84	.46	.47	.62	—		.58	.50			
INTG						—					
PRON	.89	.63	.67	.90	.59		—	.72			
DET	.73	.62	.48	.56	.45		.72	—			
PREP									—		
CONJ										—	
PTCL											—

Note. Table cells specify in terms of Distributional Usefulness how well the categories separate from each other in the SCO vector space when each category is reduced to its unambiguous members. The categories INTG, PREP, CONJ, and PTCL contain too few (less than five) unambiguous members to reasonably evaluate their SCO constellations.

Both modified confusability analyses (Table 4-2 and Table 4-3) indicate that categorial ambiguity affects the categories' pairwise separation values and thus their overall Distributional Usefulness scores. But despite this general tendency towards higher separation values, only few of these values come fairly close to their theoretical maximum 1.0. Distributional confusability between categories apparently remains an issue even when ambiguous items are removed. Thus, categorial ambiguity may be one factor contributing to the observed confusability relations but it is not the only one. Indeed, other factors (such as the categories' intrinsic distributional characteristics) appear to be more relevant than categorial ambiguity. Support for this interpretation is offered especially by the first experiment (Table 4-2 vs. Table 4-1) where the overall confusability pattern did not change very much when relevant ambiguous members were excluded. And in the second experiment (Table 4-3 vs. Table 4-2), most of the more substantial changes occurred for the smallest categories (that were still large enough to be analyzed at all) suggesting that random factors might begin to play a bigger role and blur an otherwise fairly robust confusability pattern.

For these reasons, all subsequent analyses were based on the full target lexicon, unless stated otherwise, thus also including the ambiguous category members. Given the goals of the study, this is not entirely satisfactory; but excluding them would be clearly worse. After all, categorial ambiguity is a linguistic fact that the child has to deal with. The preliminary solution offered here is to keep in mind the reported influence of ambiguous target words whenever particular confusability relations are investigated. In the long run, a more suitable solution would be to develop a benchmark coding and an evaluation scheme that do full justice to the issue of categorial ambiguity (cf. section 5.3).

For the time being, it is worth to point out that the observed confusability relations are in many cases linguistically interpretable. For instance, some of them correspond to grammatical classification problems, most prominently with respect to the category of **particles**. There is a controversy among German grammarians whether and how to discriminate particles from prepositions, conjunctions, and adverbs (cf. Helbig, 1994; Hentschel & Weydt, 1994).[89] In Table 4-1, these three categories were all found to separate less well from particles than from most other categories; and particles, in turn, show low separation values from adverbs and prepositions. For the case of conjunctions, their low separation from particles is caused to some extent by categorial ambiguity (cf. Table 4-2 vs. Table 4-1). For prepositions, the contribution of categorial ambiguity cannot be decided from the current data because almost all prepositions (12 of 15) have particle homonyms. Only for adverbs, it is safe to conclude that this category intrinsically shares many distributional properties with particles — adverbs and particles remain extremely confusable even when the categorial ambiguity between them is removed (cf. Table 4-2).

Given that particles and adverbs are distributionally so very similar, these two categories should be easier to discover from distributional information when they are merged into one single category. Corresponding tests confirmed this prediction: When compared to all nonmembers, the merged adverb–particle category achieves an overall Distributional Usefulness of .32 which lies clearly above the corresponding values for adverbs alone (.22) and particles alone (.17). However, this increased score is still fairly low — even for this merged category, distributional information is far from being highly useful. Nevertheless, these analyses suggest to group particles and adverbs (or at

[89] Several other classification problems involving the particle category became apparent in the description of how the benchmark categories were built (cf. 2.2).

least some of their subclasses) together, and it is quite likely that further explorations of distributional patterns would contribute additional evidence to the debate about the scope of the particle category.

This is an example of how co-occurrence statistics might serve as a tool to revise and refine the lexical categories in linguistic theory, or at least in the benchmark classification used here. In another example, the high separation values observed for the category of interrogative words support the decision to introduce it as an independent category. For although it combines words that are typically assigned to three different categories (adverb, pronoun, and determiner), interrogative words separate very well from these categories, and vice versa (cf. Table 4-1 and Table 4-2).

A third example where distributional confusability may be insightful for classification issues is the distinction between pronouns and determiners. As was noted earlier (cf. 2.2.1), some grammarians categorize determiners and pronouns together (e.g., Hentschel and Weydt, 1994). And indeed, both categories separate less well from each other than from most other categories (cf. Table 4-1). However, when all target words that can be used both as a pronoun and as a determiner are discarded, the confusability between these two categories essentially disappears (cf. Table 4-2). Thus, there is actually very little distributional support for defining pronouns and determiners as one category.

These examples illustrate how using co-occurrence statistics as a classification tool is attractive because it provides objective evidence and is sensitive to the statistics of how language is actually used. But caution is clearly advisable, for this approach can only contribute supplementary evidence but not by itself decide any classification issues. For instance, we observed that both particles and adverbs are fairly confusable with pronouns (cf. Table 4-1). However, this is not caused by categorial ambiguity: The ambiguity percentages between these categories are very low in the first place (cf. Table 2-2); and when this ambiguity is removed, the separation values between pronouns on the one hand and particles or adverbs on the other hand change very little (cf. Table 4-2 vs. Table 4-1). Thus, these low degrees of separation must arise from intrinsic distributional properties of these three categories. But there seems to be no independent linguistic motivation for classifying pronouns with particles, and, as far as I am aware of, there are no such proposals, at least not for German. Thus, this appears to be a case where distributional confusability points into a direction that is linguistically not very meaningful.

4.2 Robustness of the information

To posit that children can and do exploit the full amount of distributional information assessed by the default analysis means to make several strong and possibly unrealistic assumptions about the cognitive capacities of the child; for instance, that he pays attention to every single word token in his input — at least for target words — or that he indeed exploits co-occurrence relations with more than 1,000 context words to learn something about the category membership of a particular word. One important question to ask is therefore how robust the distributional information and its usefulness are when these assumptions are met only partly or not at all.

Although all investigated issues of robustness were motivated by considerations about the child — or rather about the learning mechanism underlying category acquisition — the explorations in this section are rather technical in nature as they modify specific aspects of the formal model. The general finding across all experiments will be that gradually reducing the *amount* or *quality* of the extracted distributional information results in a graceful decline or even partial improvement of the *usefulness* of this information with respect to the acquisition of lexical categories.

The section begins with an investigation of the distributional benefit from several assumptions that were made during the decoding stage and that might be disputable (subsection 4.2.1). I then turn to the issue of attention and how the distributional information is affected when the child fails to observe some portion of the available evidence (4.2.2). The next subsection explores the effects of substantially reducing the number of context words, including utterance boundary markers (4.2.3), while the final subsection is concerned with the possibility that the child does not actually exploit exact co-occurrence frequencies but rather only some rough tendencies such as *relatively rare* or *relatively frequent* co-occurrence (4.2.4).

4.2.1 Conservative decoding

The standard procedures that were used to remove transcription codes from the original corpus data involved a few decoding steps that modified the data to a degree that is potentially controversial (cf. subsection 2.1.1). Although these modifications were argued to correspond to realistic assumptions about the child, it appears vital to also study their distributional consequences in order to make sure that they did not substantially ameliorate the derived distributional information. To this end, a second,

more conservative decoding scheme (also described in 2.1.1) was applied to the original input data, avoiding the modifications in question. This alternative decoding scheme should therefore be entirely uncontroversial.

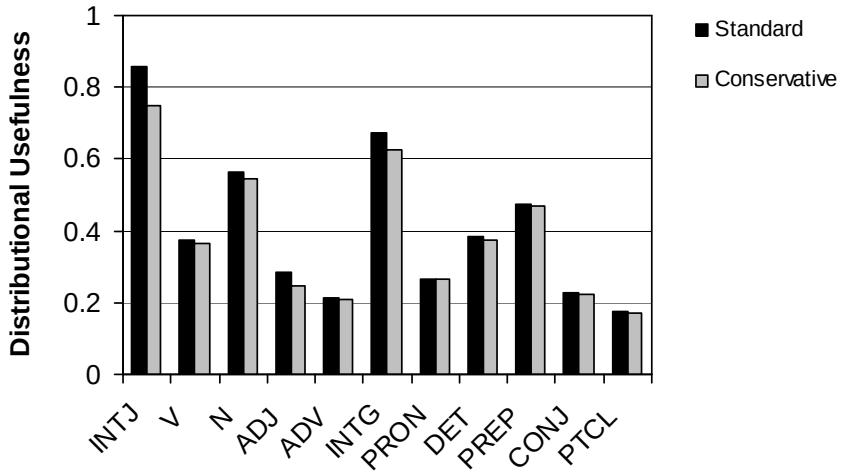

Figure 4-5: *Distributional Usefulness for standard vs. conservative decoding*

Scores for either decoding scheme were obtained by applying the default analysis to the respective decoded corpus.

By applying the default analysis (as described in chapter 3) to the conservatively decoded corpus, one obtains Distributional Usefulness scores for this decoding scheme. Figure 4-5 above shows these scores in comparison to the corresponding scores obtained for standard decoding (copied from Figure 4-1 on p. 94). Across all 11 benchmark categories, Distributional Usefulness is higher for standard decoding than for conservative decoding. However, in most cases, these differences are rather marginal. In fact, there is only one category (interjection) with a more pronounced difference; but it nevertheless clearly remains the category with the highest score such that the overall Distributional Usefulness pattern is not affected. These results indicate that the more relaxed procedures of the standard decoding scheme serve to filter out some portion of statistical noise that is still present after conservative decoding. But the statistical gain is negligible relative to the core information that can be extracted even from the conservatively decoded data.

In addition to Distributional Usefulness levels, there is another perspective from which the distributional consequences of choosing standard over conservative decoding can be studied. Applying the default analysis to both decoded corpora yields two SCO vectors for each target word; and the L_1 distance between these two vectors quantifies

the degree to which the distributional properties of the target word differ between the two decoded corpora. The 1,017 resulting L_1 values are summarized in Figure 4-6.

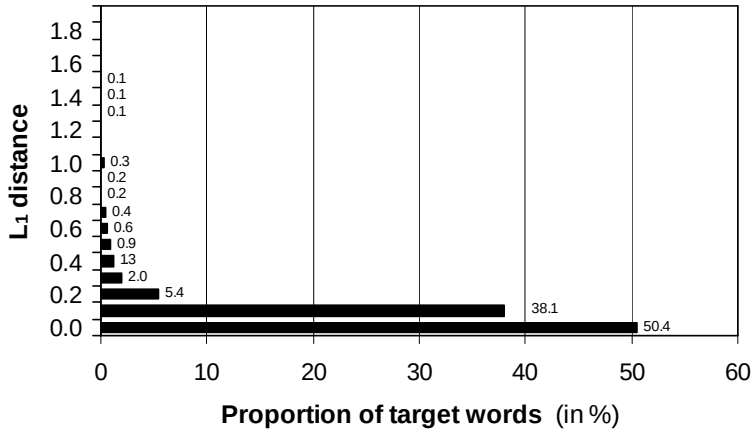

Figure 4-6: *Distribution of word-wise L_1 distances between decoding schemes*

For each target word, the L_1 distance between its two SCO vectors (one derived by standard decoding, the other by conservative decoding) were computed. The L_1 scale in the histogram is partitioned into intervals (which have length .1 and are labeled by their lower bound). Each bar specifies the percentage of all target words whose L_1 value lies within a particular interval.

As it turns out, the two SCO vectors are extremely similar for almost all target words: Their L_1 distance is less than .2 for 900 (88.5%) target words, and for as many as 997 (98.0%), it is below .6. To interpret these figures, pilot work was conducted which suggested that two *different* target words whose SCO vectors have an L_1 distance of less than .6 are as similar as synonyms (cf. Appendix C). But even for synonyms, no single pair of different target words was found to have an L_1 distance below .42.[90] Thus, for virtually all target words, their two SCO vectors are at least as similar as synonyms, and in fact even much more similar for the vast majority. This confirms at the level of SCO vectors that the difference between the two decoding schemes is but statistical noise.

Interestingly, the interjection category has by far the greatest proportion of members with higher L_1 distances between their two SCO vectors. For instance, 59.7% of all interjections have an L_1 value greater than .2, and 18.2% of all interjections even have an L_1 value greater than .6. The distributional consequences of using standard

[90] Although these reference L_1 levels are based on the corpus derived from standard decoding, they also apply to the conservatively decoded corpus as corresponding investigations revealed. In fact, the L_1 range of synonyms appears to be even higher for conservative decoding, and the lowest L_1 distance observed across all 516,636 possible pairs of target words here is as high as .48 (instead of .42 for standard decoding). Thus, if the reference L_1 levels reported in Appendix C were to be corrected at all, they would have to be higher.

rather than conservative decoding thus are the most extensive for the interjection category; and this is consistent with the earlier observations at the level of Distributional Usefulness (cf. Figure 4-5).

Together with the developmental arguments given in subsection 2.1.1, the marginal differences both in terms of Distributional Usefulness and SCO vectors justify using standard rather than conservative decoding. All subsequent analyses are therefore based exclusively on the standard decoding scheme.

4.2.2 Diluting the data

It would be unrealistic to assume that the child attends to every single utterance being uttered in his presence, even if directly addressed to him. The current subsection therefore aims at investigating the potential effects of attention on the distributional information that can be derived from the input. Of course, the *Leo* corpus is only a sample of the full linguistic input encountered by this particular child during the same three-year period — in this sense, the *attention* of the recordings is presumably gappier than that of the child. However, because the corpus captures complete one-hour sessions at full detail, it allows for studying the type and degree of the effects that attention might have *in principle*.

To this end, an experiment was conducted in which the corpus was gradually *diluted*, i.e., scaled down to smaller data samples that nevertheless stretch across the full time period. This reduces the statistical basis for extracting the distributional properties of target words. In order to ensure that this statistical basis is reduced to the same degree for all target words, the dilution procedure was designed to select target word tokens rather than full utterances. Further, to force the dilution to take place uniformly across the entire corpus, the tokens were not selected randomly but rather at a fixed *dilution rate d* (a positive integer) such that for any given target word, every *d*-th of its tokens was selected. This is a reasonable simplification of random token selection unless one has reason to assume a systematic bias of certain distributional properties to occur on, say, the even rather than the odd tokens of a target word.

This procedure thus partitions the original corpus into *d* complementary subcorpora, one starting with the first token of each target word, another one with the

second token, and the last subcorpus with the $d-1$st token.[91] A target word with base frequency f in the full corpus thus has roughly f/d tokens in each of the d subcorpora.[92] Applying the default co-occurrence method (cf. chapter 3) to these subcorpora yields one SCO vector for each target word and each subcorpus.[93]

Intuitively, one would expect the SCO vector of a particular target word to become more dependent on random factors — and thus generally less representative for its genuine distributional properties in the full input — as the statistical basis is reduced. To the extent that these random factors are unlikely to be informative about lexical categories, the prediction therefore is that Distributional Usefulness declines when the dilution rate increases. To test this prediction, a given category's Distributional Usefulness score was computed independently for each subcorpus resulting from the dilution procedure. The d scores for the same category were averaged to obtain a single score for this category and the given dilution rate d.

Figure 4-7 below shows the resulting average Distributional Usefulness scores for dilution rates 2, 4, and 10, in comparison to those for the undiluted original corpus (corresponding to dilution rate $d=1$). As predicted, the scores consistently decrease for each category as the dilution rate is gradually raised. However, scores do not drop sharply; rather, the overall pattern is best described as a *graceful decline* of Distributional Usefulness. Even for dilution rate $d=10$ — when the local contexts of a target word are recorded for only every 10th of its tokens — distributional information is in most cases still not dramatically less useful than it is for the full corpus.

---

[91] Technically, the resulting subcorpora are no real corpora in the sense of collections of utterances — instead, they are sets of target word tokens and their local contexts. To keep terminology simple, I nevertheless refer to these sets as *subcorpora*.

[92] Unless this ratio happens to be an integer, the actual token counts for the same target word may differ by ±1 between the different subcorpora.

[93] Note that this experiment is formally related to the second experiment in 4.1.1 in which the target words' base frequencies were reduced as well. However, there the reduction was not at all the point of the experiment but rather performed as the only way to obtain a corpus in which all target words have the same base frequency. In consequence, the reduction occurred at higher rates for the more frequent target words. The current experiment, by contrast, was intended to explore the effects of data reduction itself; and for this reason it applied the same rate to all target words, retaining their relative frequencies. These differences between the two experiments can explain their seemingly inconsistent results (cf. footnote 84, p. 97).

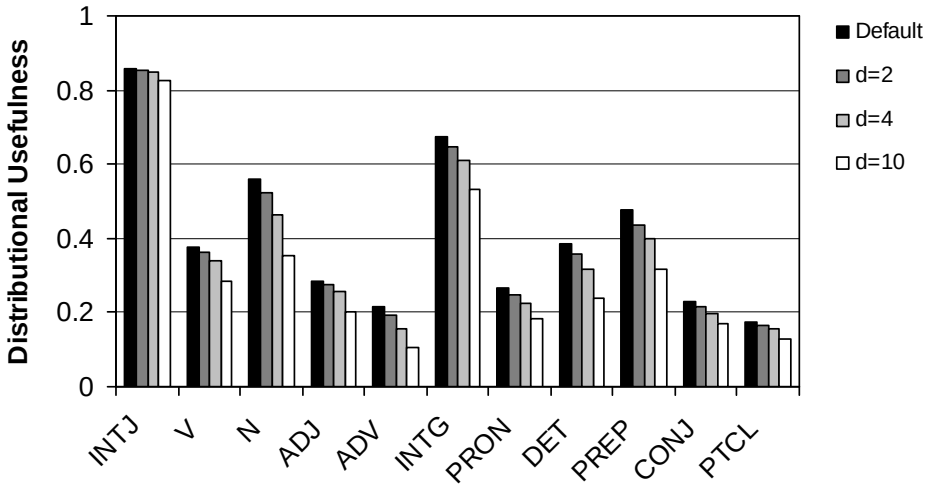

Figure 4-7: *Diluting the corpus and the effects on Distributional Usefulness*

Average Distributional Usefulness for three dilution rates $d=2$; 4; 10 (each value resulting as the mean score across the d complementary subcorpora), and, for comparison, the corresponding scores of the default analysis on the original corpus (copied from Figure 4-1). Standard deviation figures (ranging from .0003 to .03) are generally too small to be displayed.

As in the previous subsection, one can study the distributional consequences of the current experiment also more directly. By computing the L_1 distance between a target word's different SCO vectors that are derived from the various subcorpora (for the same dilution rate), one obtains a measure of how robust this word's distributional properties are across these subcorpora, i.e., a measure of how much its distributional properties depend on the choice of the data sample. For dilution rate d, there are d different SCO vectors for the same target word and therefore $d(d-1)/2$ pairs of SCO vectors for which their L_1 distance can be computed. The distribution of these L_1 values for all 1,017 target words is plotted in Figure 4-8 below for dilution rates 2, 4, and 10.

Obviously, the entire distribution of L_1 distances rises as the dilution rate increases. This confirms the earlier prediction that distributional information becomes less robust as its statistical basis is gradually reduced. This prediction can also be verified for a fixed dilution rate d where it translates to the prediction that distributional properties tend to be more robust for the more frequent target words. Figure 4-9 (p. 116) plots each target word's L_1 distance (for dilution rate $d=2$) against its base frequency (in the original, undiluted corpus). Corresponding graphs for higher dilution rates would show essentially the same pattern, only shifted towards greater L_1 values.

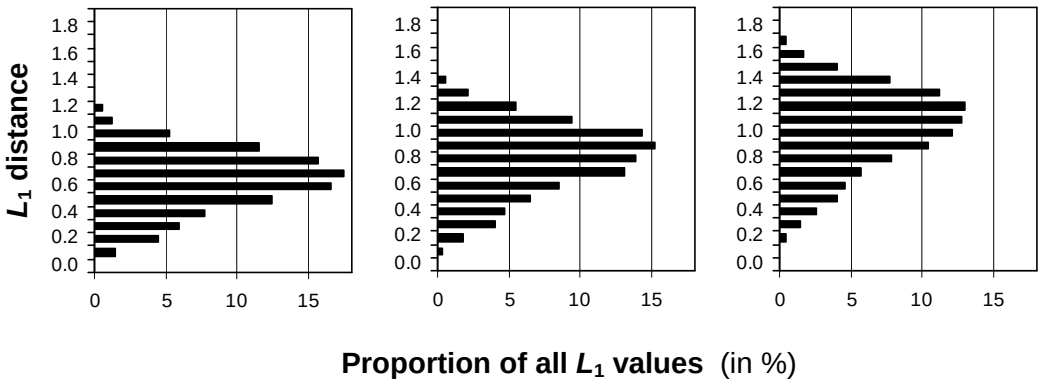

Figure 4-8: *Distribution of word-wise L_1 distances between diluted subcorpora*

For each dilution rate *d* and each target word, the SCO vectors of this target word are compared between any two of the *d* diluted subcorpora. The charts show the distribution of the resulting 1,017 L_1 distance values for $d=2$ (left), of the 6,102 L_1 values for $d=4$ (center), and of the 45,765 L_1 values for $d=10$ (right). In each histogram, the L_1 scale is partitioned into intervals (which have length .1 and are labeled by their lower bound). Each bar specifies the percentage of all L_1 values within a particular interval.

There clearly is a strong correlation between a target word's L_1 distance and the logarithm of its base frequency; and within the given frequency range, this relation is roughly linear (Pearson's $r = -.813$, $p < .0001$).[94] But beyond the mere correlation, the graph suggests that there is an upper bound of possible L_1 values and that this upper bound is a descending function of base frequency. There is no comparable lower bound — even for infrequent words, it is possible, though not likely, to have fairly robust distributional properties. These findings imply that if the frequency threshold for selecting target words is raised significantly, overall robustness would be boosted. For instance, with a frequency threshold of 1,000 word tokens, the L_1 distance would be below .6 for almost all (97.0%) of the then selected 169 target words — their two SCO vectors (for dilution rate $d=2$) can thus be regarded as being as similar as the those of synonyms would be (cf. Appendix C).

For the full target lexicon (with frequency threshold 100), by contrast, only 32.0% of all L_1 distances are below .6.[95] Thus, even though the L_1 values for $d=2$ are clearly

[94] Base frequencies of target words range from 100 to 38,518. If less frequent words were also included and if there were any words with base frequency above this range, the correlation would best be described by some hyperbolic function.

[95] The reference level $L_1 = .6$ was formulated for the full corpus. However, because the overall L_1 distribution between SCO vectors of different words rises as well when the data are diluted, the appropriate reference level for synonyms would have to be based on the two subcorpora for dilution rate $d=2$ (see the general remarks in Appendix C). However, doing so raises this reference level merely to $L_1 = .65$; and even with this corrected reference level, still only 39.7% of all L_1 distances

lower than those for any greater dilution rate (cf. Figure 4-8), they are generally not as low as might have been expected (for instance, compare the left histogram in Figure 4-8 with Figure 4-6 on p. 111). This indicates that the distributional properties of a target word, as represented by its SCO vector, depend to a nonmarginal degree on the choice of a particular subsample of the data. However, this does not appear to affect the usefulness of this information very much as is evidenced by the graceful decline of Distributional Usefulness (cf. Figure 4-7 on p. 114).

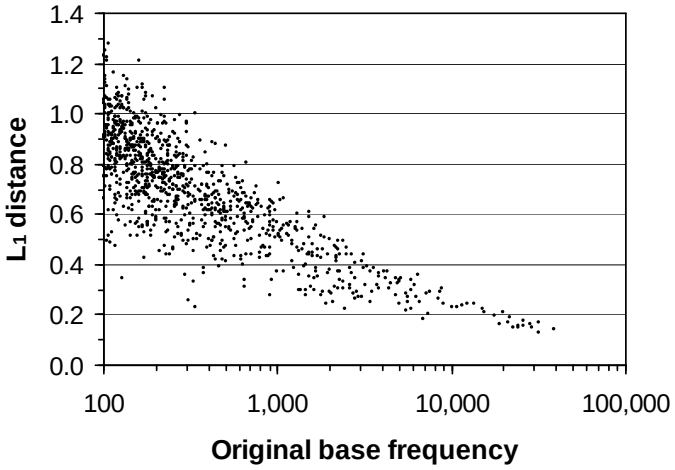

Figure 4-9: *Influence of base frequency on robustness*

Each data point represents one target word, the coordinates being the word's base frequency in the full corpus (plotted on a logarithmic scale) and the L_1 distance between its two SCO vectors (derived from the two subcorpora for dilution rate $d=2$).

The presumable reason is that even though the two SCO vectors of the same target word are not always very similar to each other, they are still clearly more similar to each other than to the SCO vectors of most other target words: 93.0% of all L_1 distances between SCO vectors of the same target word are below 1.0 whereas 93.9% of all L_1 distances between SCO vectors of different target words are greater than 1.3 (all numbers for dilution rate $d=2$). Thus, for most target words there appears to be some region in the SCO vector space such that its SCO vector derived from a particular diluted subcorpus is likely to lie within this region. These regions can be fuzzy and even fairly large, but they generally overlap with only very few regions of other words.

between the two SCO vectors of the same target word can be considered as being as low as the L_1 distance between synonyms.

Within this picture, the effect of raising the dilution rate may be that these word-specific regions become larger and possibly also fuzzier.

In sum, the effects of diluting the data are a consistent but graceful decline of Distributional Usefulness and an overall decrease of robustness of the target words' distributional properties. Extrapolating these results beyond the corpus — which can be regarded as a diluted sample of the child's full input during the same period — it is very likely that the distributional information in this full input would be even more robust and more informative than that in the corpus, at least when the analysis is carried out on the same target lexicon.

By a very conservative estimate, the *Leo* corpus captures between 5% and 10% of the full input during that same three-year period (a higher percentage during the first year, and lower rates during the other two years). Correspondingly, this full input would have comprised a total of at least 13 million word tokens which is also in line with conservative estimates by Hart and Risley (1995:132). The extrapolative prediction — namely, that Distributional Usefulness is generally greater for the full input than for the input sample captured by the *Leo* corpus — receives additional support by Redington et al. (1998:456) who found their own scores to gradually increase as the corpus size was raised step by step, up to their maximal corpus which had roughly twice the size of the *Leo* corpus.

This does not imply, however, that Distributional Usefulness could be pushed up to the maximal value of 1.0 for all categories if only the corpus were made large enough. Eventually, one is most likely to run into a *ceiling effect* for each category; that is, there is an upper limit (below 1.0) for the category's Distributional Usefulness score that can be approximated but not overcome by increasing the corpus size. Such a ceiling effect already seems to be forming in Figure 4-7 (p. 114) for the category of interjections. In other words: The corpus may still be too small to demonstrate the full potential of distributional regularities, but corpus size is not the main reason for the fact that we do not observe a Clump Scenario for each of the 11 benchmark categories. It rather seems that the *kind* of distributional information assessed here does not suffice to fully predict these categories (cf. discussion in section 5.3).

4.2.3 Reducing the context lexicon

In the default analysis, a target word's distributional properties are defined in terms of its co-occurrences with 1,017 context words and four virtual context words marking

utterance boundaries (cf. 3.1.1). The current subsection explores the distributional consequences of reducing the set of (lexical and virtual) context words. This manipulation constitutes another way of reducing the statistical basis of distributional information: Whereas in the preceding section, the statistical basis was reduced by selecting some subsample of the corpus data and holding the context lexicon constant, it is now the data sample that is held constant while the context lexicon is gradually reduced. In a first experiment, only the set of lexical context words is reduced while another experiment step by step removes the distributional information from utterance boundaries.[96]

Intuitively, the highly frequent context words are more likely to provide useful cues about the target words' category membership than are less frequent context words.[97] Therefore, in order to retain as much information as possible, the context lexicon is reduced by raising the frequency threshold for selecting context words. Figure 4-10 shows the resulting Distributional Usefulness scores when the default threshold of 100 tokens (selecting all 1,017 context words) is raised to 1,230 tokens (leaving only 150 context words), then further to 4,377 tokens (50 context words), and finally to 22,474 tokens (10 context words).[98] Note that each of these context lexica also still comprises the four utterance boundary markers in addition to the selected context words.

The general pattern is a graceful decline of Distributional Usefulness for most categories as the context lexicon is reduced in this fashion. However, the categories of determiners, prepositions, and pronouns show a temporary, though minor, increase in Distributional Usefulness before their scores finally drop sharply at 10 context words. Notably, for the category of interrogative words, Distributional Usefulness consistently improves all the way from 1,017 to 10 context words. This suggests that co-occurrences with less frequent context words are not very informative about this category such that they blur the very useful cues from co-occurrences with highly frequent context words

---

[96] In the two previous subsections, robustness was analyzed both at the level of Distributional Usefulness and at the level of SCO vectors. In the current subsection, however, the latter type of analysis is not applicable because the experiments modify the number of context dimensions.

[97] This intuition was confirmed in pilot work where three different ways of defining the usefulness of individual context words were all found to correlate highly with the context words' base frequency. It also corresponds to the effects of base frequency that were observed in the preceding section with respect to target words (in particular, Figure 4-9 on p. 116).

[98] These thresholds were chosen because Redington et al. (1998:453f) found distributional information to be most useful in general when the context lexicon consisted of the *N* most frequent word forms, with *N* between 50 and 150 words. With only 10 context words, performance was rather poor. And when the size of the context lexicon was raised above 150 words, performance was found to decline gracefully. However, my own explorations (in subsection 3.2.2) indicate that this general decline for large context lexica is a characteristic property of the similarity measure used by Redington et al. (viz., rank correlation) rather than reflecting a genuine decline in usefulness of the extracted information.

and with utterance boundary markers. Overall, one can say about all categories that reducing the context lexicon from the 1,017 to the 50 most frequent words has only marginal effects on Distributional Usefulness.

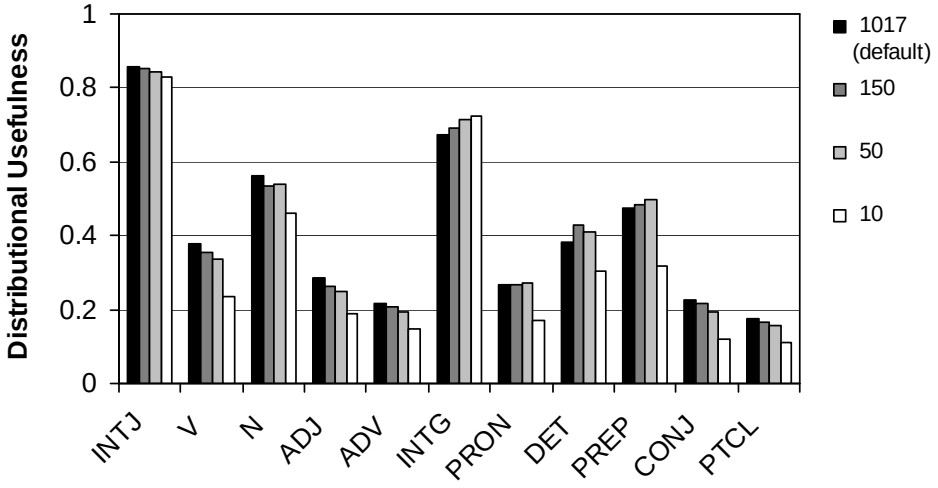

Figure 4-10: *Reducing the number of lexical context words*

Distributional Usefulness for the default context lexicon and when the context lexicon is reduced to the 150, 50, or 10 most frequent words. The four utterance boundary markers are included in each of the four context lexica.

Utterance boundaries can provide informative cues about lexical categories in three complementary ways, all of which are present in the default analysis (and throughout the above experiment). First, even distinguishing utterances can provide implicit cues because this presumably removes statistical noise in the form of lexical co-occurrences across utterance boundaries (cf. pp. 59f). Second, explicitly representing utterance boundaries as virtual context words (i.e., the utterance boundary markers) and recording a target word's co-occurrences with these markers provides the model with information about this word's probability to occur in particular serial positions of an utterance (first, second, last but one, last). And finally, for the two utterance-final positions, this information is further supplemented by distinguishing whether an utterance terminates on an intonation signaling a question, an exclamation, or a declaration — this information is represented by the three different post-utterance markers (cf. p. 60).

It should be reiterated that all three types of information from utterance boundaries are likely accessible by the child, most of all the implicit cues (cf. 2.1 and 3.1.1). In fact, the converse assumption that the child considers co-occurrences across utterance boundaries in any systematic fashion seems highly unrealistic, given that utterances tend

to be clearly delimited by pauses or by change of turn between speakers (including the child himself). Nonetheless, for the general issue of robustness, it is important to determine how much the usefulness of distributional information depends on cues from utterance boundaries.

Because the three types of information build on each other, their distributional benefit was assessed by removing them from the default analysis in reverse order. The resulting Distributional Usefulness scores are shown in Figure 4-11. Starting from the default analysis, the information about the type of utterance termination was removed by collapsing the three post-utterance markers to one single symbol, leaving a total of only two utterance boundary markers. As it turned out, this has little effect in general; only for the categories of interrogative words and interjections, Distributional Usefulness decreases to a nonmarginal degree, though not dramatically.[99]

Next, information about serial position was removed by ignoring co-occurrences with any utterance boundary marker such that the only information left from utterance boundaries are implicit cues. This removal results in a dramatic drop of Distributional Usefulness for interjections — though it still remains higher than for most other categories — and in smaller declines for interrogative words, nouns, and determiners.[100] The seven remaining categories even benefit from the removal, suggesting that cues from serial position are by far less informative about these categories than are cues from lexical co-occurrence.[101]

In the third step, even the implicit cues from utterance boundaries were discarded by concatenating all utterances (in chronological order) into a single, extremely long utterance, and thus recording co-occurrences across the original utterance boundaries. Distributional Usefulness decreases a bit for 10 categories, with the greatest drop

[99] The reason for this finding is that interjections and interrogative words share some of their most salient distributional properties but can be distinguished fairly well by the empirical fact that, among a few other differences, most utterances with an interrogative word form in the last or last but one position are questions and thus terminate on a question mark whereas the vast majority of utterances with an interjection in these serial positions are declaratives or exclamations and thus terminate on a period or exclamation mark (cf. 4.3.2).

[100] Most interjections prefer to occur in very short utterances — for instance, on average, 33.9% of all tokens of an interjection constitute one-word utterances and thus have no lexical context whatsoever. But even in longer utterances, the vast majority of interjection tokens occur close to either utterance boundary (cf. 4.3.2). This means that removing information about serial position removes a very large portion of the original statistical basis for the interjection category.

[101] In a corresponding experiment, Redington et al. (1998:457f) found virtually no effect of removing serial information. However, they only investigated this manipulation for the category system as a whole. Therefore, their results are actually consistent with the differential effects that my own analyses revealed at the category-specific level — the negative effect for interjections, interrogative words, nouns, and determiners might neutralize the positive effect for the other seven categories.

occurring for interrogative words.[102] Only the interjection category actually benefits from the removal, and it can partly recover from its initial loss.[103]

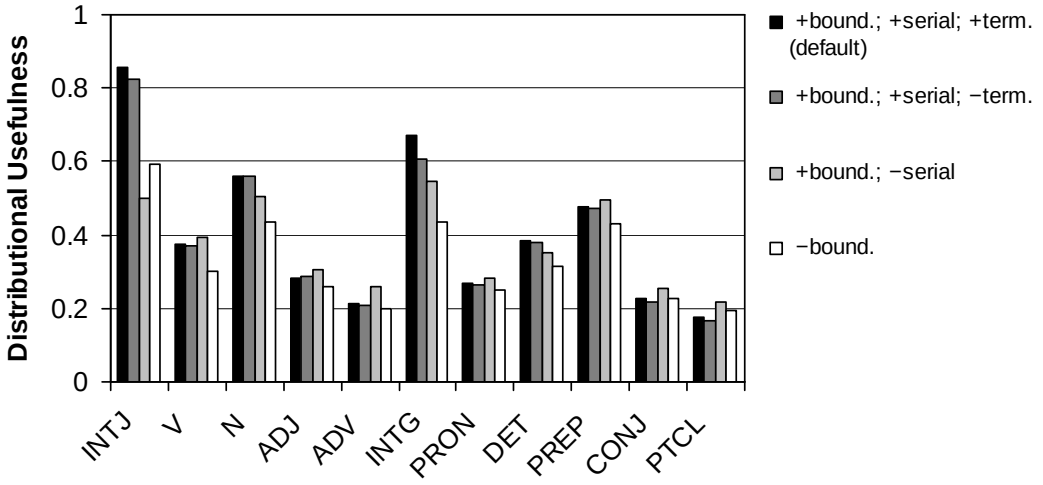

Figure 4-11: *Removing cues from utterance boundaries*

Distributional Usefulness when explicit information about serial position is recorded and type of utterance termination is distinguished (default analysis), when explicit information about serial position is recorded but the type of utterance termination is ignored (+bound.; +serial; −term.), when also explicit information about serial position is removed (+bound.; −serial), and when even implicit cues from utterance boundaries are removed by ignoring utterance boundaries altogether and treating all input as one single utterance (−bound.). All four analyses recorded co-occurrences with all 1,017 lexical context words.

Overall, with the exception of interjections and interrogative words, gradually removing cues from utterance boundaries results in a graceful decline — or even in a slight increase — of Distributional Usefulness. The severe effects on interjections and interrogative words indicate that these two categories depend on such cues more than any other category; and this interpretation is consistent with the earlier observation that interjections and interrogative words are least affected — or even benefit — when the context lexicon is reduced to the 10 most frequent word types, as long as the full information from utterance boundaries is retained (cf. Figure 4-10 on p. 119).

In order to directly compare for each category the relevance of cues from lexical co-occurrence (i.e., co-occurrence with the 1,017 context words) with that of cues from

[102] This general, slightly negative effect of removing implicit cues from utterance boundaries is consistent with findings by Redington et al. (1998:457f) at the level of the overall category system.

[103] The most likely explanation for this partial recovery concerns the fact that recording co-occurrences across utterance boundaries provides the many interjection tokens occurring in one-word utterances with a lexical context. It appears that cross-utterance co-occurrences are not statistical noise altogether but potentially informative as well. However, in general this does not reflect any syntactic relations but rather certain distributional properties of the *context words*, namely that the various words are not equally likely to occupy any of the two first or last positions of an utterance.

utterance boundaries, Figure 4-12 displays Distributional Usefulness scores for the default analysis (where both types of information are fully included) and for both types of information in isolation.

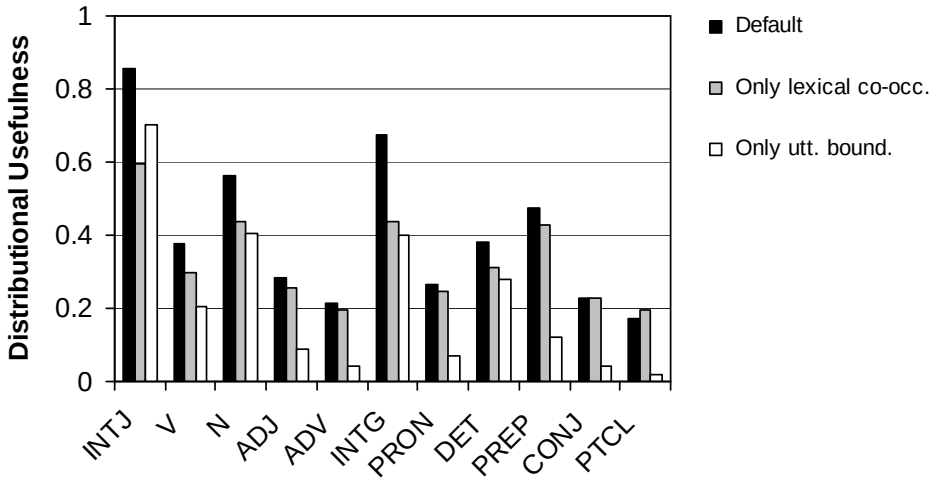

Figure 4-12: *Interacting information from lexical co-occurrence and utterance boundaries*

Effects on Distributional Usefulness when all information from utterance boundaries is discarded while lexical co-occurrences with the full context lexicon are retained (copied from Figure 4-11), or, conversely, when any lexical co-occurrence information is removed while all cues from utterance boundaries are retained.

Except for conjunctions and particles, all categories receive *some* useful cues from both types of information because they achieve the highest scores when both types are included (that is, for the default analysis). Conjunctions and particles, by contrast, depend entirely on cues from lexical co-occurrence while cues from utterance boundaries even have negative effects on Distributional Usefulness: When only lexical co-occurrence is considered, both categories achieve even slightly better scores than for the default analysis; and when only utterance boundaries are considered, these categories achieve scores that are hardly better than chance (for the distribution of random Distributional Usefulness values, see pp. 86f). Five other categories (adverb, pronoun, adjective, preposition, and to a lesser degree also the verb category) also chiefly rely on lexical co-occurrence information although they do perform best when both types of information are present.

For interrogative words, nouns, and determiners, both types of information are about equally informative, with but a slight advantage of lexical co-occurrence. And finally, the interjection category is the only one for which cues from utterance boundaries are more informative than cues from lexical co-occurrence. Nevertheless,

even lexical co-occurrence information alone is still very useful for interjections — and in fact more useful than it is for any other category.

To conclude this subsection, maybe the most impressive finding was that distributional information is so robust that reducing the context lexicon to the 50 most frequent words, while retaining the full information from utterance boundaries, has hardly any effect on Distributional Usefulness (cf. Figure 4-10 on p. 119). Even with as few as 10 context words, Distributional Usefulness is still remarkable for most categories (ibid.). Further, even when all context words are left out, a few categories still receive very useful cues from utterance boundaries alone (cf. Figure 4-12).

4.2.4 Sloppy counting of co-occurrences

If the child indeed exploits the full information from SCO vectors, he would have to be sensitive to differences in relative frequency of co-occurrence. However, without making any specific assumptions about the nature of the underlying learning mechanism, it is a realistic possibility that the child merely exploits very rough tendencies of such frequencies. That is, the learning mechanism might note which co-occurrence events were observed relatively often, without being responsive to the exact number of observations. Similarly, the learning mechanism might notice which co-occurrence events were encountered at all — provided that they are not too rare —, without paying attention to just how frequently they occur.

To formalize this idea, the exact original (i.e., nonstandardized) co-occurrence counts were transformed into *sloppy counts* which simply distinguish three frequency classes: *rare/missing*, *occasional*, and *frequent*.[104] If a particular target word was observed for less than 1% of its tokens to co-occur with a particular context word in a particular context position, the relation between these two words in this context position was classified as a *rare/missing co-occurrence event*.[105] Correspondingly, if a co-occurrence relation was observed for at least 1% but less than 5% of all tokens of the

[104] Alternatively, one might use more than three frequency classes but the goal was to reduce the distributional information as far as possible. On the other hand, using only two such classes would seem to overdo the reduction because it corresponds to a binary decision about whether or not something is possible — *frequency*, the core notion of any usage-based work, would thus effectively be removed from the model. Therefore, even the most basic distributional model should distinguish at least three different frequency levels.

[105] In formal terms, dividing an observed co-occurrence count by the target word's base frequency (i.e., by its total number of tokens) yields the corresponding *relative co-occurrence frequency* which can be interpreted as an estimate for the probability of the corresponding co-occurrence event. Thus, a co-occurrence event is considered rare/missing if its estimated probability is below 1%.

given target word, it was rated an *occasional co-occurrence event*. And finally, a co-occurrence relation observed for at least 5% of all instances of the given target word was considered a *frequent co-occurrence event*.[106]

In this fashion, the original co-occurrence vector of each target word was rewritten as a vector of sloppy counts, with rare/missing co-occurrence events represented by the integer 0, occasional co-occurrence events by the integer 1, and frequent co-occurrence events by the integer 2. This rewritten vector thus is a sloppy summary of the distributional properties of the corresponding target word. To compare these *sloppy vectors* between different target words, it is reasonable to apply the L_1 distance as usual.[107] Thus, when comparing two target words in a particular context dimension, there are only three possible cases: Either the two words do not differ at all (when they are both in the same frequency class such that this context dimensions contributes the value 0 to their overall distance), or they differ somewhat (when one word is in the *occasional* class and the other in the *rare/missing* or *frequent* class, thus contributing the value 1), or they differ greatly (when one word is in the *rare/missing*, and the other in the *frequent* class, thus contributing the value 2).

This illustrates that using sloppy vectors instead of SCO vectors removes a lot of distributional information. At the level of individual context dimensions, two SCO vectors can differ at any conceivable real-valued degree whereas sloppy vectors can only differ *greatly*, *somewhat*, or *not at all*. Mapping the real-valued differences onto three discrete differences might in some cases remove important cues to category membership, while in other cases it might simply level out differences generated by random factors. The first type of cases would presumably affect Distributional Usefulness while the second type would actually improve it. Therefore, it was unclear a priori whether the reduction of information would harm or improve the overall usefulness of this information.[108]

---

[106] In pilot work, I also experimented with other frequency ranges. The particular percentages 1% and 5% were chosen because they yield, for most target words, a reasonable number of *occasional co-occurrences* (median across all target words: 46 context dimensions) and also a reasonable number of *frequent co-occurrences* (median: 10) while the bulk of context dimensions (median: 4,109) are *rare/missing co-occurrences* (also see footnote 110, p. 126).

[107] The resulting distance values could be standardized to the range [0;1] by dividing them by $2N$ where $N = 1{,}017$ is the total number of context dimensions. However, this would not alter the resulting similarity structure among the sloppy vectors.

[108] As in the preceding subsection, I here assess robustness only in terms of Distributional Usefulness. Evaluating the robustness of vectors (as in subsections 4.2.1 and 4.2.2) would not be meaningful because SCO vectors and sloppy vectors reside in different vector spaces.

As it turns out, Distributional Usefulness drops dramatically for all categories except interjections (Figure 4-13, the two left-most bars for each category). At first this seems to imply that sloppy vectors are in principle much less informative about lexical categories than are SCO vectors. The real-valued differences between target words in particular context dimensions appear to provide some crucial cues that are lost when moving to sloppy counting. However, this conclusion may be premature as there is another potential explanation for the substantial drop.

For a highly frequent target word that occurs, say, at least 10,000 times in the corpus, no fewer than 100 co-occurrences need to be observed with a particular context word in a particular context position for this co-occurrence relation to exceed the 1% threshold and thus to be rated at least an occasional co-occurrence event. But almost half of all 1,017 context words occur less than 200 times and are unlikely to co-occur 100 times with this single target word (in the given context position). By contrast, an infrequent target word can potentially show occasional or frequent co-occurrence events with any context word, including the infrequent context words. The most extreme case is a target word with only 100 tokens in the corpus; and for such a word, a single co-occurrence observation suffices for reaching the 1% threshold and thus being rated at least an occasional co-occurrence event. This demonstrates that within the *sloppy counting* paradigm, co-occurrences with infrequent context words are of little help or even obstructive for discovering lexical categories, because they introduce — as a statistical artifact — distributional differences between highly frequent and less frequent target words, even when the distributional preferences of these words may be almost identical.

One way of removing this artifact is to restrict the context lexicon to highly frequent words as in the preceding subsection.[109] Two additional sets of sloppy vectors were computed again, this time using only the 50 or 10 most frequent words as context words (occurring at least 4,377 or 22,474 times, respectively), and still retaining utterance boundary markers as virtual context words. Because for some target words, this left hardly any occasional co-occurrence events, the boundary between rare/missing

[109] Alternatively, one could reduce the percentages by which occasional and frequent co-occurrence events were defined. But this would result in a general explosion of the number of these events and mask information from the most robust co-occurrence patterns that are instantiated very frequently (cf. footnote 106, p. 124). Another alternative would be to increase the number of frequency classes to be distinguished. But the general goal of investigating issues of robustness is to see how far down one can push the extracted information while keeping its usefulness as high as possible (cf. footnote 104, p. 123).

and occasional events was lowered from 1.0% to 0.5%.[110] The resulting Distributional Usefulness scores are also shown in Figure 4-13.

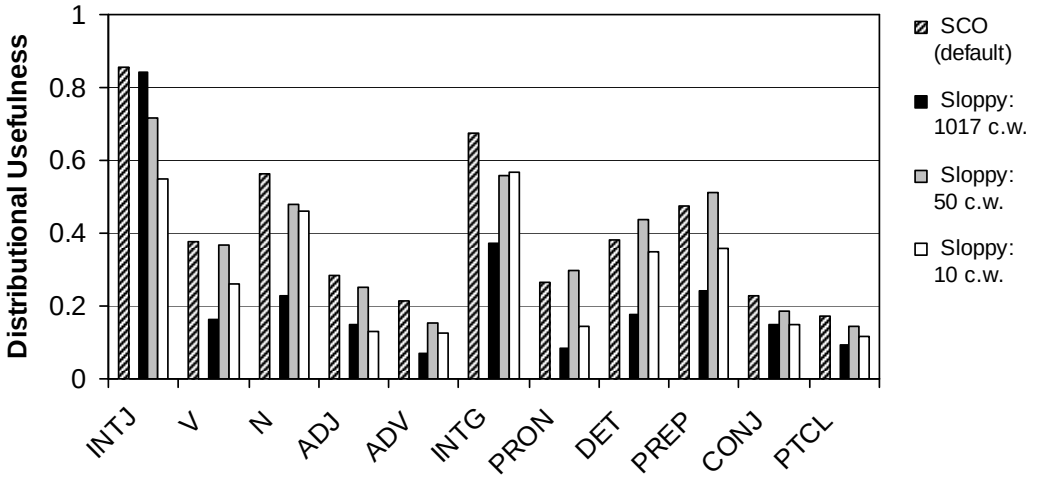

Figure 4-13: *Sloppy counting and the effects on Distributional Usefulness*

Distributional Usefulness for the default analysis (based on SCO vectors) and for sloppy vectors (using the 1,017, 50, or 10 most frequent words as context words). The boundary between frequent and occasional co-occurrence events was 5.0% (for all three analyses with sloppy counts) while the boundary between occasional and rare/missing co-occurrence events was 1.0% (for 1,107 context words) or 0.5% (for 50 and 10 context words). The four utterance boundary markers were included in each condition.

With the exception of interjections, Distributional Usefulness increases substantially when sloppy vectors are computed for the 50 most frequent context words rather than for the full context lexicon. The scores generally compensate for most of their earlier losses (between the default analysis and the sloppy vectors for the full context lexicon); in fact, for three categories (pronoun, determiner, and preposition) this earlier loss is even clearly overcompensated for. This is a remarkable finding because *both* steps (reducing the size of the context lexicon and moving from real-valued SCO vectors to discrete sloppy vectors) constitute a reduction of the extracted information. It appears that both types of reduction interact in a way that they can efficiently remove some noise while retaining the crucial distributional information.

When reducing the context lexicon further to only the 10 most frequent words (apart from utterance boundary markers), Distributional Usefulness declines again but

[110] In general, best results in terms of Distributional Usefulness were obtained when the median number of *frequent co-occurrences* across target words was between five and 10, and no smaller than two for any target word; and when simultaneously, the median number of *occasional co-occurrences* was between 20 and 60, and no smaller than four for any target word (also see footnote 106, p. 124).

in most cases remains clearly greater than for the sloppy vectors computed on all 1,017 context words.[111] Interestingly, when the context lexicon is held constant at the 50 (or the 10) most frequent words, Distributional Usefulness is roughly as high for sloppy vectors as for SCO vectors, for all categories except for interjections and interrogative words (compare Figure 4-13 with Figure 4-10 on p. 119).

In sum, the general observation is essentially the same that was repeatedly made throughout this section on robustness, namely that the usefulness of distributional information declines gracefully (if at all) as the amount or quality of this information is gradually reduced. The only sudden drops in Distributional Usefulness occurred for interjections (when removing serial information, cf. 4.2.3) or as a side-effect of statistical artifacts (when moving from SCO vector to sloppy vectors on the full context lexicon, cf. the current subsection).

Finally, it should be stressed that the concepts of exact and sloppy counting of co-occurrences are used here only to distinguish two different kinds of distributional information; but they are not intended to insinuate that a distributional learning mechanism would actually have to engage in any counting, be it exact or sloppy. In fact, if the child indeed exploits patterns of lexical co-occurrence and serial position, he is likely to use a learning mechanism that differs fundamentally from the formal co-occurrence model presented here. In particular, this mechanism presumably does not operate with representations such as SCO vectors or sloppy vectors. Nevertheless, such vectors can be interpreted as roughly summarizing the child's distributional experience with the respective target words (for further discussion, see section 5.2).

[111] This overall pattern of Distributional Usefulness is reminiscent of the sensitivity curve that was observed when testing rank correlation as a possible measure of similarity between vectors (cf. 3.2.2). In both cases, the overall usefulness of the information increases as the frequency threshold for selecting context words is increased until the context lexicon becomes too small which is somewhere between 50 and 10 context words. These corresponding patterns are no coincidence — transforming an SCO vector to the rank order of its vector elements is closely related to computing a sloppy vector. In both cases, the basic idea is to map exact frequencies onto a set of consecutive integers. This mapping may differ considerably between the two concepts (e.g., involving very different numbers of integers), but both concepts nevertheless perform poorly on infrequent context words for very similar reasons; namely that these infrequent context words increase the overall proportion of co-occurrence counts close to zero, and that the distribution of these small co-occurrence counts is highly susceptible to random factors. This adds random noise to the distributional similarity between target words. The difference, however, is that for sloppy vectors, this does not concern the similarity between highly frequent target words.

4.3 Exploring the information

While the focus of the previous two sections was the usefulness and robustness of the extracted distributional information, the current section aims at uncovering in detail what this information actually consists of. This issue is approached by investigating for each category the questions *Where can the informative distributional cues to this category be found?* (subsection 4.3.1) and *What exactly are these informative cues?* (subsections 4.3.2 and 4.3.3). While the first question only asks about the location (i.e., context position) of the most informative cues, the second question asks more specifically about the particular co-occurrence relations (i.e., context words in a context position) that give rise to these cues. In particular, this second question will shed some light on the confusability relations between different categories that were identified in 4.1.2, and likewise on the overall Distributional Usefulness score obtained for each category. These observations will in turn lead to some general insights about the role and relation of two basic types of cues (subsection 4.3.4). In contrast to the section about robustness, the analyses in the current and all subsequent sections are based on the full amount of distributional information extracted by the default analysis, unless stated otherwise. The distributional cues identified in the current section are in many cases traced back to the typical underlying linguistic constructions observed in the corpus, and occasionally illustrated by examples which are all taken from the corpus.

4.3.1 Location of informative cues

Which of the four context positions provide the most informative cues to a particular benchmark category? Without prior knowledge of these cues, this question can best be addressed by applying the general co-occurrence method to each context position separately. This amounts to simply skipping the concatenation of the four individual vectors computed for each target word (cf. 3.1.1), and to use each of these vectors as the co-occurrence vector of the given word for the corresponding context position. Taking the 1,017 vectors for one particular context position, the Distributional Usefulness score for a given category then quantifies how useful the cues in this context position are with respect to this category. The resulting scores are displayed in Figure 4-14.

For many categories, Distributional Usefulness varies considerably across the four context positions. The most uneven patterns can be observed for **interjections** and **interrogative words** — coincidentally the categories with the greatest overall

Distributional Usefulness scores — which both receive by far the most and least useful information from cues in the positions [−1] and [−2], respectively. **Verbs** receive the most useful information from left context (i.e., from the positions [−2] and [−1]) while **nouns** rely mostly on immediate context (i.e., on the positions [−1] and [+1]).

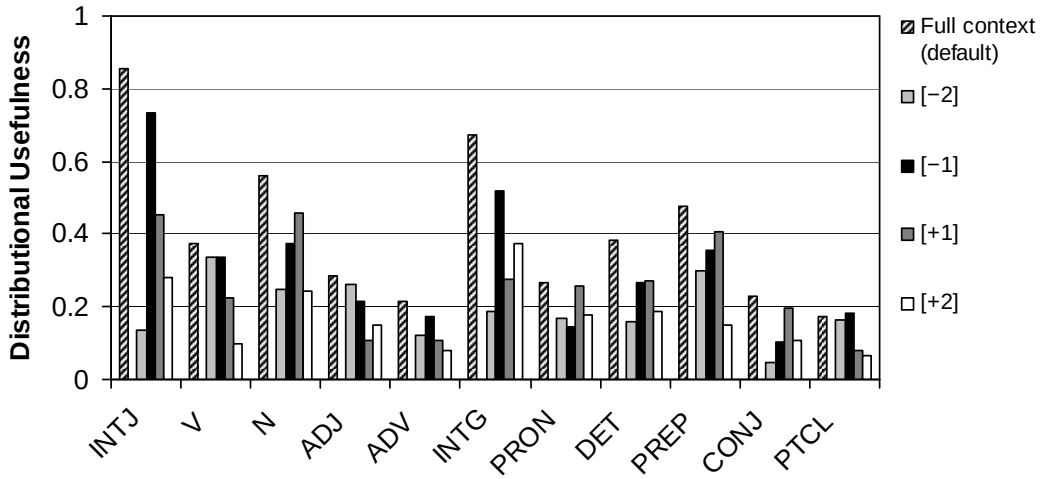

Figure 4-14: *Distributional Usefulness by context position*

Distributional Usefulness scores for the default analysis (based on co-occurrence counts in all four context positions) and separately for each context position (based on co-occurrence counts only in this context position).

The **adjective** category, like verbs, also performs best on the two left context positions. Not surprisingly, **adverbs** and **particles** — the two categories that were found to be indistinguishable in terms of distributional information (cf. 4.1.2) — show very similar patterns of Distributional Usefulness across context dimensions, with some mildly useful information arising from cues in the position [−1] and hardly any useful information being found in the two right context positions.

Pronouns rely mostly on cues in the position [+1], **determiners** on cues from immediate context, and **prepositions** from cues in all context positions, except for [+2]. Finally, **conjunctions** achieve their only appreciable Distributional Usefulness value in the position [+1] while the information in the position [−2] is no better than chance.

In sum, for all categories but adjectives, the most informative context position is always found in immediate context (i.e., in either [−1] or [+1]), while the least informative position is always one of the two more distant context positions [−2] and

[+2].[112] An additional analysis revealed that, for the category system in its entirety, the most and least informative context positions are [−1] and [+2], respectively.[113] This result was obtained by computing — for each context position — an overall goodness score as the weighted mean of the eleven category-specific Distributional Usefulness values, with the weights being proportional to the individual category sizes.[114]

Figure 4-14 can also be interpreted from the perspective of robustness. First of all, each category receives better-than-chance information in each context position — with the single exception of conjunctions in the position [−2]. Thus, even a distributional learner that considers co-occurrence relations only in, say, the position immediately to the left of target words, would not learn much less about the 11 benchmark categories than does a learner relying on all four context positions. Nevertheless, all four context positions together are more informative about each category than is any single position in isolation — with the only exception of particles which perform slightly better when confined to the position [−1]. This indicates that — although there may be redundancy (i.e., correlations) across context positions and single cues — the cues from all four context positions combine in a way that boosts their overall informativeness.

Correspondingly, additional analyses revealed that when the two left context positions [−2, −1] are assessed together, without the two right context positions [+1, +2] (or vice versa), Distributional Usefulness is generally lower than for all four context positions in combination. With respect to robustness, these findings suggest that there is a general effect of declining Distributional Usefulness as the number of context positions is reduced. However, this is only true within the default context window. When this context window is further extended the comprise the three, four, or more context positions to either side of a target word, Distributional Usefulness gradually decreases for almost all categories. The only striking exception is the category of interrogative words which consistently improves as the size of the context window becomes wider (up to eight context positions to either side of the target word). But the default context window [−2, −1, +1, +2] is the most informative about the overall category system; and for interjections and prepositions, the smaller window of the two immediately adjacent context positions [−1, +1] is even more informative.

[112] To be more precise, for the case of the verb category, the most useful information is found in the immediately adjacent context position [−1], but equally well in the more distant position [−2].

[113] These observations are in line with the results of Redington et al. (1998:449ff) who found for their English data, that immediate context is overall more informative than distant context, and left context more informative than right context. This general pattern also extended to the wider context window [−4, −3, …, +3, +4].

[114] The pattern remains the same if one instead computes simple arithmetic (i.e., nonweighted) means.

4.3.2 Distributional preferences: Potential cues

Now we know for the various benchmark categories to which extent they rely on each of the four context positions. But which particular context words constitute the actual cues in these positions? To approach this question, the distributional properties of all categories were studied in detail. Of special interest were, of course, their salient properties, that is, the specific context words that the members of a given category prefer the most as their co-occurrence partner in a particular position?

These salient properties were determined as follows. Each SCO vector was re-standardized such that each of its four components — corresponding to the four context positions — individually has unit mass (cf. 3.1.2). This yields for each target word *t* and for each context position [*cp*] = [−2]; [−1]; [+1]; [+2] a standardized vector

$$v_t^{[cp]} = \left(v_{t,\,cw}^{[cp]}\right)_{cw \in \{\text{context words}\}} .$$

The vector element $v_{t,\,cw}^{[cp]}$ estimates the conditional probability for target word *t* to co-occur with context word *cw* in the context position [*cp*], given the event that a co-occurrence is recorded in this context position.[115] In the following, whenever I speak of distributional properties by using terms such as *more likely*, *percentage*, or even *probability*, these are without exception intended to refer to the respective conditional co-occurrence probabilities, even when the condition (namely, that any co-occurrence is recorded in the specified context position) is not explicitly mentioned.

Next, these values were averaged across the target words of a particular benchmark category Γ, while holding [*cp*] and *cw* constant:

$$\text{pref}^{[cp]}(\Gamma, cw) = \frac{1}{C}\sum_{t \in \Gamma} v_{t,\,cw}^{[cp]} . \qquad (15)$$

[115] Recall that a target word's co-occurrences were only recorded with words selected as context words and with utterance boundary markers. Furthermore, for word tokens occurring in the first or last position of an utterance, no co-occurrences were recorded in context position [−2] or [+2], respectively (cf. 3.1.1). While the base probabilities (which take into account all tokens of a target word and not just those for which a co-occurrence is recorded) may be more insightful linguistically, I was primarily interested in the conditional probabilities because they are the ones on which similarities between SCO vectors are based. Since the default context lexicon captures 87.1% of all word tokens in the corpus, these two probabilities should generally be reasonable approximations of each other, with the conditional probabilities being slightly higher. Only for words which frequently occur in the first or last position of an utterance, the approximation is rather poor, due to the missing co-occurrences in the context position [−2] or [+2], respectively.

This average value $\text{pref}^{[cp]}(\Gamma, cw)$ will be called the *preference* of Γ to co-occur with *cw* in [*cp*].[116] The set of preference values for all context positions and all context words will be called the *distributional profile* of category Γ. This profile reflects the distributional properties that are characteristic of the members of Γ.

Because of the large number of context words, it is not possible to display the complete distributional profiles within this dissertation. However, for current purposes, a category's salient properties — that is, context positions and context words with particularly high preference values — are the most interesting portions of its profile. The most salient properties of each of the 11 benchmark categories are listed in Appendix D.

Additionally, the full profiles were summarized by applying the benchmark lexicon also to context words. Adding up the preference values across all context words of the same category Δ, while holding the context position [*cp*] fixed, yields a quantity

$$\text{pref}^{[cp]}(\Gamma, \Delta) = \sum_{cw \in \Delta} \text{pref}^{[cp]}(\Gamma, cw) \ . \qquad (16)$$

I will refer to $\text{pref}^{[cp]}(\Gamma, \Delta)$ as the *cumulative preference* of Γ to co-occur with Δ in [*cp*]. It estimates the average conditional probability for target words of category Γ to co-occur in context position [*cp*] with any context word of category Δ.

It should be stressed that this summarizing across context word categories is merely used here as a means of investigation. This step allows one to scan the detailed distributional properties for more general patterns and relating them to the underlying linguistic regularities, just as zooming out of a digital image allows the viewer to make out the relevant shapes and contours in the overall picture and not just look at isolated pixels. At the same time, when interpreting these general patterns, one should keep in mind that the context word categories are not the level at which the model operates, for it has only access to co-occurrences between individual words. Presenting a category's distributional properties at the level of individual context words as done in Appendix D therefore captures more closely the perspective of the model.

The summarized profiles of the 11 benchmark categories are presented in Table 4-4 below, showing only the three most likely context word categories in each context position. Interpreting these profiles will provide some insights into the categories'

[116] The term *preference* is meant to insinuate that — in contrast to relative co-occurrence frequencies — these average values quantify the degree to which a particular co-occurrence relation is preferred *only among* the set of co-occurrence relations that were actually considered (cf. footnote 115 above).

confusability relations identified earlier (cf. 4.1.2, esp. Table 4-1 on p. 101). The most salient properties across categories are highlighted.

Among the most salient properties across all categories are co-occurrence preferences with utterance boundary markers. This is not surprising because these markers occur by far more frequently than any context word.[117] The category of **interjections** is the one that relies most extensively on co-occurrences with utterance boundary markers, in each of the four context positions. Interjections thus like to occur in any of the two first and the two last positions of an utterance, but most frequently in the very first (on average, 73.9% of all instances for which a co-occurrence was recorded) and the very last position (51.6%) — this only reflects the fact that these words are frequently used as one-word utterances.[118] Interjections have only few other salient properties; but these former two properties are very pronounced and their combination is fairly unique such that they alone might explain why interjections separate so very well from any category in the SCO vector space (cf. Table 4-1 on p. 101), and ultimately, why their overall Distributional Usefulness is so high (cf. Figure 4-1 on p. 94).

Interrogative words occur in the two first positions of an utterance almost as often as interjections. Much less frequently than interjections, they are also used in the last or last but one position, but even when they do so, the utterance is likely to be a question and thus terminate on the utterance boundary marker _?_ whereas interjections co-occur with this particular marker relatively rarely (cf. pp. 249 and 254 in Appendix D).[119] But rather than occurring utterance-finally, interrogative words clearly prefer to be succeeded by a verb (47.6%) and then by a pronoun (30.8%) which reflects the default word order for *wh*-questions in German.[120]

---

[117] After all, each of the 251,080 utterances in the corpus has a pre-utterance marker, and most utterances (96.9%) terminate on one of the three post-utterance markers, whereas the most frequent context word (*das*; English: *the*$_{\text{neut.}}$, *that*$_{\text{neut.}}$) occurred 38,518 times overall (cf. 2.1.2).

[118] When they do combine with other words, they relatively often occur before the boy's name *leo* (see p. 249 in Appendix D) which underscores the highly communicative function of at least some interjections.

[119] In most such cases, the question will be fairly short, consisting of no more than three word tokens (e.g., *"warum nicht ?"* and *"und wen noch ?"*; English: *"Why not?"* and *"And who else?"*, respectively; examples are taken from the corpus, the relevant target word is underlined). But longer and more complex constructions are also encountered (e.g., *"weisst du auch noch warum ?"*; English: *"Do you (also) still know why?"*).

[120] Appendix D (p. 254) shows more specifically that the most likely verb forms to immediately follow an interrogative word are inflected auxiliary verbs; and the most preferred pronouns in the position [+2] relative to interrogative words are *du* (English: you$_{\text{sg.nom.}}$), *das* (English: *that*$_{\text{neut.:nom.+acc.}}$, *the*$_{\text{neut.}}$), and *wir* (English: *we*).

Table 4-4: *Distributional profile of each category (cumulative summary)*

Left context				Category	Right context			
[−2]		[−1]			[+1]		[+2]	
40.5	<Bnd>	73.9	<Bnd>		51.6	<Bnd>	31.7	V
10.3	V	8.2	INTJ	INTJ	8.3	PRON	22.9	<Bnd>
9.8	DET	3.1	N		6.7	V	11.3	PRON
28.8	<Bnd>	21.9	PRON		29.9	PRON	21.2	<Bnd>
11.6	DET	15.4	ADV	V	23.5	<Bnd>	16.7	PTCL
10.1	PRON	13.5	<Bnd>		11.9	PTCL	14.0	ADV
16.0	V	54.9	DET		34.6	<Bnd>	34.8	<Bnd>
14.5	PREP	8.1	PRON	N	23.5	V	11.7	V
13.3	PTCL	7.4	PTCL		10.1	ADV	9.8	PRON
22.3	V	26.5	PTCL		33.7	<Bnd>	39.5	<Bnd>
14.1	PTCL	25.6	DET	ADJ	25.2	N	13.2	V
13.8	PRON	10.5	ADV		11.2	V	8.7	PRON
22.1	V	20.9	ADV		27.9	<Bnd>	31.2	<Bnd>
19.9	PRON	17.2	PTCL	ADV	19.2	V	11.8	PRON
13.8	DET	16.7	PRON		12.1	ADV	11.3	DET
48.6	<Bnd>	51.0	<Bnd>		47.6	V	30.8	PRON
12.2	V	12.6	CONJ	INTG	20.5	N	16.9	V
8.6	PTCL	6.8	PREP		8.8	<Bnd>	14.6	<Bnd>
21.4	V	35.1	V		20.1	PTCL	21.0	<Bnd>
16.5	<Bnd>	17.5	PRON	PRON	17.8	ADV	13.9	PTCL
16.3	PRON	9.3	PREP		17.2	V	13.5	ADV
19.3	V	23.2	PREP		55.8	N	32.3	<Bnd>
16.6	PRON	14.8	PTCL	DET	11.6	ADJ	17.2	V
13.2	ADV	12.5	V		8.9	<Bnd>	10.9	ADV
18.4	V	18.6	ADV		50.3	DET	36.7	N
17.8	DET	14.8	PTCL	PREP	15.6	PRON	22.8	<Bnd>
14.9	PRON	13.9	<Bnd>		11.5	N	11.3	V
20.2	<Bnd>	29.5	<Bnd>		32.5	PRON	25.3	PRON
15.3	PTCL	17.5	V	CONJ	20.0	V	13.1	DET
13.8	V	11.5	PTCL		14.7	DET	12.0	ADV
25.5	V	19.9	V		26.0	<Bnd>	28.2	<Bnd>
21.9	PRON	17.6	PRON	PTCL	15.5	PTCL	13.5	V
11.5	DET	14.8	PTCL		12.5	ADV	9.6	DET

Note. The distributional properties of a target word category (central column) are summarized by context word categories. The symbol <Bnd> represents all four utterance boundary markers. Only the three most likely context word categories are shown in each context position. The percentages next to them specify cumulative preference values which estimate the average conditional probability at which members of the target word category co-occur with any word of the context word category in the respective context position. Probabilities above 30% are shaded.

A third category preferring to occur at the utterance beginning is that of **conjunctions** (29.5% in the first, and 20.2% in the second position). Conjunctions share with interrogative words also a preference to be followed by verbs and pronouns. Although the average percentages are quite distinct, a number of individual conjunctions and interrogative words may nevertheless be very similar in these preferences; and this would explain why these two categories are more confusable with

each other than with almost any other category (cf. Table 4-1 on p. 101).[121] In fact, the only category from which conjunctions separate even less well is that of **verbs**. These share with conjunctions not only the tendency to occur in utterance-second position but also their strongest preference to be immediately succeeded by a pronoun (29.9% for verbs and 32.5% for conjunctions)[122]. Accordingly, verbs also separate worse from conjunctions than from most other categories.

What is remarkable about verbs is their apparent flexibility to occur in any of the two first and the two last positions of an utterance, with the second position (28.8%) being slightly preferred over the other three. This degree of flexibility to occur at either end of an utterance is only matched by interjections. But while these often occur in one- or two-word utterances and thus close to both utterance boundaries simultaneously, verbs are more frequently used in longer utterances such that their preference to occur close to either utterance end must in fact arise from essentially distinct sets of tokens. This is a first hint that the verb category is distributionally rather complex, and later analyses will make this more explicit (cf. 4.4.1).

The most salient feature in the distributional profile of the **noun** category is the nouns' preference to be immediately preceded by a determiner (54.9%), reflecting the structure of a simple NP. Although this involves a fairly large range of individual determiners, the most common ones (*die*, *der*, and *das* — the three basic forms of the definite article in German) are quite salient by themselves (cf. Appendix D, p. 251).[123] But individually even more salient are the nouns' preferences to occur in the last or last but one position of an utterance (34.6% and 34.8%, respectively), with nearly half of these utterances being questions (again, see p. 251). When they occur in the last but one position, the final word is most typically a verb form. Nouns also show a tendency to co-occur with verbs and prepositions two to their left, with one verb form (*ist*; English: *is*) and three prepositions (*auf*, *in*, *mit*; English: *on*, *in*, *with*, respectively) being individually fairly salient.[124] I shall return to this characteristic set of distributional preferences in subsection 4.4.2 where they are interpreted relative to the underlying syntactic privileges of nouns.

---

[121] From the perspective of interrogative words, the absolute separation value is still fairly high (.75).

[122] In either case, the most salient individual pronoun forms are subject pronouns (cf. pp. 250, 258).

[123] The word *das* was classified as a pronoun, but most of its tokens that occur right before a noun reflect determiner usage. If *das* were classified as a determiner, the cumulative probability of nouns to immediately follow a determiner would thus even rise to 60.0%.

[124] Not surprisingly, these three prepositions are the most frequent ones in the corpus. Likewise, *ist* is by far the most frequent individual verb form, with almost five times as many tokens as the second most frequent verb form.

What matters for current purposes is that the combination of these salient properties set nouns apart from most other categories. Only adjectives and to a lesser degree also adverbs share several of these preferences with nouns, most notably the tendency to occur in the last two utterance positions; and this explains why nouns were found to separate least well from these two categories (cf. Table 4-1 on p. 101). Correspondingly, adjectives and adverbs are also fairly confusable with nouns and with each other.

The most salient properties of the **adjective** category are those that are consistent both with predicative usage and attributive usage of adjectives: namely, to occur close to the utterance end, to co-occur with the verb form *ist* (English: *is*) in the context position [−2], and to co-occur with the scaling particle *ganz* (English: *quite*, *very*) in the position [−1] (cf. Appendix D, p. 252).[125] Their average tendency to immediately follow a determiner (overall 25.6%) is clearly less pronounced than that of nouns, because this preference is inconsistent with predicative usage. Conversely, adjectives are likely to be immediately preceded by a verb form when used predicatively, but rather unlikely to do so in attributive usage, unless the adjective modifies a plural noun form.

Mirroring the noun profile, **determiners** most strongly prefer to be followed by a noun (55.8%), but individually, no single noun is highly preferred in this position (cf. Appendix D, p. 256).[126] More salient are the determiners' tendency to immediately precede one of the three most frequent prepositions (*mit*, *in*, *auf*; English: *with*, *in*, *on*, respectively). Like nouns and adjectives, determiners are very likely to occupy the last but one position of an utterance (32.3%) — and they do so both in declaratives (on average 19.7%) and questions (12.1%) —, but they occur much less often utterance-finally (8.9%) than do nouns and adjectives. Also remarkable are a number of subject (i.e., nominative) pronoun forms (*das*, *du*, *ich*, *wir*; English: *that*$_{\text{neut.}}$, *you*$_{\text{sg.}}$, *I*, *we*) with which determiners like to co-occur two positions to their left (again, see p. 256).

The distributional profile of **prepositions** shows unmistakable traces of the structure of a simple PP, for prepositions strongly prefer to be followed by a determiner

[125] Note that attributive usage generally requires an inflected adjective form whereas predicative usage only applies to the uninflected base forms. Since in this study, different forms of the same lexeme are treated as different word types, the adjective category here essentially falls into two non-overlapping sets of word forms. It would be interesting to distributionally analyze these two sets as categories of their own but this proposal was not pursued here. On another note, one may at first be surprised to read that even in attributive usage, adjectives frequently occur in the last position of an utterance. This preference arises from the fact that the modified noun is often omitted — ellipsis is a very common phenomenon in spoken language — e.g., *"noch eine grosse ."* (English: *"Another big (one)."*).

[126] The reason is that most nouns have a relatively low base frequency. Only because there are so many different nouns, do the individually not very salient preferences of determiners to be followed by a particular noun add up to the exceedingly high cumulative preference value of 55.8%.

(on average, 50.3%) and then a noun (36.7%). Several individual determiners (mostly definite articles for different grammatical genders and cases) constitute fairly salient preferences by themselves (cf. Appendix D, p. 257).[127] By contrast, such high preference values were not found for the co-occurrence relation of prepositions with any individual noun (again, cf. p. 257). The single most salient property of prepositions is their tendency to be used utterance-initially (13.9%) which they do most typically in utterances that consist of nothing but a PP. When they do not occupy this position, prepositions most frequently follow an adverb (18.6%) or particle (14.8%), and a few specific adverbs and articles constitute somewhat salient preferences individually (cf. p. 257). This set of distributional properties is fairly unique which explains why prepositions separate fairly well from any other category (cf. Table 4-1 on p. 101).

The three remaining categories (adverb, pronoun, and particle) were found to be mutually highly confusable (cf. Table 4-1 on p. 101). In fact, between **adverbs** and **particles**, there appeared to be virtually no distributional difference at all. Inspection of their distributional profiles confirms this at both analytical levels, that is, for individual context words (Appendix D, pp. 253, 259), as well as for summarized context word categories (Table 4-4 above). Adverbs and particles are very similar not only regarding the set of context words they prefer to co-occur with, but also with respect to the specific preference values of most of these co-occurrence relations. As their most salient properties, both categories frequently occur in the last two positions of an utterance (likewise in declaratives and questions). Although they take less frequently the first or second utterance position, these are not rare co-occurrence events, either. Moreover, both like to be immediately preceded by other adverbs and particles such as *nicht*, *mal*, *auch*, *noch* (English: *not*, *once*, *also*, *still*, respectively).[128] Another set of properties that adverbs and particles have in common are their preferences to co-occur with subject pronoun forms and verb forms (again, most frequently with *ist*) one or two words to their left, and with determiners in the context positions [−2] and [+2].

Pronouns share all these preferences but differ more substantially in the particular preference values. Most notably, they appear much less frequently than adverbs and particles in the last two positions of an utterance; instead, they occupy more often the utterance-second position. These commonalities in their salient properties and the

[127] Note that each preposition in German requires its dependent NP to take a specific grammatical case. Some prepositions are associated with more than one case, but even then, each case reflects a different semantic relation.

[128] These translations should not be taken too literally — especially particles often have very complex and context-dependent meanings for which no obvious lexical counterparts exist in English.

differences in exact probabilities explain why pronouns are mutually fairly confusable with adverbs and particles but at the same time clearly less confusable than are adverbs and particles with each other.

4.3.3 Distributional discriminators: Identifying informative cues

These analyses demonstrate that comparing only the few most salient distributional properties of target word categories can be quite revealing as to why and in which ways a given category is distributionally more confusable with some categories than it is with others. It was hardly surprising to observe that two categories tend to be the more confusable, the more of their salient co-occurrence preferences they share. But how do these findings project to the categories' overall Distributional Usefulness scores? A particular preference that any given category has in common with one or two other categories may not be shared by the majority of other categories and thus still be useful to learn something about this category. So the question is how a category's distributional differences and commonalities with other categories combine to set this category apart from all other categories *simultaneously*, rather than pairwise.

To study this question with respect to a category Γ, it is useful to treat the set of target words that are *not* a member of Γ as a single category of their own and to compute the distributional profile of this *inverse category* Λ\Γ (where Λ denotes the lexicon of all target words). The obvious intuition is that the more similar this profile is with that of Γ itself, the less useful should distributional information be for acquiring Γ. Accordingly, those specific properties in which Γ and its inverse category differ substantially should be particularly useful for discriminating the members of Γ from the nonmembers.

In order to detect these *discriminators* (i.e., discriminating properties), the profile of the inverse category was subtracted from the profile of Γ. That is, for each context word *cw* (lexical or virtual) and each context position [*cp*], I computed the difference

$$\begin{aligned} &\mathrm{pref}^{[cp]}(\Gamma, cw) \;-\; \mathrm{pref}^{[cp]}(\Lambda \setminus \Gamma, cw) \\ &\qquad = \quad \frac{1}{C} \sum_{t \in \Gamma} v_{t,\,cw}^{[cp]} \;-\; \frac{1}{L-C} \sum_{t \in \Lambda \setminus \Gamma} v_{t,\,cw}^{[cp]} \;, \end{aligned} \tag{17}$$

where L = 1,017 denotes the size of the target lexicon Λ. This difference will be called the *relative preference* of Γ to co-occur with context word *cw* in context position [*cp*]. It quantifies how much more likely it is, on average, for a member of Γ to co-occur with

cw in [*cp*] than it is for a nonmember. When the difference is negative, the category members enter this co-occurrence less likely than do the nonmembers. In such a case, the context word *cw* constitutes, in the context position [*cp*], a *negative discriminator* with respect to Γ. Accordingly, a context word and context position for which the relative preference value is positive will be called a *positive discriminator*. In either case, when the relative preference value for *cw* and [*cp*] is substantially different from zero, it is safe to conclude that the co-occurrence relations with *cw* in [*cp*] provide some useful information about Γ. Note that testing for the statistical significance of relative preference values being different from zero would be rather pointless because, due to the large number of data points (1,017), even tiny values tend to be highly significant.

Yet, even a very high relative preference value does not guarantee that it is the entire category that benefits from this information. Consider the following example. Nouns on average have a fairly high preference to be immediately preceded by any of the three definite articles *die*, *den*, and *das* (cf. Appendix D, p. 251). However, in German, determiners and nouns have to agree in number, case, and gender when they combine to an NP. The inflectional system of determiners is somewhat ambiguous; but when restricted to singular number, each of the three determiners *die*, *den*, and *das* is unambiguously associated with one particular gender: *die* with feminine, *den* with masculine, and *das* with neuter.[129] Due to agreement, these associations also become evident in the co-occurrence preferences of grammatical subclasses of the noun category defined by number and gender (Table 4-5 below).

Feminine singular noun forms have an exceedingly high preference to be preceded by *die* but they almost never follow *den* or *das*. Likewise, masculine singular nouns on average occur frequently right after *den* but only rarely after *die* and *das*, whereas neuter singular nouns strongly prefer to follow *das* but not *die* or *den*.

The fact that feminine singular nouns co-occur with *den* and *das* at all (and likewise masculine singular nouns with *die* and *das*, neuter singular nouns with *die* and *den*) arises from at least four factors. The least relevant one concerns speech and transcription errors. Second, some German nouns (e.g., *finger* and *fenster*; English: *finger*, *window*, respectively) have identical singular and plural forms such that they like

[129] Although the target word *das* was classified as a pronoun, the context word *das* is discussed here in its determiner function. Further note that while *die* and *das* may have either nominative or accusative case, *den* is unmistakably accusative, when used with a singular noun form. For the purpose of the current example, *den* was chosen instead of the corresponding nominative form *der* because *der* is highly ambiguous with respect to gender as it can also form an NP with feminine nouns in genitive or dative case. In the corpus, singular feminine nouns on average prefer to follow *der* at a rate of 12.1%.

to follow *die* or *den* in their plural reading, irrespective of their gender. The third factor concerns lexically ambiguous nouns such as *see* which has a masculine homonym (for English: *lake*) and a feminine homonym (for English: *sea*, *ocean*). And finally, there is the phenomenon of categorial ambiguity which may concern either the noun or the determiner. For instance, *runde* (English: *round*, in either case) can be a feminine singular noun but also an inflected adjective form that is unmarked for gender and thus likes to follow determiners of any gender. As for determiners, *die*, *den*, and *das* — like most other determiners (cf. Table 2-2 on p. 52) — can also be used as pronouns; and as such they are not unlikely to appear right before singular nouns that have a mass noun interpretation (e.g., *spass* as in the corpus example *"hat das spass gemacht ?"*; English: *"Was that fun?"*). The important observation is that, despite these four factors, the preference of each noun gender subclass to be preceded by the corresponding determiner is one or two orders of magnitude greater than that to be preceded by the other two determiners.

Table 4-5: *Selected preferences of noun gender subclasses*

Noun subclass [a]		**Context word in position [−1]**		
		die	*den*	*das*
Feminine singular	(60)	31.2	< 0.1	0.1
Masculine singular	(82)	0.7	9.3	0.3
Neuter singular	(57)	1.3	0.3	22.3
All nouns	(268)	12.5	3.3	5.0
All non-nouns	(749)	2.3	0.6	3.1

Note. Co-occurrence preferences of noun gender subclasses to immediately follow a particular determiner form. Preference values are given in percent. Agreement associations between gender class and determiner form are shaded. For the purpose of comparison, the two bottom rows present the corresponding co-occurrence preferences of the overall noun category and its inverse category.

[a] Number of target words is given in parentheses.

These are very pronounced examples of how category-wide averages such as preference and relative preference can be deceiving. Particularly the preference of the overall noun category to immediately follow the context word *die* (12.5%) is not very

representative of most nouns. And this is not simply a statistical problem in the form of some large variance around the overall mean value; rather, the noun subclasses differ systematically in their co-occurrence preferences with *die* — in fact, they differ so much that, on average, masculine and neuter singular nouns follow *die* even less likely than do non-nouns (0.7% and 1.3% vs. 2.3%, respectively). Co-occurrences with *die* might therefore be extremely useful for discovering the subclass of feminine singular nouns — probably together with plural noun forms which prefer to follow *die* at a similar rate — but at the same time, *die* discriminates these nouns from most masculine and neuter singular nouns even more than from many non-nouns. Of course, by achieving the former, this determiner definitely does constitute a very useful cue that could help to learn something about the noun category — but not about the noun category in its entirety (cf. 4.3.4 for a discussion of partially informative cues).

The relative preference value of the overall noun category to follow *die* (10.2% = 12.5% − 2.3%; cf. Appendix D, p. 251) does not capture this since it merely considers the category-wide preference across all nouns (12.5%), in comparison to the corresponding preference across all non-nouns (2.3%).

Therefore, an additional measure, *discriminative power*, was defined that quantifies how well a particular cue by itself discriminates an entire category from all other categories. It considers the full distribution of relevant co-occurrence probabilities of all members and nonmembers of a given category, and not just the averages of these probabilities. Formally, the discriminative power of a context word *cw* for a category Γ, with respect to a given context position [*cp*], is given by

$$\mathrm{dp}^{[cp]}(\Gamma, cw) \quad = \quad \frac{1}{C(L-C)} \sum_{t \in \Gamma} \sum_{s \in \Lambda \setminus \Gamma} \left| v^{[cp]}_{t,\,cw} - v^{[cp]}_{s,\,cw} \right| \quad - \quad \frac{2}{C(C-1)} \sum_{i=1}^{C-1} \sum_{k=i+1}^{C} \left| v^{[cp]}_{t_i,\,cw} - v^{[cp]}_{t_k,\,cw} \right| \quad . \qquad (18)$$

This difference determines, with respect to the target words' individual co-occurrence probabilities with *cw* in [*cp*], how much more similar these probabilities are, on average, between any two members of Γ than they are between any member and any nonmember. Alternatively, $\mathrm{dp}^{[cp]}(\Gamma, cw)$ can also be described as the average L_1 distance between the members of Γ (along the single context dimension given by *cw* and [*cp*]) minus the average L_1 distance between members and nonmembers (along this same dimension). Discriminative power is thus related to the overall distance values from which Distributional Usefulness is computed; and it can be interpreted as a very

rough measure for the contribution of the given context dimension to the overall Distributional Usefulness of Γ.

When $\mathrm{dp}^{[cp]}(\Gamma, cw)$ is clearly positive, co-occurrences with *cw* in [*cp*] are indeed informative about the category as a whole, whereas in the earlier example, *die* in the context position [−1] should achieve a negative discriminative power for the noun category, reflecting that, by itself, it does not very well separate the noun category from non-nouns. Importantly, these interpretations hold regardless of whether one examines positive or negative discriminators which are defined in terms of relative preference, and not by discriminative power (cf. p. 138).

Discriminative power was used only as a secondary measure to get a better understanding of what a clearly nonzero relative preference value actually means for the overall category. For this purpose, equation (18) is a sufficient implementation of the intuitive notion of the discriminative power of a given cue, even though it has some weaknesses.[130] What matters most about discriminative power is whether or not it is greater than zero (coupled with a clearly positive or clearly negative relative preference). Pilot explorations suggest that, as an additional rule of thumb, a context dimension with discriminative power greater than +.02 can be considered fairly informative about the respective category in its entirety. Applying statistical significance tests would not be very insightful here: In general, even very small discriminative power values are significantly different from zero, due to the large number of data points (cf. related considerations with respect to relative preference, p. 139).

Note that the scale of discriminative power depends on the size of the given target word category, with values tending to be greater for larger categories. As a consequence, discriminative power values for different context dimensions (i.e., different combinations of context word and context position) should only be compared when the category is held constant. This problem could in principle be overcome by

[130] One weakness is that a particular cue may show its usefulness for the overall category only in interaction with some other cues (cf. 4.3.4); and discriminative power is completely blind to these interactions. Although relative preference also looks at individual cues only in isolation, this measure can at least assess whether these cues bring to the table any distinct information that has the potential to interact with other cues in the first place. A second weakness of the way discriminative power was implemented is that it dissociates differences between co-occurrence probabilities from the specific pairs of target words they were observed for. It might therefore miss some relevant aspects of how target words are lined up on the given context dimension. This is essentially the same problem that was briefly discussed for ANOVA-based evaluation scores that implement the notion of discriminative power at the level of multidimensional SCO vectors (cf. 3.3.5). However, it is not as dramatic in the current situation where only one dimension is evaluated, since now at least the arrangement of values for category members can be recovered completely from the set of differences between them.

taking into account the variance of all target words along any given context dimension. But such standardization is not desirable because it would overemphasize the role that context dimensions with small variance play for the model. In this sense, discriminative power quantifies quite naturally the contribution of the given context dimension to the overall Distributional Usefulness of the given benchmark category (cf. the earlier interpretation on pp. 141f).

As was done earlier for the original distributional properties, it is worthwhile to also summarize discriminators by the context word categories, in order to reveal some linguistically meaningful patterns across discriminators. To this end, I summed up, for each target word category, all its relative preferences with context words of the same category and in the same context position. For the case of discriminative power, however, there is no correlate at the level of context word categories — it would not be meaningful to add up the individual discriminative power values. The most important discriminators of each target word category can be looked up in Appendix D while the summarized cues are presented in Table 4-6 (positive discriminators) and Table 4-7 (negative discriminators) below, showing only context word categories for which the cumulative relative preference is above 5% or below −5%, respectively.[131] In the following, these three resources (the two tables and Appendix D) are interpreted simultaneously.

For **interjections**, their salient preference to appear utterance-initially is indeed a very strong positive cue to this category — on average, an interjection does so 63.2% more frequently than does a non-interjection. No other category has a cue of a similarly high relative preference level. Furthermore, the discriminative power of this cue (+.414; cf. Appendix D, p. 249) is exceedingly high, indicating that it is indeed the entire category that benefits from it. The same context position [−1] also accommodates a large number of negative discriminators for interjections; in particular, interjections are much less likely to be preceded by a pronoun or a determiner than are non-interjections. Given these observations, it is not surprising at all that Distributional Usefulness for interjections was found to be greater for this context position in isolation than for any of the other three context positions (cf. Figure 4-14 on p. 129). Other strong positive discriminators are the interjections' preference to take the second or last position of an utterance, the latter only when the utterance terminates on a period or exclamation

[131] Context word categories were assigned to either table only by their cumulative relative preference. That is, any context word category listed in Table 4-6 might nevertheless comprise context words that individually are negative discriminators (likewise for Table 4-7 and positive discriminators).

mark. In fact, co-occurrences with the question mark in context positions [+1] and [+2] even constitute strong negative discriminators for the interjection category.

Table 4-6: *Positive discriminators of each category (cumulative summary)*

Left context		Category	Right context	
[−2]	[−1]		[+1]	[+2]
+22.7 <Bnd>	+63.2 <Bnd>	INTJ	+25.4 <Bnd>	+20.6 V
+12.9 <Bnd> +5.5 INTJ	+12.1 PRON +8.3 N +6.7 ADV +6.2 INTG	V	+25.9 PRON	+9.1 PTCL +6.3 ADV
+12.5 PREP	+47.5 DET	N	+13.7 V +8.8 <Bnd>	+8.7 <Bnd>
+7.8 V	+16.9 PTCL +6.3 DET	ADJ	+17.9 N +6.1 <Bnd>	+12.4 <Bnd>
+8.4 PRON +7.5 V	+11.4 ADV +6.6 PTCL +5.7 V	ADV	+6.4 V	
+29.6 <Bnd>	+36.1 <Bnd> +8.9 CONJ	INTG	+34.7 V +11.7 N	+20.6 PRON +5.1 DET
+6.3 V	+28.6 V +6.0 PREP	PRON	+11.2 PTCL +9.9 ADV	
	+21.0 PREP +5.3 V	DET	+49.8 N +10.2 ADJ	
+7.7 DET	+8.1 ADV +5.6 V	PREP	+44.2 DET	+30.6 N
	+14.2 <Bnd> +10.0 V	CONJ	+21.4 PRON +8.1 DET +6.6 V	+14.9 PRON
+10.8 V +10.2 PRON	+13.0 V	PTCL	+6.5 PTCL +5.9 ADJ	

Note. The positive discriminators of a target word category (central column) are summarized by context word categories and listed in the column for the respective context dimension. The symbol <Bnd> represents all four utterance boundary markers. Only context word categories with cumulative relative preference above +5% are shown. These cumulative relative preference values (given to the left of each context word category) estimate how much *more* likely, on average, a member of the target word category co-occurs with any word of the context word category in the respective context position, than does a nonmember. Values above +30% are shaded.

In comparison to other words, **interrogative words** occupy by far more often the first two utterance positions, and at the same time clearly less frequently the last two positions in utterances that terminate on a period. These two strong positive and two strong negative cues are all highly informative about the category of interrogative words as a whole (discriminative power is very high in each case, cf. p. 254), and their combination should contribute a lot to the high Distributional Usefulness for interrogative words which was also confirmed by Figure 4-12 (p. 122). But the same

Table 4-7: *Negative discriminators of each category (cumulative summary)*

Left context		Category	Right context	
[−2]	[−1]		[+1]	[+2]
−5.4 PRON	−20.0 DET	INTJ	−7.3 V	−5.9 <Bnd>
−5.4 V	−12.4 PRON		−6.7 PTCL	
	−8.8 PTCL			
	−8.5 ADV			
−10.1 V	−19.2 DET	V	−12.1 V	−9.9 <Bnd>
	−5.3 V		−8.6 N	−7.4 V
			−6.4 <Bnd>	
−10.0 <Bnd>	−12.6 <Bnd>	N	−12.7 PRON	
	−9.3 ADV		−8.2 N	
	−6.9 PRON			
	−6.0 N			
	−5.9 V			
	−5.1 PTCL			
−7.7 <Bnd>	−7.6 <Bnd>	ADJ	−10.2 PRON	−5.4 PTCL
−11.3 <Bnd>	−18.2 DET	ADV	−9.6 PRON	
	−8.6 <Bnd>		−5.5 N	
−7.2 DET	−19.5 DET	INTG	−19.6 <Bnd>	−14.0 <Bnd>
−5.5 PRON	−10.6 PRON		−5.9 PTCL	
	−6.5 ADV			
−5.7 DET	−16.3 DET	PRON	−17.0 <Bnd>	−7.6 <Bnd>
−5.4 PTCL	−7.2 <Bnd>		−5.5 N	
	−5.1 PTCL			
−7.4 <Bnd>	−13.6 DET	DET	−20.4 <Bnd>	−7.8 PRON
	−6.6 <Bnd>		−10.5 PRON	
			−7.4 V	
			−6.1 PTCL	
			−5.4 ADV	
−10.2 <Bnd>	−16.2 DET	PREP	−21.2 <Bnd>	−8.4 PRON
			−12.4 V	−5.6 <Bnd>
				−5.6 PTCL
	−19.1 DET	CONJ	−18.7 <Bnd>	−19.4 <Bnd>
	−7.0 PRON		−6.7 N	
−9.5 <Bnd>	−16.8 DET	PTCL	−7.8 PRON	
			−5.2 N	

Note. The negative discriminators of a target word category (central column) are summarized by context word categories and listed in the column for the respective context dimension. The symbol <Bnd> represents all four utterance boundary markers. Only context word categories with cumulative relative preference below −5% are shown. These cumulative relative preference values (given to the left of each context word category) estimate how much *less* likely, on average, a member of the target word category co-occurs with any word of the context word category in the respective context position, than does a nonmember. Values below −20% are shaded.

figure indicates that this category relies almost as much on lexical co-occurrences as it does on the above cues from serial position. Strong positive cues of this kind result from co-occurrences with verbs in the context position [+1] and with pronouns in [+2]. At the level of individual context words, there are several verb and pronoun forms that constitute fairly strong positive cues but none of them achieves a positive discriminative

power which indicates that the various interrogative words differ as to which particular verbs and pronouns they like to be followed by.[132] The strongest negative cues from lexical co-occurrences is the relatively low preference of interrogative words to be preceded by a determiner or pronoun (be it one or two words to their left).

Two very interesting positive cues with a clearly positive discriminative power are the high preferences of interrogative words to be immediately preceded by the conjunction *und* (English: *and*; relative preference +10.1%, cf. p. 254) and to co-occur with the modal particle *denn* (no English counterpart) in context position [+2] (relative preference +4.6%).[133] What is remarkable about these cues is that they are highly idiosyncratic — interrogative words do not frequently co-occur with any other conjunctions in [−1] nor with any other particles in [+2].[134] Overall, most of the stronger positive cues to interrogative words concern the context positions [−1] and [+2] which is consistent with the earlier finding that interrogative words achieve the most useful information from these two positions (cf. Figure 4-14 on p. 129).[135]

The most important cues to the category of **conjunctions** are their fairly high preference to occur utterance-initially (relative preference +14.2%), and their rather low tendency to occupy the two utterance-final positions, irrespective of the kind of utterance termination (overall relative preference −18.7% for the last, and −19.7% for the last but one position), reflecting that they are generally followed by some other words which is hardly surprising. At the level of context word categories, conjunctions, on average, are clearly more likely to be followed by a pronoun (immediately or two words to their right) than are non-conjunctions. And there are many pronouns that individually constitute positive cues (cf. p. 258) but none of them is informative about the conjunction category as a whole. By contrast, various individual negative cues that instantiate the low preference of conjunctions to immediately follow a determiner are informative about the overall conjunction category. Nonetheless, most of these cues, especially the positive ones, are individually not very strong.

---

[132] A quite intuitive example to illustrate this is the interrogative pronoun *wer* (English: *who*) which is very naturally followed by a third-person singular verb form and pronoun (as in the corpus utterance *"wer hat das gesagt ?"*; English: *"Who (has) said that?"*); but there is no example in the corpus for the odd-sounding combinations of *wer* with other inflected forms of the same verb *haben*.

[133] The particle *denn* is frequently used in questions, serving a number of communicative functions, but it was not expected that it would so reliably occur two words after the interrogative word.

[134] To be precise, interrogative words actually do co-occur fairly often with one other particle, namely with *'n* which is a contracted form of several different words (e.g., *denn*, *ein*, *einen*, *ihn*), but by far the most frequently of *denn*. This contracted form even shows up among the most dominant positive cues to interrogative words (p. 254) but its discriminative power is slightly below zero.

[135] It is unclear, however, why the context position [−2] is not more informative overall, despite the highly useful cue from occurrences of interrogative words in the second utterance position.

Verbs on average occur clearly more frequently right before (cumulative relative preference +25.9%) or after (+12.1%) a pronoun than do non-verbs. But no single pronoun has a positive discriminative power for the overall verb category; most of these values are even clearly below zero (cf. p. 250). In fact, the only positive cue that appears to be at least mildly useful for the verb category as a whole is a preference of verbs to take the second position of an utterance (relative preference +12.9%). The most prominent negative cues for verbs consist of their comparatively low preferences to occupy the last two positions of an utterance, to succeed determiners, and to occur right next to other verbs (on either side). But none of these negative cues is particularly pronounced, be it in terms of relative preference or discriminative power. In sum, the discriminators for the verb category tend to be rather weak — an important result I examine in detail in subsection 4.4.1.

The cumulative relative preference of **nouns** to immediately succeed a determiner form (+47.5%) is impressive. But although several specific determiners (most of all the definite articles *der*, *die*) achieve fairly high relative preference values individually, none of them is actually informative about the entire noun category (cf. p. 251). Of course, this is just what was predicted by the earlier example (cf. Table 4-5 on p. 140) which motivated the distinction between relative preference and discriminative power in the first place. The positive cues of co-occurrences with prepositions in the position [−2] essentially share the same fate. The only positive cues with a clearly positive discriminative power are the nouns' high preferences to occupy one of the two utterance-final positions, regardless of the utterance termination type. At the same time, nouns occur much less frequently than non-nouns in the two first positions of an utterance; and these two strong negative cues achieve the highest discriminative power values of all cues to the noun category. Most other negative cues concern the context positions [−1] and [+1], and here particularly co-occurrences with adverbs and pronouns.

Adjectives have only few strong cues. On average, they are more likely to occur in the last but one or the last position of an utterance, most of all in declarative utterances. Other positive cues are co-occurrences with the verb form *ist* (English: *is*) two words to their left, and with scaling particles such as *ganz*, *ziemlich*, and *sehr* (English for all three: *quite*, *very*) immediately to their left (cf. p. 252). Adjectives like to be preceded by a determiner and followed by a noun but neither nouns nor determiners are particularly prominent among the individual discriminators for adjectives. Finally,

although adjectives occupy quite frequently the first two utterance positions, they are clearly less likely to do so than are non-adjectives, on average.

Relative to all other words, **determiners** are, on average, 49.8% more likely to be immediately followed by a noun and 21.0% more likely to be preceded by a preposition. Prepositions are also the strongest positive cues at the level of individual context words although none of them is informative about the determiner category as a whole (cf. p. 256); and this is probably due to the fact that each preposition requires some particular grammatical case such that different prepositions co-occur with different though overlapping sets of determiners.[136] The impressive cumulative relative preference of determiners to precede nouns is not based on very strong cues from individual nouns (the strongest one being *seite*; English: *side, page*). The only positive cues that are useful for the overall determiner category arise from a slightly above-average preference to occur in the last but one position of an utterance (be it a question or a declarative; cf. p. 256). However, with respect to the first, the second, and especially the last utterance position, determiners are in fact much less likely to occupy these positions than are non-determiners, and these negative cues are by far the strongest discriminators for this category. Additionally, determiners are relatively unlikely to be immediately preceded by other determiners, and to be succeeded by a pronoun.

Maybe the most noteworthy observation with respect to **prepositions** is that this category and the category of pronouns are the only ones that simultaneously receive negative cues from occurrences in any of the two first and the two last utterance positions (irrespective of utterance termination type; cf. p. 257). This is even more remarkable since one of these relations — namely, occupying the first position in an utterance — is, after all, the most salient individual property of prepositions (cf. 4.3.2 and p. 257); but as a consequence, it is just a fairly weak negative cue. The strongest positive cues to the preposition category are to be followed by a determiner (cumulatively +44.2%) and then by a noun (+30.6%). At the level of individual context words, determiners in the context position [+1] are also the strongest cues, but due to the before-mentioned case-agreement between preposition and dependent NP, most of these determiners are not very useful for the overall preposition category.[137] There are

[136] Also see footnote 127 (p. 137).

[137] The only exception with at least a slightly positive discriminative power is the definite article *den* which can be either accusative (in combination with masculine singular) or dative (plural, regardless of gender) and thus is a likely successor of most prepositions.

even two individual nouns (*bett* and *gegend*, both feminine singular; English: *bed* and *surroundings*, respectively) for context position [+2] among the dominant positive cues; but these are exceptions that arise from significant collocations in the corpus.[138] Most nouns are individually not very strong cues to the preposition category, which might explain why Distributional Usefulness for prepositions is so low in context position [+2] (cf. Figure 4-14 on p. 129).

Interestingly, some determiners constitute positive cues that are at least mildly useful for the overall preposition category when they occur two words to the left of a preposition (cf. p. 257) although the cumulative relative preference for such cues is not too high (+7.7%).[139] Cumulative relative preference is also not very high for the prepositions' preference to be immediately preceded by an adverb (+8.1%) or particle (+3.7%); but at the item-specific level, there are a few strong positive cues with positive discriminative power.[140] Other negative cues involve a comparatively low preference of prepositions to occur right after a determiner, and to be followed by a pronoun (one or two words to their right). In sum, there is a large number of informative cues to the preposition category, but most of them are not very strong (both in terms of relative preference and discriminative power) which might explain the intermediate Distributional Usefulness score for this category.

In subsection 4.3.2, the distributional profiles of **adverbs** and **particles** were discussed together because these categories are distributionally so very similar. At the same time, they are also the two categories with the lowest overall Distributional Usefulness scores (cf. Figure 4-1 on p. 94). Now it turns out that, while all other categories — with the exception of pronouns (see below) — have at least one

---

[138] Both these nouns do not occur very frequently in the corpus but they entertain significant collocations which all take the form *PREP DET N*: *"in 's bett"* (English: *"into (the) bed"*), *"unter 'm bett"* (English: *"under the bed"*), or *"im bett"* (English: *"in (the) bed"*; note that *im* is a blend between the preposition *in* and the determiner *dem*) which together account for 40.6% of all occurrences of *bett*; and *"durch die gegend"* (English: *"around the place"*) which by itself accounts for 82.4% of all occurrences of *gegend*.

[139] There are several linguistic constructions underlying this reliable co-occurrence relation. Most important are utterances in which the determiner is used pronominally as in *"weil die naemlich unter die erde fahren"* (English: *"(That is) because they go under (the) ground"*). In another relevant type of utterances, the target word is not actually used as a preposition but instead as a separable verb prefix which is a common secondary usage of many prepositions. For instance, the preposition *unter* in the particle verb *untergehen* is separated in utterances such as *"hier geht die sonne unter ."* (English: *"Here the sun is going down."*). A third type concerns complex NPs of the form *DET N PREP NP* such as *"der elefant auf dem skateboard"* and *"der papa von bobo"* (English: *"the elefant on the skateboard"*, *"the daddy of Bobo"*, respectively).

[140] These co-occurrences can arise from default word ordering in declarative sentences such as *"das ist zu kalt hier auf dem boden ."* (English: *"It is too cold here on the floor."*). Additional co-occurrences of this sort arise from cases where the preposition actually is a separable verb prefix such as *nach* in *"guck mal nach"* (English: *"(now) go and check/look"*).

reasonably strong positive discriminator, adverbs and particles do not. Across all positive cues to either category, the greatest relative preference value is +4.2%, and the greatest discriminative power value is +.008 (cf. pp. 253, 259). Inasmuch as positive cues reflect what is characteristic about a category, there appears to be very little that would characterize adverbs and particles.[141]

Relative to other categories, adverbs and particles are less likely to occupy the first, second, or last position of an utterance but only barely more likely with respect to the last but one utterance position. Other positive cues involve co-occurrences with verbs in the two left context positions and with particles and adverbs (as context words) in context position [−1]. Furthermore, adverbs and particles (as target words) are more likely than words of any other category to co-occur with pronouns in the context position [−2], and likewise with determiners such as *das* or *die* that can be used as pronouns. In the context position [+1], co-occurrences with pronouns constitute negative cues to these two categories; but they are individually not very strong.

As was noted earlier, **pronouns** — like prepositions — receive negative cues from occurring less frequently than other words in any of the two first and two last utterance positions (regardless of utterance termination type; cf. p. 255). Like adverbs and particles, pronouns have only very weak positive cues. However, at the cumulative level, they are by far more likely to be preceded by a verb than are non-pronouns. At the level of dominant individual context words, *hat* (English: *has*) is the only example of this type of cue. In the context position [−2], pronouns receive positive cues from a range of context word that belong to various categories (especially conjunction, adverb, and interrogative word) which in turn appear to have in common that they frequently occur utterance-initially.[142] Pronouns are also more likely than non-pronouns to be immediately followed by an adverb or particle, and to be immediately preceded by a preposition, corresponding to simple pronominal PPs. Pronouns also receive cues from co-occurrences (or lack thereof) with other pronouns which mainly concerns subject

[141] In the case of particles, this reflects the special status of this category which combines a heterogeneous collection of subclasses that linguists classify as one single category not so much because of shared syntactic privileges but rather because they do not fit into any other category (cf. Helbig, 1994). Further analyses on the *Leo* corpus (similar to those carried out for verbs in 4.4.1) revealed that these subclasses themselves have distinct distributional profiles that are not consistent with each other. In consequence, these distinct profiles level each other out to a rather generic distributional profile for the entire particle category that shares some features with many other categories, especially with closed classes. To a lesser degree, this also holds for the adverb category.

[142] Together with the cue from verbs, this partly arises from constructions of the form *ADV/INTG V PRON …* where the adverb or interrogative word occupies the default slot for subjects. In German, this results in *inversion* such that the subject pronoun is moved into the first postverbal position (cf. p. 167).

(i.e., nominative) pronoun forms. These co-occurrences result in positive cues to the pronoun category in context position [−1], and in negative cues in [+1].

All these observations for the various target word categories lead to four general conclusions. First, while most of the stronger positive cues arise from salient properties of the respective category, the converse is not true. In fact, highly salient co-occurrence preferences may even constitute negative cues. For instance, nouns and adjectives occupy the first two positions of an utterance quite frequently, but still less frequently than do non-nouns and non-adjectives such that these salient properties are nevertheless negative cues about these two categories (cf. pp. 251f). Likewise, although the most salient properties of pronouns and prepositions are occurrences in any of the two first and the two last utterance positions, they receive negative cues from all of these serial positions (cf. pp. 255, 257).

Second, when the cumulative relative preference of a target word category to occur with nouns or verbs (as context words) is substantially different from zero, there are generally no strong cues at the level of individual nouns or verbs. Apparently, these two categories are too large and most individual verbs and, particularly, nouns occur too infrequently to constitute strong cues by themselves. Although even very weak cues could potentially combine to more useful evidence (cf. 4.3.4), the full potential of co-occurrences with nouns or verbs cannot be exploited until at least some rudimentary noun category or verb category is acquired (cf. the experiment in 4.5.2).

The third general conclusion is closely related to the previous one. Across target word categories, the dominant positive and negative cues involved either utterance boundary markers (i.e., information about serial position and utterance termination type) or context words that occur very frequently in the corpus (mostly closed class items). Almost all of them are among the 100 most frequent words in the corpus; and the negative cues by themselves even derive entirely from the 25 most frequent words. This can explain the graceful decline of Distributional Usefulness when the context lexicon is gradually reduced to the more frequent words (cf. Figure 4-10 on p. 119). In order to constitute a strong cue to a particular category, a context word needs to co-occur frequently with a substantial portion of category members (for a positive cue) or with an even larger portion of nonmembers (for negative cues). Therefore, less frequent context words might simply not have enough instances in the corpus in order to meet these conditions. At the same time, these considerations suggest that it is more likely to find a less frequent context word to constitute a strong positive cue when the target word category is relatively small. And, indeed, among all positive discriminators listed in

Appendix D, the least frequent context words were found in cues to interrogative words and prepositions which are, together with conjunctions, by far the smallest of the 11 benchmark categories.

The final conclusion addresses the original question by which the current subsection was introduced. How can the sets of individual discriminators that were just identified for the 11 benchmark categories explain the obtained pattern of Distributional Usefulness? When the category is held constant, its Distributional Usefulness scores that were observed separately in each of the four context positions (cf. Figure 4-14 on p. 129) indeed seems to be predicted fairly well by the number and degree (in terms of relative preference and discriminative power) of the individual positive and negative cues that were found within a particular context position (cf. Appendix D). This can be confirmed especially for categories whose Distributional Usefulness pattern across context positions is rather skewed (e.g., interjections, verbs, interrogative words, and conjunctions).

However, when attempting to explain a category's overall Distributional Usefulness level (cf. Figure 4-1 on p. 94) by its total number and degree of discriminators, a satisfactory answer turns out to be harder to formulate than one may have expected.[143] There is a trend that categories with a relatively high overall Distributional Usefulness score have at least *some* strong cues; but there are also cases that clearly violate this trend. For instance, compared to determiners, conjunctions have more and stronger positive and negative discriminators (both in terms of relative preference and discriminative power; cf. pp. 256, 258) but determiners achieve a clearly higher Distributional Usefulness level. It is for examples like this, that reliable predictors of overall Distributional Usefulness are difficult to be formulated in terms of the categories' individual discriminators. A more promising approach would therefore consider the interaction between individual cues (cf. the related discussion on p. 156).

As a very preliminary exploration in this direction, the relation between Distributional Usefulness and strong discriminators was assessed in terms of cumulative relative preference.[144] At this level, the strongest positive discriminators were the following (cf. the shaded cells in Table 4-6 on p. 144): **Interjections** are, on average, 63.2% more likely to occur in utterance-initial position than are non-interjections.

[143] In part, this may have to do with the dependence of discriminative power values on category size (cf. p. 142).

[144] This approach does not actually consider interactions but at least, it allows correlated weaker cues to add up and show some of their joint usefulness.

Determiners are immediately succeeded by a noun 49.8% more frequently than are non-determiners.[145] Conversely, **nouns** are immediately preceded by a determiner 47.5% more frequently than are non-nouns. The average probability of **prepositions** to be succeeded by a determiner is 44.2% higher than that for non-prepositions. Simultaneously, in the context position [+2], they co-occur with nouns 30.6% more frequently than do non-prepositions.[146] And finally, **interrogative words** are 36.1% more likely to occur in utterance-initial position and 34.7% more likely to be immediately followed by a verb form than are non-interrogative words.

Negative discriminators are generally less pronounced than the positive cues. The strongest ones at the level of context word categories were the following (cf. the shaded cells in Table 4-7 on p. 145): On average, **determiners** and **prepositions** are by 20.4% and 21.2% (respectively) less likely to occur utterance-finally than are other words. And **interjections** immediately succeed a determiner 20.0% less frequently than do non-interjections.

These strongest discriminators at the cumulative level are cues to only five different categories. The interesting observation is now that these five categories all achieve higher Distributional Usefulness scores than any of the remaining categories **verb**, **adjective**, **adverb**, **pronoun**, **conjunction**, and **particle** which do not have cues with the same cumulative relative preference.[147] This suggests that, as a very preliminary rule of thumb, one or two sufficiently strong cues at the cumulative level suffice for a category to achieve relatively high Distributional Usefulness.

4.3.4 Implications: Positive and negative cues

One striking observation across all target word categories is that the discriminators with the greatest discriminative power are generally positive cues while at the same time the majority of positive cues actually have a negative discriminative power. Negative cues, by contrast, all appear to be informative for the respective category as a whole, indicated by their positive discriminative power. The explanation for this finding is that positive and negative cues are intrinsically related in two complementary ways. And these follow from statistical rather than linguistic facts.

---

[145] But as was pointed out in the second general observation (pp. 151f), cumulative relative preference values from co-occurrences with nouns probably overestimate the joint usefulness from these cues, unless some rudimentary noun category is already available to the distributional learner.

[146] See footnote 145 above.

[147] Note that Distributional Usefulness for determiners is only slightly higher than that for verbs.

First, a very strong positive cue to a category directly gives rise to a whole set of negative cues to the same category. The most obvious example is the very strong positive cue that interjections receive from occurring utterance-initially. In fact, interjections do this so frequently (73.9%, on average) that there are only relatively few interjection tokens left that may enter other kinds of co-occurrences in context position [−1]. And indeed, the interjection category does not have any other salient distributional properties in this context position. Rather, the preference of interjections to co-occur with a particular context word in [−1] is necessarily quite low for most context words. And the more frequent context words among these are very likely to constitute negative cues to the interjection category since they often occur before words other than interjections. For verification of this prediction, recall that for interjections — as well as for any other category — the dominant negative cues across all four context positions only involve the 25 most frequent context words and utterance boundary markers (which are even more frequent), whereas strong positive cues can also arise from less frequent context words (cf. p. 151). The effect can also be verified in Table 4-7 (p. 145) and Appendix D (p. 249): Interjections have the largest number of dominant negative discriminators in context position [−1]. Likewise, nouns and determiners have the largest number of dominant negative discriminators in context position [−1] and [+1], respectively (cf. pp. 251, 256), because these are the positions in which they also receive their strongest positive cues, at least at the cumulative level.

The second way by which positive and negative cues are intrinsically related is more indirect. A strong positive cue to one category is likely to, by itself, constitute a negative cue to other categories; and it seems that most of the strong negative cues to any category arise in this fashion. This effect can be observed, for instance, for co-occurrences with various determiners in the context position [−1], and with several pronouns in [−1] and [+1] which are highly preferred by nouns and verbs, respectively: At the level of individual context words, these three types of co-occurrence relations are the most repetitive ones among the negative cues to any category (cf. Appendix D, pp. 250f). Apparently, determiners (as context words) occur so often right before nouns (as target words) that they do not have a sufficient amount of tokens left to also frequently occur in the same context position relative to other target words (and likewise for pronouns as cues to verbs). However, this effect is not as straightforward as the first one. If determiners occurred much more frequently overall, they could be as highly preferred by nouns in [−1] and still have many tokens left that can enter co-occurrences in the same context position with target words from other categories.

Whether or not a strong cue to one target word category Γ_1 becomes a negative cue to another category Γ_2, essentially depends on the sizes and frequency distributions of Γ_1 and Γ_2, and on the base frequency of the particular context word that constitutes this cue. And it is for these dependencies that the second relation between positive and negative cues is merely an indirect effect.[148]

When both kinds of effects operate together, they tend to result in particularly strong negative cues. To pick up the earlier examples, interjections strongly prefer to occur utterance-initially and are therefore less likely to be preceded by any lexical context word (the direct effect). At the same time, nouns are fairly likely to be preceded by a determiner such that other categories might be less likely to be preceded by a determiner (the indirect effect). Putting both effects together, interjections should be particularly unlikely to be preceded by determiners; and this prediction is confirmed at the level of individual determiners (cf. p. 249), and also in terms of cumulative relative preference (cf. Table 4-7 on p. 145).[149]

Of course, each of the negative cues mentioned above is more easily accounted for by a linguistic explanation. However, what I mean to illustrate by this more technical line of argument is the fact that, these negative cues are highly likely, if not inevitable, to occur, given the specified positive cues. Distributional cues are intrinsically related to each other, and not just loose and unstructured collections of information. In other words, even if there were no obvious linguistic reasons for the fairly low probability of interjections to be preceded by a determiner, this probability should be low anyway.

Having established these two relations between positive and negative cues, I am now in a position to address the general observation with which this current subsection started out. Negative cues to a category generally have a positive discriminative power and are thus quite informative about the overall category, whereas most positive cues are not. A prominent example of such positive cues is, once again, that of nouns being preceded by determiners. Although at the cumulative level, almost all nouns (except for some proper names) show this strong preference (cf. Table 4-11 on p. 177), they diverge considerably as to which particular determiners they prefer to co-occur with, due to agreement with respect to gender, case, and number. For instance, feminine singular

[148] This reasoning does not apply to the first relation between positive and negative cues because SCO vectors are standardized relative to target words and not relative to context words. The difference is that the first relation mainly depends on the available tokens of target words of the given category while the second one also depends on the available tokens of context words underlying the given cue.

[149] For the individual determiners, recall that *das* is also a determiner, though it was classified as a pronoun.

nouns frequently occur right after *die* but hardly ever after *das* while the contrary is true of neuter singular nouns (cf. Table 4-5 on p. 140). Both cues are *parallel cues* (rather than correlated cues) in that they (i) instantiate the same abstract combinatorial relation *DET N*, but (ii) are specialized on different subclasses of nouns and therefore almost mutually exclusive. Therefore, despite the impressive co-occurrence relation of nouns with determiners at the cumulative level, no single determiner is actually by itself a strong cue to the noun category as a whole.

This insight was stated before (p. 141), but it now directly leads to the following important result: The individual determiners may constitute parallel cues that are by themselves not very useful about the overall noun category; but what all these individual cues do have in common is that the determiners frequently occupy the [−1] context position relative to some nouns. And by doing so, all determiners contribute — by the direct effect described above — to the same kinds of negative cues to nouns in this context position.

Thus, although it may initially have appeared as if positive cues with negative discriminative power are *in principle* not useful — or even harmful — for the overall category, their potential contribution via negative cues demonstrates that this is not necessarily the case.

Additionally, the overall category might benefit from such positive cues even more directly, namely, via interaction. Suppose, some specific category has a number of positive cues, each of which singles out only some subset of category members that frequently enter the relevant co-occurrence relation that defines this cue, while other category members, like most nonmembers, do so hardly ever. If these subsets of category members always take the same boundaries across cues, with several cues possibly favoring the same subsets while others single out completely different subsets, the overall cohesion of category members in the SCO vector space might indeed be at stake. If, however, the subsets of different cues partly overlap and tend to take different boundaries, most members would share some of these cues that set them apart from most nonmembers. The positive cues with negative discriminative power would still shape the internal structure of the overall category, but as a whole, it would separate fairly well from most other categories.

Cues interact, and in a constellation such as the hypothetical one that was just described, their interaction provides crucial information that is not contained in any single cue in isolation. And psycholinguistic evidence suggests that this extra

information from partially correlated cues might indeed solve certain puzzles of category acquisition (cf. the general discussion 5.2).

Let me conclude this section by emphasizing the importance of negative cues. Positive cues are certainly the concept that first comes to mind when one is thinking of distributional cues because they consist of context words which the members of a category Γ strongly prefer to co-occur with and which therefore more naturally characterize this category. Negative cues, by contrast, are missing or underrepresented properties of Γ, relative to other categories, and thus appear to say more about these other categories than about Γ itself.

This view of positive cues as the primary type of distributional evidence may be challenged by the finding that many positive cues to a category are parallel cues and achieve much of their overall informativeness via negative cues (to the same category). It is, of course, a very different question to what extent and in which ways children exploit either type of cue.

4.4 Categories under the microscope: Verbs and nouns

Building on the previous section, the distributional properties of the two major categories noun and verb are analyzed at greater detail in the following. These two categories were chosen here for three reasons. First, they are arguably the most fundamental categories in German and most other languages. Second, they are by far the largest categories among the target words derived from the *Leo* corpus. And third, they illustrate very well — and especially by the contrasts between them — how a category's distributional profile arises from the combination of grammatical regularities and usage preferences. It would be desirable to conduct similar analyses for the other benchmark categories as well, but this would go beyond the scope of this dissertation.[150]

As a starting point, let me briefly recap what the study brought to light so far about nouns and verbs. First, according to the default analysis (cf. Figure 4-1 on p. 94), distributional information is substantially less informative about the verb category than

[150] One particularly interesting candidate pair of categories would be adjectives and adverbs, because they have almost the same size, they share certain distributional commonalities, and both probably have some interesting substructure defined by semantic (e.g., temporal adverbs, locative adverbs, color adjectives) and grammatical distinctions (e.g., inflected vs. unmarked adjectives, which are associated with attributive vs. predicative usage, respectively).

about the noun category. Second, the earlier confusability analysis (cf. Table 4-1 on p. 101) showed that verbs separate rather poorly from almost any other category whereas nouns separate very well from most categories. And finally, the results in the previous section revealed that no single distributional cue to the verb category achieves particularly high discriminative power whereas the noun category receives some fairly strong cues (cf. pp. 147f; Appendix D, pp. 250f).

What is causing these divergent findings for nouns and verbs? In particular, what are the underlying linguistic reasons? To approach these questions, it is useful to first study the internal structure of both categories in the SCO vector space. In particular, do nouns and verbs form an Intermediate Scenario or rather a Hybrid Scenario with several isolated clusters (cf. 3.3.1)? This is precisely the question for which the two alternative evaluation scores Global Coherence and Local Coherence were introduced (cf. 3.3.4). Recall that both scores should be used only when comparing categories of roughly the same size. Since there are 288 verb forms and 268 nouns in the target lexicon, this condition is met here.

As it turns out, Global Coherence is much higher for nouns (.66) than for verbs (.27) whereas both categories achieve relatively high Local Coherence scores (73.7 for nouns, 67.7 for verbs). This indicates that the noun category is fairly coherent in its entirety, looking somewhat like an Intermediate Scenario with a single coherent core cluster and fuzzy boundaries. The verb category, by contrast, is best depicted by the Hybrid Scenario which means that verbs tend to form several clusters which are by themselves quite coherent, but located in rather distant regions in the SCO vector space.[151] This is a very important result, for it entails that, even though the distributional model fails to provide very useful cues about the verb category as a whole, it does pick up some substantial substructure.

But what exactly does this substructure consist of? And in particular, do the isolated verb clusters correspond to some meaningful linguistic distinctions within the verb category? These questions are addressed in subsection 4.4.1. Analogous investigations were also carried out for the noun category to reveal why it does not form a Hybrid Scenario like the verb category (subsection 4.4.2). The implications of these

[151] This divergent result for nouns and verbs is consistent with observations by Redington et al. (1998) and Mintz et al. (2002) for English. In their default analysis, Redington et al. found verbs, including auxiliaries, to substantially contribute to five distinct clusters whereas virtually all nouns were found in one large cluster. And in the various analyses conducted by Mintz and colleagues, the general pattern was similar in that verbs inhabited by far more clusters than did nouns.

findings, especially in regard to language acquisition, are discussed in the final subsection.

4.4.1 The distributional structure of the verb category

It turns out that the verb category essentially partitions into two large clusters which correspond to the two grammatical classes of nonfinite verb forms (infinitives and past participles) and finite verb forms (imperatives, first person singular forms, second person singular forms, etc.).[152] This becomes apparent when each of these two classes is treated as a category by itself which yields three kinds of Distributional Usefulness scores that are displayed in Figure 4-15.

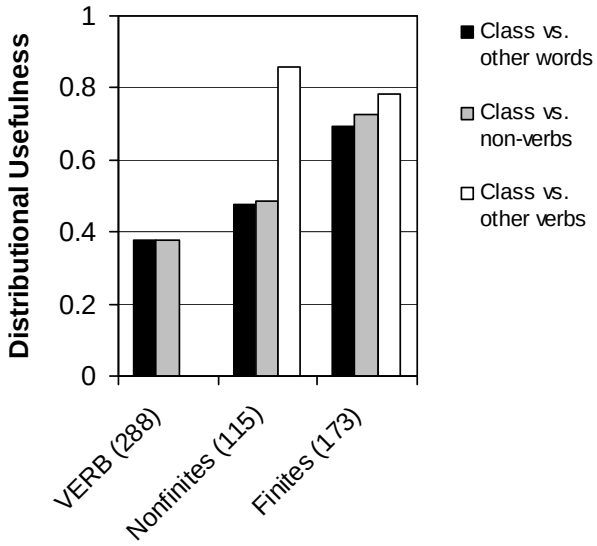

Figure 4-15: *Separation and performance of finite and nonfinite verb forms*

Three kinds of Distributional Usefulness scores are shown for the verb category and its two major classes: nonfinite forms and finite forms. These scores quantify how useful distributional information is to separate the given class from all target words outside that class (be they verbs or non-verbs; black), to separate it from all non-verbs (gray), or to separate it from all verbs outside the class (white). Note that for the full verb category, the third value is not defined, whereas the first two values are necessarily identical. Class sizes are given in parentheses.

[152] Note that German is equipped with richer verb morphology than English, with unique forms in most inflectional subclasses, despite some degree of ambiguity between some of them. In coding verb form target words for their grammatical class, these ambiguities were resolved in the same way as for the major benchmark categories (cf. 2.2.2). Note further that, in general, present participles are used much less frequently in German than in English. In the corpus, no single present participle occurred often enough to enter the target word lexicon.

The first score specifies how well nonfinite (or finite) verb forms separate in the SCO vector space from any other words — that is, from both finites (nonfinites, respectively) and non-verbs. It thus quantifies how easy the class of nonfinite (finite, respectively) verb forms would be to discover among all target words. The second score assesses how well nonfinites (finites) separate only from non-verbs while the third score evaluates the separation of nonfinites (finites) from finites (nonfinites, respectively). For comparison, the Distributional Usefulness score for the overall verb category is also shown.

The first striking observation to make is that distributional information is clearly more useful for discovering each of the two classes (.48 for nonfinite forms and .69 for finite forms) than it is to discover the global verb category (.38). Surprisingly, however, nonfinite verb forms separate very well from finite forms (.86); in fact, they do so much better than they separate from non-verbs (.48). Finite verb forms do separate fairly well from non-verbs (.72) but also separate even better from nonfinite verb forms (.78). These findings indicate that nonfinite and finite verb forms occupy two unique and distant regions. The space between them, however, is not empty but filled by the bulk of non-verbs which partly overlap with nonfinites but much less so with finites.

In Figure 4-16 below, this pattern can also be observed in the same two-dimensional projection of the SCO vector space that was shown before (Figure 3-5 on p. 65). In this projection, nonfinites (circles) and finites (triangles) form two clusters which essentially do not overlap and only barely touch, even though both clusters cover fairly large regions by themselves. Most non-verbs are located either in the area between these two clusters, or they overlap with nonfinite verbs — an observation which is fairly representative of the situation in the full SCO vector space, since finites (.72, see above) separate much better from non-verbs than do nonfinites (.48, see above). Crucially, from such a topographic configuration, no reasonable learning algorithm would assign nonfinite and finite verbs to the same category — unless virtually all target words would be assigned to one single category, resulting in a trivial and useless category system.

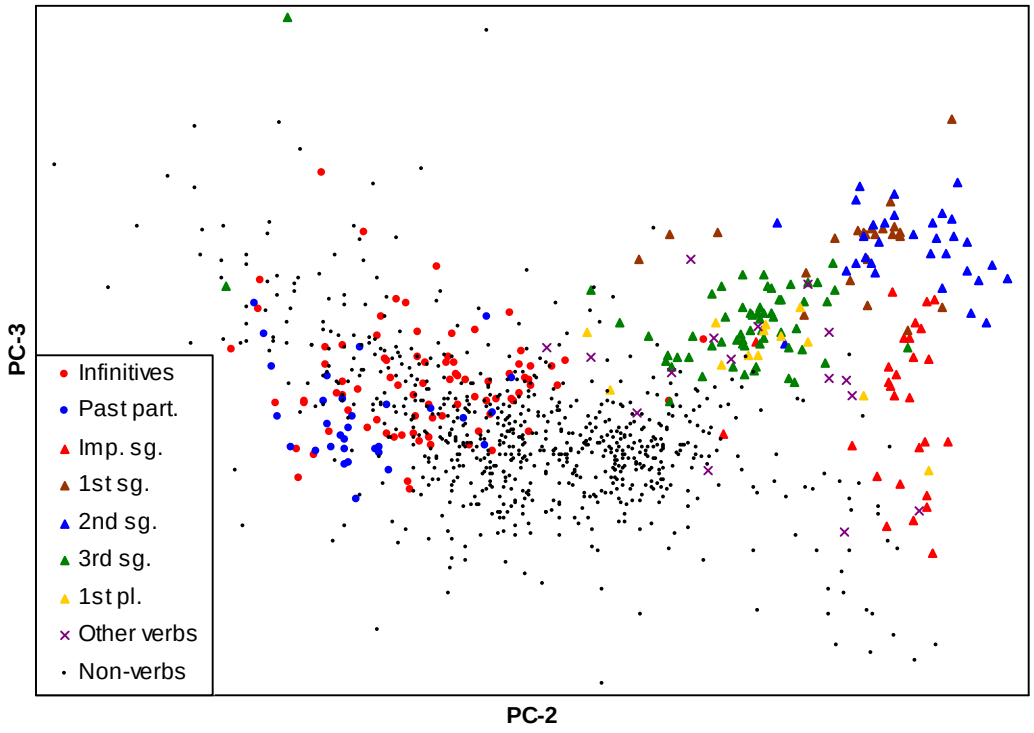

***Figure 4-16**: Verb subclasses in the SCO vector space (two-dimensional projection)*

Grammatical subclasses of the verb category, shown in the second and third principal component (PC-2 and PC-3) of the full SCO vector space. Vectors are marked for their grammatical subclass (with all finite subclasses being displayed as triangles, and all nonfinite subclasses as circles). Subclasses with less than 10 members are summarized as *other verbs*.

The two-dimensional projection further suggests that, within the cluster of finite verb forms, a number of subclusters can be identified that roughly correspond to the individual finite subclasses: imperative singular forms (imp. sg.), first person singular forms (1st sg.), second person singular forms (2nd sg.), etc.[153] And indeed, the existence of these subclusters is confirmed for the full high-dimensional SCO vector space by Figure 4-17 which plots, at the level of individual finite (and nonfinite) subclasses, the

[153] It is worth to mention here that imperative verb forms are treated as being nonfinite by some linguists because they only inflect for number but not for person (e.g., Eisenberg, 2000, Vol. 2: 100f). Others view *imperative* as a verb mode for which inflection for person exists in principle, but all forms other than second person singular and plural are simply missing (e.g., Bußmann, 1990:325). For the current study, I followed this second view. But nothing crucial hinges on this decision; it merely helps to keep terminology simple when discussing the findings in the current section, since imperative singular forms cluster together with the finite verb forms. But independently of such practical considerations, this constitutes another example of how the distributional approach might support the investigation of grammatical classification problems (cf. p. 108). The empirical fact that imperatives are distributionally by far more similar to finite than to nonfinite verb forms can be taken as independent, though certainly not critical, support for the view that they are finite themselves.

same three types of Distributional Usefulness values that were computed earlier for the two superordinate classes (cf. Figure 4-15).[154]

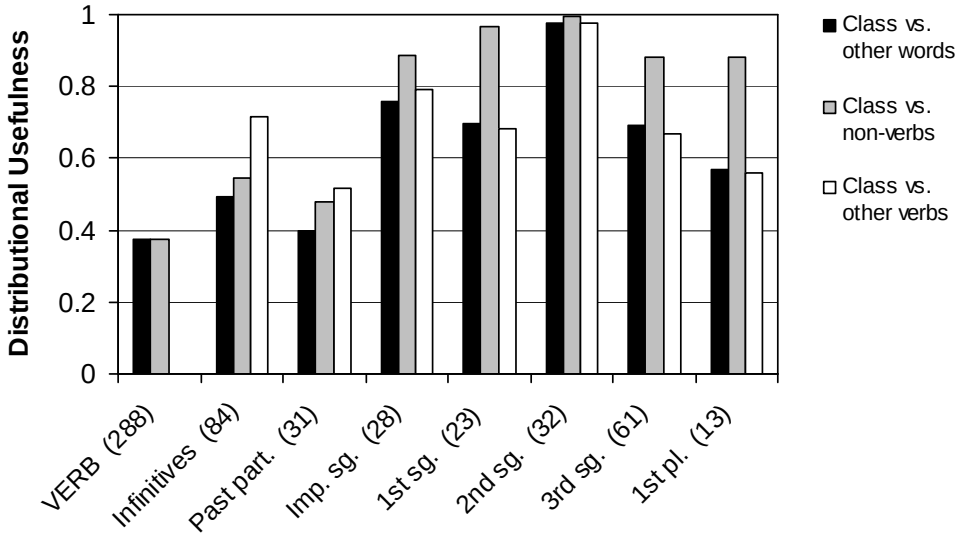

Figure 4-17: *Separation and performance of verb subclasses*

Three kinds of Distributional Usefulness scores are shown for the verb category and its basic grammatical subclasses. As before, these scores quantify how useful distributional information is to separate the given subclass from all target words outside that subclass (be they verbs or non-verbs; black), to separate it from all non-verbs (gray), or to separate it from all verbs outside the subclass (white). Only subclasses with more than 10 members are shown. Subclass sizes are given in parentheses.

Like the superordinate class of all finite forms, each of the finite subclasses would be much easier to discover than the global verb category (Distributional Usefulness values ranging from .57 to .98, compared to .38 for the full verb category; black bars in the chart). Moreover, each individual finite subclass separates almost maximally from all non-verbs (values ranging from .88 to 1.00; gray bars) and thus by far better than does the superordinate class of all finites (.72, see above). This implies that the superordinate class spreads across a much larger region in space than each of its individual subclasses. This together with the fact that the superordinate class separates

[154] Further investigations revealed that other inflectional properties such as tense or aspect have hardly any influence on the distributional structure of the verb category. Likewise, verbs are not organized by their valency properties. There is, however, some substructure reflecting functional verb class distinctions (full verb, auxiliary, modal verb, or copula verb) and semantic similarity. But these aspects are clearly predominated by the grammatical features finiteness, person, and number which characterize the major verb clusters.

fairly well from all words outside this class (.69, see above) implies that the individual finite verb subclasses occupy adjacent regions in the SCO vector space.

But some of these adjacent regions partly overlap. This follows from Table 4-8 which presents confusability analyses (similar to those in subsection 4.1.2) between the individual subclasses of the verb category. The separation values between any two finite subclasses (i.e., the lower-right 5x5 table cells) are fairly high, ranging from .70 to 1.00, but in most cases still clearly below 1.00. The only exception is the subclass of second person singular verb forms which practically does not overlap with any other verb subclass, nor with any non-verbs (cf. Figure 4-17 above; gray bar).[155] To some extent, this pattern of adjacent and partially overlapping subclusters of finite verb subclasses can also be observed visually in the two-dimensional projection (Figure 4-16 above) but the degree of overlap appears higher here than in the full SCO vector space.

Table 4-8: *Pairwise separation between verb subclasses*

Subclass Γ_1	**Subclass Γ_2**						
	Infinitives	Past part.	Imp. sg.	1st sg.	2nd sg.	3rd sg.	1st pl.
Infinitives	—	.54	.97	.96	.96	.89	.79
Past part.	.41	—	.99	.96	.99	.92	.98
Imp. sg.	.99	1.00	—	.77	.89	.89	.90
1st sg.	.99	1.00	.74	—	.95	.73	.94
2nd sg.	1.00	1.00	1.00	1.00	—	.98	.99
3rd sg.	.89	.94	.89	.70	.78	—	.74
1st pl.	.64	1.00	.90	.91	.86	.80	—

Note. Table cells specify Distributional Usefulness of verb subclass Γ_1 when the target lexicon is restricted to members of Γ_1 and Γ_2 , thus quantifying how useful distributional information is to distinguish Γ_1 from Γ_2. Only subclasses with more than 10 members are included. The two nonfinite subclasses are set off from the finite subclasses by horizontal and vertical space.

[155] It is worth pointing out that those pairs of finite subclasses that overlap the most in the SCO vector space — first person singular forms and imperative singular forms on the one hand, and first person singular forms and third person singular forms on the other hand (cf. Table 4-8) — also are the ones with the highest proportion of ambiguous verb forms between them (ranging from 29.5% to 92.9%). Lexical ambiguity — in this case more precisely: *syncretism*, i.e., homonymy between inflected forms of different grammatical specification — also explains why the separation value between first person plural forms and infinitives is clearly lower than between any other pair of finite and nonfinite subclasses. In fact, considering that 98.8% of all target words that were classified as infinitives are homonymous with the corresponding forms for first person plural, and 61.5% of all target words classified as first person plural forms can also be used as infinitives, their mutual separation values even appear rather high.

The projection further suggests that there are only very few non-verbs intruding the overall region filled by finite verb forms; and the fact that each of the individual finite subclasses separates extremely well from all non-verbs (cf. Figure 4-17) confirms this for the full SCO vector space. However, because the overall region of finite verbs is so large, finite verbs close to the boundary of this region are closer to many non-verbs than they are to other finite forms located in the opposite side of this region. This is why the superordinate class of all finite verb forms does not separate quite as well from non-verbs than do the individual finite subclasses, as was noted above. This suggests that if all finites together would occupy a smaller region — for instance, the region currently mainly inhabited by first person singular verb forms — their separation from non-verbs would undergo a sizable boost.

In contrast to finite verb forms, the cluster of nonfinite verb forms does not appear to consist of linguistically interpretable subclusters. The two nonfinite subclasses — infinitives and past participles — do not separate very well from each other (separation values .54 and .41, cf. Table 4-8 above); that is, they occupy largely overlapping regions in the SCO vector space which is also reflected, though to an exaggerated extent, in the projection (Figure 4-16 on p. 161).[156] Like their superordinate class of all nonfinite forms, infinitives and past participles separate better from all other verb forms (separation values ranging from .79 to .99, cf. Table 4-8) than from non-verbs (.55 for infinitives and .48 for past participles, cf. Figure 4-17 on p. 162).

But what kinds of non-verbs are distributionally so very similar to the nonfinite verb forms? Additional confusability analyses revealed that nonfinite verbs in fact separate very well from six of the 10 non-verb categories (separation values ranging from .88 to .97) but not quite as well from nouns (.70), adjectives (.64), particles (.61), and especially adverbs (.46). Hence, in the SCO vector space, these four latter categories overlap to some extent with nonfinite verb forms. While the other six non-verb categories are not problematic for the nonfinite class, they are for the overall verb category (cf. Table 4-1 on p. 101) which suggests that these six categories fill the space between nonfinite and finite verb forms.

Putting the pieces together, these analyses revealed a peculiar topography of the verb category, and they imply that the low Distributional Usefulness for the global verb

[156] Note that there is very little syncretism between infinitives and past participles. Only 9.7% of all target words classified as past participles are homonymous with the corresponding infinitive form, and not a single target word classified as an infinitive can be used as a past participle. The overlap of infinitives and past participles in the SCO vector space thus has to arise from genuine distributional causes.

category essentially arises for three reasons. First and most importantly, nonfinite and finite verb forms are distributionally very different from each other such that they form two distinct clusters in the SCO vector space, with many non-verbs occupying the space between them. The second reason only concerns nonfinite verb forms, as these are very similar to many non-verbs, most prominently to adverbs. The third reason in turn only involves finite verb forms and is much less problematic than the first two: The whole class of all finite verb forms already separates very well from non-verbs, but this separation would be even much higher if the individual finite subclasses were more similar to each other such that all finites together would occupy a much smaller region in the SCO vector space.

The next step therefore was to identify the properties of verbs that are causing these problems. This means to ask (i) *Why are nonfinites so different from finites?*; (ii) *Why are they at the same time so similar to some of the non-verb categories?*; and (iii) *What distinguishes the individual finite subclasses from each other?* On the following pages, these three questions are addressed in terms of the subclasses' distributional profiles. Where possible, the relevant findings are linked back to the underlying causes in terms of linguistic structure and usage preferences. As in the previous section, these are often presented together with specific examples which are all taken from the corpus.

Table 4-9 below shows the distributional profiles of all substantial subclasses of the verb category, summarized by context word category. Their most salient distributional properties at the level of individual context words can be found in Appendix E. In the following, I freely interpret the numbers from both sources without always providing explicit references to either one of them.

To begin with question (i), it becomes immediately obvious that the profiles of the nonfinite subclasses are highly incompatible with those of the finite subclasses; and this directly explains why all cues to the overall verb category were found to be fairly weak (cf. pp. 147, 250). Infinitives and past participles — the two nonfinite classes — share distributional properties in all four context positions. In particular, the most salient preferences of both are to occupy the last two positions of an utterance — regardless of the type of utterance termination — which reflects their typical positions in simple questions and main clause declaratives (for the final utterance position), and in tag questions and subordinate clauses (for the last but one position). These properties distinctly set nonfinites apart from the various subclasses of finite verb forms which occur in these positions much less frequently.

Table 4-9: *Distributional profiles of verb subclasses (cumulative summary)*

Left context				Verb subclass [a]	Right context			
[−2]		[−1]			[+1]		[+2]	
17.4	PRON	20.9	ADV	Infinitives (84)	45.8	<Bnd>	32.3	<Bnd>
17.3	DET	20.1	PTCL		11.1	PRON	16.4	PRON
13.3	PTCL	15.1	N		8.4	V	10.2	DET
22.8	DET	22.2	N	Past participles (31)	50.6	<Bnd>	44.7	<Bnd>
18.3	PRON	22.1	ADV		13.5	V	11.6	PRON
14.4	V	18.5	PRON		6.3	CONJ	10.8	V
46.5	<Bnd>	38.1	<Bnd>	Imperative sg. (28)	43.3	PTCL	26.2	PTCL
18.8	INTJ	15.6	CONJ		31.6	PRON	18.3	ADV
6.1	V	12.4	PRON		7.5	DET	14.4	<Bnd>
42.6	<Bnd>	57.5	PRON	1st singular (23)	55.6	PRON	25.6	PTCL
18.7	INTJ	11.1	<Bnd>		11.4	PTCL	17.4	ADV
7.7	PTCL	8.2	ADV		10.8	ADV	13.8	V
44.5	<Bnd>	27.4	PRON	2nd singular (32)	68.4	PRON	22.6	PTCL
9.6	INTJ	24.9	<Bnd>		7.6	PTCL	20.9	ADV
9.1	CONJ	15.3	INTG		6.3	<Bnd>	15.4	PRON
38.9	<Bnd>	24.4	PRON	3rd singular (61)	33.4	PRON	20.8	PTCL
11.2	INTJ	13.9	INTG		14.7	PTCL	17.1	ADV
10.3	DET	12.7	ADV		14.4	DET	12.6	<Bnd>
36.3	<Bnd>	25.8	PRON	1st plural (13)	54.9	PRON	27.6	PTCL
12.2	INTJ	16.1	ADV		11.2	PTCL	23.4	ADV
9.2	V	15.6	<Bnd>		8.6	DET	14.0	DET

Note. The distributional properties of each verb subclass (central column) are summarized by context word categories. The symbol <Bnd> represents all four utterance boundary markers. Only the three most likely context word categories are shown in each context position. The percentages next to them specify cumulative preference values which estimate the average conditional probability at which members of the verb subclass co-occur with any word of the context word category in the respective context position. Probabilities above 30% are shaded.

[a] Size of subclass is given in parentheses.

At the same time, all finite subclasses strongly prefer to occur in utterance-second position which they do most typically in main clause declaratives and *wh*-questions.[157] Finites also frequently occupy the utterance-initial position — which they do in *yes/no*-questions and several kinds of elliptic utterances with the subject being omitted — but in this case, their precise preference values (ranging from 10.8% to 38.1%) vary considerably across subclasses. In any case, nonfinites occupy these two utterance-initial positions much less frequently.[158]

[157] This interpretation also applies to imperative singular forms since 92.9% of them are homonymous with the corresponding first person singular forms. But even when used in their imperative function, they often occur in utterance-second position, after a conjunction, interjection, adverb, or pronoun.

[158] Infinitives occupy these positions more frequently than do past participles. The reason is that nearly all (98.8%) of the verb forms classified as infinitives are homonymous with the corresponding forms for first and third person plural. But given this exceedingly high degree of syncretism, it is remarkable how clearly these co-occurrences distinguish infinitives from first person plural forms, for the latter occupy the two first utterance positions roughly three times as often as do infinitives.

A third important difference between the two superordinate classes concerns co-occurrences with pronouns. While all finite forms strongly prefer to be immediately preceded or followed by a pronoun, infinitives and past participles enter these co-occurrence relations much less frequently overall — even though for some individual pronoun forms, this difference is not very pronounced.[159] The following facts are important to make sense of these findings. In the vast majority of cases in the corpus, when finite verb forms co-occur next to a pronoun, this is a subject pronoun form.[160] The default subject position in German declaratives is in the *Vorfeld*, that is, right before the main verb, as for English. However, when this first position is occupied by some other constituent — this could be an adverb or even a full PP or object NP — the subject moves into the position right after the finite verb. Such inversion also takes place for the majority of questions: virtually always for *wh*- and *yes/no*-questions, but not generally for tag questions. In sum, inversion is the main cause for the high preference values of finite verb forms to be followed by a subject pronoun.

Co-occurrences with utterance boundary markers and pronouns thus constitute the most influential — though not the only — kinds of cues that together so sharply distinguish nonfinite from finite verb forms. In fact, some of these differences are so pronounced that the strongest positive cues to nonfinites are also the strongest negative cues to finites, and vice versa, as additional analyses in terms of relative preference revealed (not shown here; cf. 4.3.3). The way how these different preferences of nonfinites and finites relate to particular utterance-level constructions (see the brief hints above) suggests the conclusion that these crucial differences between nonfinites and finites are a direct consequence of their syntactic privileges. Finite and nonfinite forms simply fill different kinds of syntactic slots in constructions.

[159] This preference is least salient for imperative singular forms; and the fact that they show it at all arises partly from their high degree of syncretism with first person singular forms (cf. footnote 157 above), and in part from a peculiarity of spoken German where imperatives occasionally do take a subject pronoun. Infinitives are followed by pronouns more frequently than one might expect; but the majority of these co-occurrences concerns the subject pronoun *wir* (English: *we*) which again is a trace of the high degree of syncretism between infinitives and first person plural forms (cf. footnote 158 above). By contrast, the surprisingly high preference of past participles to be preceded by pronouns is most likely not accounted for by syncretism — even though 16.1% of all past participles are homonymous with the corresponding third person singular form, and 9.7% are homonymous with the corresponding first person plural form. More relevant are structural reasons: A subject (nominative) pronoun often appears right before a past participle in questions, and object pronoun forms (dative or accusative) are likely to do this in almost any utterance construction.

[160] The only exception is the subclass of imperative singular verbs which are about equally likely to be followed by subject pronouns and by other pronouns. But for context position [−1], imperatives show the same strong bias towards subject pronouns as do the other finite verb subclasses.

This answers why nonfinites are so different from finites, which was the first of the three questions (p. 165). To turn to the second question — why are nonfinite verb forms so similar to adverbs and to a lesser degree also to particles, adjectives, and nouns — consider the distributional profiles of these four benchmark categories (cf. Table 4-4 on p. 134, and Appendix D). All four have as their most salient properties frequent occurrences in the two final utterance positions; and at the same time, they take much less often the two utterance-initial positions. All four categories like to be preceded by verbs, most frequently by the verb form *ist* (English: *is*) and preferably in context position [−2]. Adjectives, adverbs, and particles further tend to be immediately preceded by (other) adverbs and particles. Finally, adverbs and particles — and to a lesser degree also adjectives, but not nouns — often co-occur with subject pronouns and some basic determiner forms (that can be used as subject pronouns) two words to their left.

Virtually all of these preferences also show up in the distributional profiles of infinitives and past participles (cf. Table 4-9 above) — with the only exception being that infinitives do not frequently co-occur with verb forms two words to their left — and each of these commonalities with adverbs, particles, adjectives, and nouns is instantiated by several examples at the level of individual context words (cf. Appendix D and Appendix E).

The kinds of constructions that might plausibly give rise to these preferences are very different ones for nonfinite verbs than for each of these major categories. It is therefore quite likely that a more powerful distributional method that uses a wider context window or even has access to some rudimentary information about phrase boundaries, would be much better at discerning nonfinite verb forms from these non-verb categories.[161] Support for this conjecture comes from Mintz et al. (2002) who found that access to even a vague notion of phrase boundaries can significantly improve the usefulness of distributional information about verbs — though it should be added that this offers only weak support because the investigated language was English, and the specific distributional reasons for the improvement were not determined.

Nevertheless, the distributional similarity between nonfinite verb forms on the one hand and adverbs, particles, adjectives, and nouns on the other hand, appears to be a

[161] Assuming the child to already be capable of roughly detecting phrase boundaries would not be circular in the given context; that is, this assumption would not presuppose that the child had already mastered a good deal of category acquisition. For it was found that there is a variety of prosodic cues that together provide a reasonably reliable basis for detecting phrase boundaries (cf. p. 22).

problem that could be considerably reduced, if not overcome, by fine-tuning the method. But it is hard to imagine how any modifications to the simple distributional model used here could also substantially reduce the extreme dissimilarity between finite and nonfinite verb forms. The differences between their most salient preferences will essentially not disappear when information about phrase boundaries is added; and, more importantly, such a modification will not introduce any relevant distributional features that finites and nonfinites actually have in common. The strong prediction to derive from this reasoning is that finite and nonfinite verb forms do not share a single strong distributional cue, no matter by which particular model the distributional properties are derived.

To complete the analysis of the distributional structure of the verb category, let me turn to the last of the three initial questions (p. 165). Why do the different finite subclasses form distinct, though partly overlapping, subclusters in the SCO vector space? In a nutshell, there are three fundamental reasons, namely grammatical agreement, usage preferences between possible constructions, and some lexical markers serving certain pragmatic functions. To illustrate how these factors indeed conspire to discriminate the individual finite subclasses, one prominent example — rather than an exhaustive list — for each factor is given below.

In German, as in principle also in English, a finite verb form and its subject agree in person and number. In particular, when the subject is a pronoun, different finite subclasses actually co-occur with very different kinds of subject pronouns.[162] While Table 4-9 masks this fact, Appendix E shows that each subclass co-occurs very frequently with the matching subject pronoun form in the context positions [−1] and [+1].[163] Co-occurrences of finite verb forms with subject pronouns thus constitute another example of parallel cues (cf. p. 156), and as such, they substantially contribute to differentiating the finite subclasses from each other.

The second important factor contributing to these distinctions concerns usage preferences. Maybe with the exception of some imperative forms, all finite verb forms

[162] There is only very little ambiguity in German between the sets of possible subject pronouns which are defined by number and person.

[163] For third person singular, there are actually multiple matching subject pronouns, but the most frequent ones are not the default pronouns *er*, *sie*, *es* (English: *he*, *she*, *it*, respectively), but rather *das*, *der*, and *die*, which are commonly construed as definite articles but may also be used as demonstrative subject pronouns. Also note that infinitives show some preference to co-occur next to *wir* (English: *we*) which only reflects their nearly perfect homonymy with the corresponding verb form for first person plural. Likewise, imperative singular forms like to occur next to *ich* (English: *I*) because most of them are identical with the corresponding first person singular verb form.

can in principle fill the same syntactic slots in the same kinds of constructions.[164] But due to pragmatic and other nonsyntactic factors, the individual subclasses differ considerably as to how frequently they actually occur in each of these possible constructions. For instance, in S data, speakers ask questions about the addressee rather than about themselves (with respect to *yes/no*-questions, see Cameron-Faulkner, Lieven, & Tomasello, 2003).[165] And from a pragmatic point of view, this is a very plausible bias, particularly when talking to a child. The distributional consequences of this bias are sizable: The most important effect with respect to *wh*-questions is that second person singular verb forms are much more likely to occur right after an interrogative word (on average 15.3%, cf. Table 4-9) than are first person singular verb forms (2.6%, not shown in Table 4-9). For the case of *yes/no*-questions, the most relevant distributional consequence is that second person singular verb forms are more likely to occur utterance-initially (24.9% vs. 11.1%, cf. Appendix E).[166] And both effects contribute to the impressive mutual separation values between first and second person singular verb forms (cf. Table 4-8 on p. 163).[167]

The third relevant factor mainly applies to imperative singular verb forms. This subclass is strongly associated with a set of pragmatic markers that can serve a variety of communicative functions, such as emphasizing or toning down a request. The most prominent one among these markers is the modal particle *mal* (which is a short variant of *einmal*, English: *once*). On average, imperative singular forms are followed by this single context word at an astonishing rate of 38.4% in context position [+1], and still very frequently (15.4%) in [+2] (cf. Appendix E). By comparison, other finite verb forms (1.5% and 2.8%, respectively), nonfinite verb forms (0.2% and 2.0%), and non-verbs (0.9% and 1.1%) are much less likely to enter these co-occurrence relations with

[164] Only two imperative singular forms have to be excluded here, since the other 26 (92.9% of all) are homonymous with the corresponding verb form for first person singular.

[165] In analyzing usage frequencies of item-specific sentence-level constructions, Cameron-Faulkner et al. (2003:857, Table 4) found across 12 corpora of English child-directed speech that, on average, 50% of all yes/no-questions were of the form *"AUX you …?"*, with AUX representing only four different auxiliary verbs. Most other *yes/no*-questions were asked in the third person, singular or plural.

[166] The fact that first person singular forms do not occupy the first utterance-position even less frequently, is a result of elliptic utterances where the subject pronoun is omitted, and from cases where the verb form is used as an imperative — 52.2% of all first person singular verb forms are homonymous with the corresponding imperative singular form.

[167] It is worth pointing out that another striking consequence of the skewed usage preferences of first and second person singular verb forms to be used as the main verb of *yes/no*-questions, is that the second person forms are more likely to be immediately succeeded by the respective default subject pronoun (60.6% vs. 37.4%), and at the same time less likely to be immediately preceded by this pronoun (22.6% vs. 48.7%). However, because these are two different pronouns (*du* vs. *ich*), the different percentages do not contribute to the formation of distinct subclusters for the two finite subclasses. But they nonetheless underscore the strong influence that usage preferences at the construction level have on distributional properties.

mal. A second important marker is the conjunctive adverb *dann* (English: *then*, with a causal and a temporal meaning). Imperative singular forms have a fairly high preference to occur right after *dann* (13.3%) which is clearly above the corresponding preference value for other finite verb forms (3.8%), nonfinites (2.0%), and non-verbs (0.2%). Together, these three unique co-occurrence preferences of imperative singular forms (*mal* in [+1] or [+2], and *dann* in [−1]) constitute strong and reliable positive cues to this special verb subclass.[168]

4.4.2 The distributional structure of the noun category

To obtain a corresponding distributional description of the noun category, I begin again by providing a clearer picture of the category's topography in the SCO vector space. The primary contrast between common nouns (238 target words) and proper names (30) is not well reflected in the SCO vector space. With regard to distributional information, both are rather confusable with each other (mutual separation values .40 and .32, respectively). Likewise, the distinction between animate nouns (63 target words) and inanimate nouns (205) also does not capture the vector constellation of nouns very well (mutual separation values .31 and .20, respectively). Compare these distinctions with the profound distributional contrast between finite and nonfinite verb forms observed before (mutual separation values .86 and .78, respectively; cf. Figure 4-15 on p. 159).

Nevertheless, there are some linguistic distinctions within the overall noun category that are captured fairly well in the SCO vector space. The model may not be very sensitive to the overall contrast between common nouns and proper names; but it does produce noun clusters that correspond to certain subclasses of these two noun classes. For common nouns, these subclasses are defined by gender and number, and the relevant proper name subclasses are names for individuals and names for places. Most of what follows is therefore formulated for the six subclasses feminine singular (60 target words), masculine singular (82), neuter singular (57), plural nouns (39), names for individuals (23; mainly people, toy animals, and cartoon figures) and names for places (7; mostly cities). A consistent grammatical description would further subdivide the class of plural nouns into three gender subclasses, paralleling the three gender subclasses of singular nouns. However, as will become apparent later (p. 181), the

[168] Discriminative power is positive for the two *mal* cues, no matter whether imperative singular forms are compared with other finite verb forms, nonfinites, or non-verbs. For the *dann* cue, discriminative power is positive when imperative singular forms are compared with non-verbs, but it becomes slightly negative for the comparison with other finite forms and nonfinites.

model is essentially insensitive to these gender distinctions among plural nouns, much in contrast to those among singular nouns.

Figure 4-18 below presents for each of these six noun subclasses the same three kinds of Distributional Usefulness scores that earlier guided the analysis of verb subclasses. Distributional information is about as useful for discovering the entire noun category (.56) as it is for discovering each of the six noun subclasses when they are treated as independent categories by themselves (scores ranging from .42 to .65; black bars in the chart), with the greatest deviation occurring for the smaller subclasses.

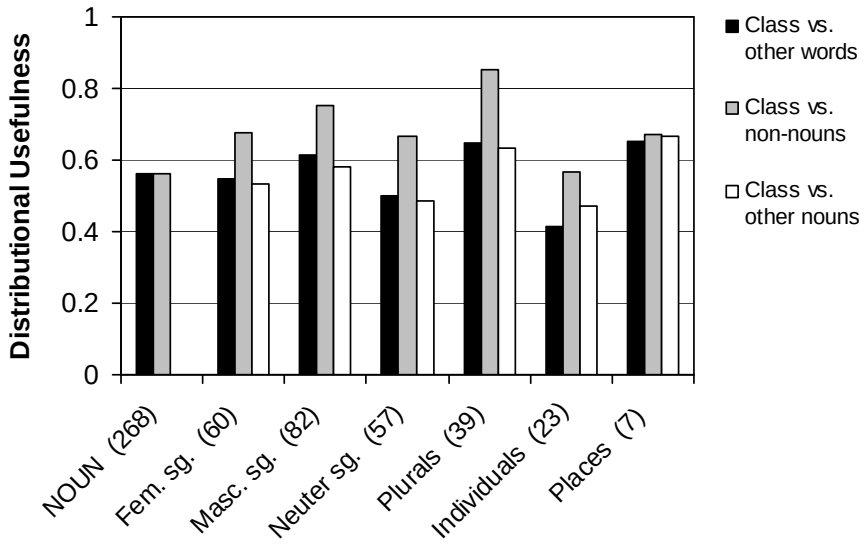

Figure 4-18: *Separation and performance of noun subclasses*

Three kinds of Distributional Usefulness scores are shown for the noun category and its basic subclasses. As before, these scores quantify how useful distributional information is to separate the given subclass from all target words outside that subclass (be they nouns or non-nouns; black), to separate it from all non-nouns (gray), or to separate it from all nouns outside the subclass (white). Note that for the full noun category, the third value is not defined, whereas the first two values are necessarily identical. Subclass sizes are given in parentheses.

The figure also shows that, with the exception of the tiny class of names for places, each individual noun subclass separates much better from non-nouns (separation values ranging from .57 to .85; gray bars) than it separates from all other nouns outside that subclass (values ranging from .47 to .63; white bars). Moreover, except for names for individuals, all noun subclasses individually separate clearly better from non-nouns (for this group, separation values range from .67 to .85; gray bars) than does the global noun category (.56).

At least for the four common noun subclasses, this pattern is very much like the one that was observed earlier for the individual finite verb subclasses (cf. Figure 4-17 on p. 162); there, it was only more pronounced and occurred at a higher range of Distributional Usefulness values. Therefore, a similar reasoning as for finite verbs now leads to the conclusion that the individual subclasses of common nouns occupy adjacent regions in the SCO vector space, such that all common nouns together occupy a much larger region than does each of its subclasses. And just as for finite verbs, these adjacent regions overlap with each other, as the confusability analyses in Table 4-10 below indicate. The particular degree of overlap between these regions varies but is generally higher (as the separation values are lower) than it is between the different finite verb subclasses (compare Table 4-10 with Table 4-8 on p. 163). The highest degree of mutual overlap is found between masculine singular and neuter singular nouns, and likewise between feminine singular nouns and plural nouns.[169]

Table 4-10: *Pairwise separation between noun subclasses*

Subclass Γ_1	**Subclass Γ_2**					
	Fem. sg.	Masc. sg.	Neuter sg.	Plurals	Individuals	Places
Feminine sg.	—	.63	.70	.48	.62	.66
Masculine sg.	.71	—	.54	.75	.73	.79
Neuter sg.	.71	.43	—	.72	.70	.77
Plurals	.59	.87	.87	—	.88	.94
Individuals	.59	.56	.67	.71	—	.75
Places	.69	.69	.70	.71	.71	—

Note. Table cells specify Distributional Usefulness of noun subclass Γ_1 when the target lexicon is restricted to members of Γ_1 and Γ_2, thus quantifying how useful distributional information is to distinguish Γ_1 from Γ_2. The four common noun subclasses are set off from the two proper name subclasses by horizontal and vertical space.

The two subclasses of proper names also form clusters that are adjacent to the other noun clusters and partly overlap with them — but in these two cases, this pattern does not follow as straightforward from the quantitative results. First of all, the mere fact that both subclasses have some degree of mutual confusability with two or three common

[169] It is remarkable to find both masculine singular and neuter singular nouns to separate so well from plural nouns (and vice versa), as 25.6% of all masculine singular nouns and 19.3% of all neuter singular nouns are homonymous with their corresponding plural form. For all other pairs of noun subclasses, the degree of lexical ambiguity between them is very low, in most cases even zero.

noun subclasses indicates that names for individuals and names for places indeed do overlap with the overall region occupied by common nouns — rather than surrounding this region, or being surrounded by it. The subclass of names for places additionally separates fairly well from non-nouns (.67; cf. Figure 4-18 above); and this ensures for such a small subclass that most of its members form some kind of core cluster.[170] This and the overlap with the common noun region implies that names for places are located somewhere in this same region, or attached to its edge.

Names for individuals may not separate quite as well from non-nouns (.57); but they still do so clearly better than they separate from nouns other than themselves (.47). This together with the relatively high degree of mutual confusability with some common noun subclasses entails that at least a substantial portion of names for individuals is found in, or attached to, the region of common nouns. Since most common nouns and even the overall noun category separate better from non-nouns than do names for individuals, this subclass cannot be very consistent in its distributional properties. And indeed, inspecting the L_1 distances between the 23 names for individuals revealed that, although most of them actually cluster fairly well, there are at least six clear outliers, with two of them being located extremely remote from the main cluster.[171]

Thus, in sum, all six noun subclasses essentially form adjacent clusters by themselves, but these clusters overlap. This pattern can also be observed, though in an overstated way, in the two-dimensional projection of the SCO vector space (Figure 4-19 below). Crucially, all nouns together inhabit a coherent region — even though it is a large one — and not two or more separate clusters as was the case for verbs. This fundamental topographic difference is what makes the SCO vector constellation of the

[170] This subclass has one single outlier, namely *ostsee* (English: *Baltic Sea*). It differs from the other names for places semantically in that most others are names of cities. And, crucially, it differs from them in terms of its usage preferences since *ostsee* is virtually always preceded by a feminine singular form of the definite article (at a rate of 29.4% by the nominate form *die*, and for 40.2% by the dative and genitive form *der*) or by *zur* which is a blend of the preposition *zu* (English: *to*) and *der* (22.5%). The other six names for places do not share these preferences but many feminine singular common nouns do (cf. Appendix F) which is why *ostsee* clusters with these.

[171] Two prominent outliers are actually the most frequent nouns in the corpus, *leo* and *wilhelmine*, which are the first names of the target child and his sister who was born during the three-year period of the recordings. It is not surprising that in a corpus of child-directed speech, names of children are used differently — particularly, in different constructions and lexical contexts — than other names for persons or toy animals. The most extreme outliers, however, are *leo+hartwig* (first and middle name of the target child) and *schatz*. This latter noun is the German word for *treasure* but was classified as a proper name because in the corpus, it is most frequently used to address the target child. Both *leo+hartwig* and *schatz* are used as alternative names of *leo* mainly on a specific range of pragmatic occasions (e.g., when he does not respond right away) which explains why their distributional properties deviate even more from those of other names for individuals than do those of the name *leo* itself.

noun category an Intermediate Scenario, and that of verbs a Hybrid Scenario (cf. the initial evaluation in terms of Global and Local Coherence, p. 158).

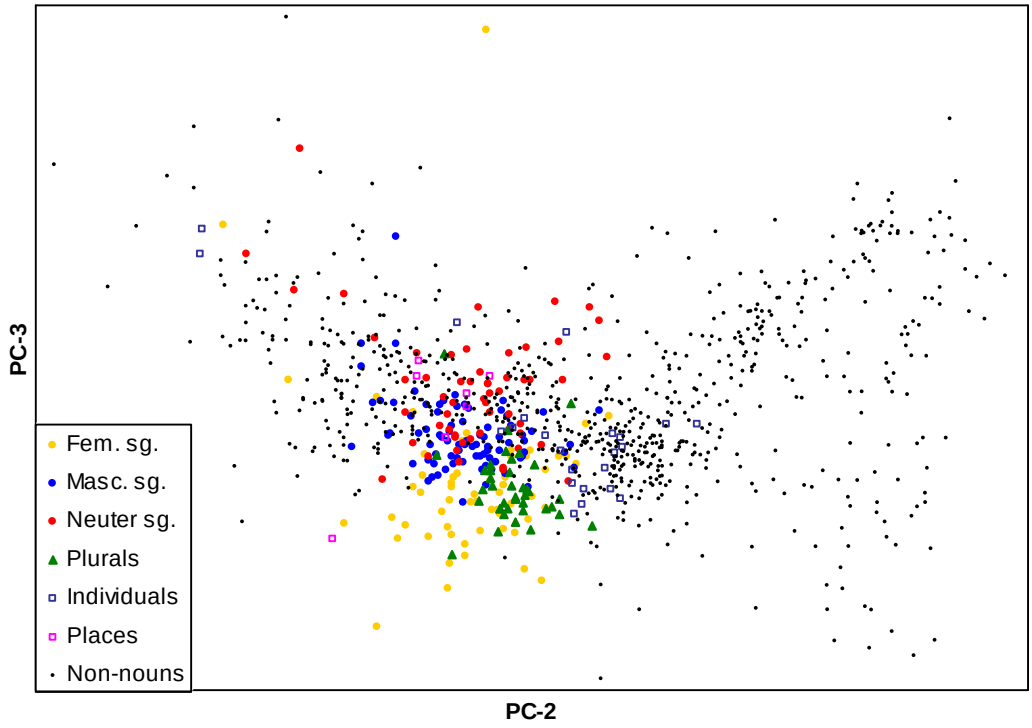

***Figure 4-19**: Noun subclasses in the SCO vector space (two-dimensional projection)*

Grammatical subclasses of the noun category, shown in the second and third principal component (PC-2 and PC-3) of the full SCO vector space. Vectors are marked for their subclass (with all singular common nouns being displayed as solid circles, and all proper names as empty squares).

But there are nevertheless qualitative similarities between the topographies of noun and verb category, as the global noun cluster combines features of the two clusters of finite and nonfinite verbs. The noun cluster is subdivided into adjacent but overlapping subclusters — just as was the case for the cluster of finite verb forms. But whereas finite verbs essentially have a large region all to themselves, the noun cluster overlaps also with some non-noun categories — as was also observed, though to a more pronounced degree, for the cluster of nonfinite verb forms.[172]

The kinds of words that the overall noun cluster overlaps with are mainly adjectives and adverbs (separation values .51 and .56, respectively; cf. Table 4-1 on

[172] Compare the separation values of noun subclasses from non-nouns (cf. Figure 4-18 on p. 172) with those of finite and nonfinite verb forms from non-verbs (cf. Figure 4-17 on p. 162).

p. 101) and also, coincidentally, these very same nonfinite verb forms (.55; not shown there). The distributional commonalities of nouns with these sets of words were already discussed (cf. p. 136 for adjectives and adverbs; p. 168 for nonfinite verbs).

The most important conclusion to make from these numbers is that the constellation of the noun category in the SCO vector space is much less complex than that of the verb category (cf. the discussion in 4.4.3). At the same time, however, the situation for finite verbs, when treated as an independent category of their own, is even better than that for the overall noun category (Distributional Usefulness for finite verbs is .69, for nouns .56).

In the remainder of this subsection, the distributional profiles of the individual noun subclasses are compared with each other, in order to account for the overall coherence of the noun region, and for the overlapping subclusters within this region. A specific aim is to consider the relevant factors that were found to be problematic for the coherence of the overall verb category (syntactic structure, usage preferences, grammatical agreement), and determine their influence on the noun category.

Table 4-11 below presents the distributional profiles of the six subclasses of the noun category, summarized by context word category. Their most salient distributional properties at the level of individual context words can be found in Appendix F. In interpreting the subclasses' profiles, I freely make use of both sources without always providing explicit references to either one of them.

The profiles are indeed remarkably consistent across subclasses, especially across the four subclasses of common nouns. I therefore discuss these common noun subclasses first. All four of them strongly prefer to occupy either of the two utterance-final positions (mostly irrespective of the kind of utterance termination), and even their precise preference values for these two serial positions are fairly consistent. At the cumulative level, all common noun subclasses show a very strong preference to be immediately preceded by a determiner. The comparatively lower cumulative preference value for neuter singular nouns solely results from the fact that the corresponding definite article (for nominative and accusative case), *das*, was classified as a pronoun in the benchmark. On average, the likelihood of neuter singular nouns to be immediately preceded by *das* is 22.3% (cf. Table 4-5 on p. 140).

Common nouns of all subclasses also show a preference to occur next to verbs, most frequently in the context positions [+1] and [−2]. The single most frequent verb form to occur in these context positions is *ist* (English: *is*) for the three singular common noun subclasses, and *sind* (English: *are*) for the subclass of plural noun forms.

Another preference mainly shared by the singular common noun subclasses consists of frequent co-occurrences with prepositions in [−2]. At the level of individual context words, all three classes prefer essentially the same basic set of prepositions (*auf*, *in*, *mit*; English: *on*, *in*, *with*, respectively), although the particular preference values vary.[173]

Table 4-11: *Distributional profiles of noun subclasses (cumulative summary)*

Left context				Noun subclass [a]	Right context			
[−2]		[−1]			[+1]		[+2]	
20.1	PREP	71.1	DET	Feminine sg. (60)	34.0	<Bnd>	34.6	<Bnd>
14.9	V	9.3	ADJ		24.3	V	10.9	V
12.6	DET	3.8	<Bnd>		9.5	ADV	10.7	PRON
16.3	V	63.6	DET	Masculine sg. (82)	36.7	<Bnd>	37.5	<Bnd>
16.2	PREP	11.3	PTCL		22.2	V	12.3	V
14.3	PTCL	6.1	ADJ		10.2	ADV	8.8	DET
16.6	V	41.0	DET	Neuter sg. (57)	36.4	<Bnd>	35.3	<Bnd>
15.0	PREP	28.9	PRON		22.2	V	11.8	V
13.5	PTCL	8.8	PTCL		10.3	ADV	10.0	PRON
16.3	V	55.5	DET	Plurals (39)	29.0	<Bnd>	35.1	<Bnd>
15.3	PTCL	7.8	<Bnd>		23.9	V	10.1	PRON
14.8	ADV	6.0	ADJ		12.9	ADV	9.8	V
19.4	<Bnd>	28.3	DET	Individuals (23)	30.9	<Bnd>	22.2	<Bnd>
15.7	V	15.5	<Bnd>		27.5	V	15.1	V
10.7	PRON	12.9	V		8.8	ADV	10.9	PRON
18.6	ADV	61.2	PREP	Places (7)	44.9	<Bnd>	39.0	<Bnd>
15.7	V	13.8	DET		27.9	V	10.5	PRON
15.2	PTCL	7.3	V		5.4	CONJ	10.4	V

Note. The distributional properties of each noun subclass (central column) are summarized by context word categories. The symbol <Bnd> represents all four utterance boundary markers. Only the three most likely context word categories are shown in each context position. The percentages next to them specify cumulative preference values which estimate the average conditional probability at which members of the noun subclass co-occur with any word of the context word category in the respective context position. Probabilities above 30% are shaded.

[a] Size of subclass is given in parentheses.

All these shared preferences reflect that common nouns have essentially the same syntactic privileges. They occur in NPs which have the same set of possible positions in an utterance, for instance as subject NPs right before the main verb, or as object NPs or in PPs at the utterance end and soon after the main verb.[174] And independent of the

[173] The lower co-occurrence preference of plural nouns with prepositions in context position [−2] mainly is a side-effect of a tendency of plural nouns to occur without a determiner (cf. p. 180). A minor additional reason may be that, in comparison to singular nouns, plural nouns are more frequently used as subjects and thus less frequently in object NPs and, particularly, in PPs.

[174] In subclauses, the main verb will occur after any PPs and object NPs, at the utterance end. Nonfinite verb forms can occur in such a position in many constructions, whenever the main verb is a modal verb (for infinitives) or an auxiliary (for past participles). These structural possibilities are also consistent with the salient distributional preferences of nouns.

position, these NPs have the same set of possible internal structures, the only exception being that plural nouns do not require a determiner (in the indefinite case). Thus, **syntactic structure** substantially contributes to a fairly homogeneous topography of the common noun class in the SCO vector space.

But even though all common nouns essentially share the same syntactic privileges, the spectrum of syntactic possibilities is vast — in theory, it is infinitely large. Even the internal structure of NPs can take any conceivable degree of complexity. And different possibilities in- and outside the NP result in systematically different local contexts around the noun. Averaging local contexts across all these possibilities would therefore likely yield very blurred distributional profiles.

However, the different structural possibilities are not equally likely to occur in the corpus; in fact, only very few of them have a nonmarginal probability of being used. And, crucially, most common nouns — irrespective of gender and number — are very similar with respect to which structural possibilities they prefer to occur in; that is, they have very similar **usage preferences** at the structural level. Most strikingly, they tend to be used in simple NPs of the form *DET N* or in simple PPs of the form *PREP DET N*, and these preferences hold even for plural nouns. This means that some of the most reliable distributional cues for common nouns occur close to the noun and generally in the same relative position to the noun, no matter where it occurs in an utterance (cf. Braine, 1987).

With respect to the position of the NP, there are at least three kinds of preferences at the construction level that, at the level of local contexts, conspire to a strong preference for common nouns to occur in the last two utterance-positions — a preference which was indeed observed in the profiles of all noun subclasses. The first reason is that common nouns tend to occur in PPs or object NPs rather than subject NPs.[175] And this is chiefly an epiphenomenon of speakers' tendency to use pronouns as the grammatical subject, or to omit the subject entirely.[176] When used within a PP or object NP, common nouns are very likely to occur in the last or last but one position of the utterance, across many different types of constructions. Second, even when they do appear in the subject slot, common nouns nevertheless tend to occur at the end of the utterance, for instance in many types of questions which constitute a large portion of

[175] This was verified by looking at random subsamples of the corpus.
[176] A bias towards pronominal subjects is characteristic of spoken language in general (cf. Tomasello, 2003b).

utterances in the corpus (35.3%; cf. 2.1.2).[177] Finally, many utterances are isolated phrases or larger sentential constituents; and when these are NPs or PPs, the noun is, once again, bound to occur utterance-finally.[178]

The tendency of common nouns to be used in PPs and object NPs and PPs also results in their preference to occur soon after the main verb (in questions and simple main clauses) or right in front of one or multiple verbs (in infinitival constructions and subordinate clauses). And this, too, was observed in the distributional profiles of the individual common noun subclasses.

Usage preferences also explain why all four common noun subclasses co-occur relatively often with the individual adverb *noch* (English: *still*) two words to their left. These arise predominantly from the highly preferred construction *noch DET N* where the determiner typically is an indefinite article (e.g., *"noch eine erbse"*, English: *"(still) a/another pea"*) or an indefinite numeral (by far most frequently the fixed combination *ein+paar* as in *"noch ein+paar nuesse"*; English: *"(still) some (more) nuts"*).

These examples illustrate how various salient distributional properties of common noun subclasses arise from profound usage preferences among the vast set of their structural possibilities, and how their fairly consistent preferences at this structural level directly result in very consistent preferences at the distributional level. Individual nouns may deviate from these usage preferences, due to statistical noise and semantic or pragmatic factors — e.g., animate nouns tend to be used in subject position more frequently than inanimate nouns —, but overall, usage preferences contribute to the homogeneity of the noun category, rather than supporting the formation of subclusters.

Thus, both syntactic structure and structural usage preferences do not only contribute towards a coherent noun cluster, but also to a very homogeneous structure inside this cluster. But such degree of homogeneity is not what we observed — for the noun cluster did show some internal substructure, and the main organizational factors among common nouns were found to be gender and number.

The primary linguistic explanation for this substructure is **grammatical agreement**. As was pointed out earlier, in German NPs, the head noun and any corresponding determiner have to agree in grammatical gender, number, and case. And despite a considerable degree of syncretism in the inflectional system of determiners, their agreement with nouns links each of the four common noun subclasses with a

[177] Such a large proportion of questions is very typical of CDS in general (e.g., Newport, Gleitman, & Gleitman, 1977; Cameron-Faulkner et al., 2003).
[178] Isolated phrases are very common among CDS utterances (e.g., Cameron-Faulkner et al., 2003).

specific set of determiner forms that are, at least in their composition, unique. Thus, although all four subclasses strongly prefer to occur in simple NPs and, hence, to be immediately preceded by *some* determiner, any two subclasses can be roughly distinguished by at least one specific determiner.

Of course, this is a textbook case of what was called *parallel cues* (cf. p. 156), and when the concept of parallel cues was first introduced, we already observed three individual determiners that set apart the three subclasses of singular common nouns (cf. pp. 139f). Comparing all four distributional profiles at the detailed level (Appendix F) yields a number of further individual determiners that distinguish some of these subclasses from each other. To name a few examples, masculine and neuter singular nouns share many determiners that they prefer to follow (e.g., *dem*, *ein*) which distinguish both from feminine singular nouns and from plural nouns; but masculine and neuter singular nouns differ in that only the former occur frequently after the definite article *der* (English: *the*$_{\text{masc.sg.nom./fem.sg.:gen.+dat./fem.+masc.+neut.:pl.gen.}}$) and after the indefinite article *einen* (English: *a*$_{\text{masc.sg.acc.}}$).[179] Plural nouns share a number of preferred determiners with feminine singular nouns (such as *die*, *deine*, *keine*) which set both apart from masculine and neuter singular nouns; but only feminine singular nouns frequently follow *eine* (English: *a*$_{\text{fem.sg.:nom.+acc.}}$; *one*$_{\text{fem.sg.:nom.+acc.}}$) or *'ne* (enclitic of *eine*), while only plural nouns like to follow *viele* (English: *many*$_{\text{fem.+masc.+neut.:nom.+acc.}}$) and the fixed combination *ein+paar* (English: *a few*).[180]

Agreement of determiners with nouns is the explanatory key to the internal structure of the cluster of common nouns. It is the main source of variation in the distributional profiles of the four different subclasses of common nouns, and thereby gives rise to the corresponding subclusters. At the same time, this implies that the overlap between two subclusters should be the higher, the more the corresponding subclasses converge in the sets of determiners they prefer to occur with. Therefore, the

[179] Note that *der* also is a plural determiner for genitive case, irrespective of gender. Therefore, neuter singular nouns that are homonymous with their plural form can take *der* as a valid determiner. However, since this only applies to 19.3% of all neuter singular nouns in the target lexicon, and because genitive case is rarely used in the corpus, co-occurrences with *der* are no salient preference of neuter singular nouns.

[180] One additional specialty of plural nouns is that they often do not take a determiner at all in the indefinite case — just as in English. While this by itself may not constitute a cue to plural nouns (at least as long as the determiner category is not yet acquired), it has the consequence that in simple PPs, the preposition can occur in context position [−1] relative to the noun, and therefore occurs less frequently in context position [−2] where it most frequently occurs for the singular nouns. Another consequence is that, due to cases where the noun is the head of a subject NP, plural nouns are more likely to occur utterance-initially than are singular nouns. All these differences can be verified both at the cumulative level (Table 4-11) and in terms of individual context words (Appendix F).

large proportion of determiners that both masculine and neuter singular nouns frequently co-occur with, explain the lower mutual separation values between these two subclasses (cf. Table 4-10 on p. 173). A comparatively low mutual separation was only observed between plural nouns and feminine singular nouns (again cf. Table 4-10), and this is in line with the large portion of preferred determiners that these two subclasses have in common.

Indeed, when the sets of determiners that different grammatical subclasses of common nouns co-occur with are effectively identical, the distinction between these subclasses is missed entirely by the model. This is the case for the gender subclasses among plural noun forms. As was remarked earlier, these classes are essentially not differentiated in the SCO vector space, unlike their singular counterparts. Although agreement for gender in principle also applies to determiners and plural nouns, all plural determiner forms are homonymous across the three genders. Consequently, they constitute strong positive cues to the overall class of plural common nouns; but they do not further discriminate the gender subclasses among these plural nouns.

A second type of grammatical distinctions to which the model is essentially insensitive, although they play a role in agreement, concerns grammatical case. However, this time it is the nouns themselves that are to be blamed, for the determiners actually are marked for case, though not always unambiguously (cf. earlier remarks on syncretism). But case is not systematically marked on common nouns; and in fact, setting aside the genitive case which is hardly ever used in the corpus, the vast majority of nouns do not inflect for case at all. Therefore, most common noun forms are used in nominative, dative, and accusative case alike; and insofar as different nouns have similar usage preferences to occur in each of these grammatical cases, they must have roughly the same probabilities to co-occur with the nominative, dative, or accusative determiners appropriate for their gender and number. Like this, distributional distinctions of grammatical case that are potentially provided by agreement between determiners and nouns are greatly blurred and do not result in the systematic formation of noun clusters defined by case.

Yet, these distinctions are not lost entirely. Traces of them show up whenever some set of common nouns prefers a particular case more than do other nouns. But the consequences are rather weak and only occur inside individual clusters of particular common noun subclasses. The reason is that these nouns still co-occur with the same sets of determiners as do other nouns within the same subclass — what differs are only the specific preference values for those determiners of the relevant grammatical case.

One intuitively plausible example is found in animate nouns. They are generally more natural agents than inanimate nouns; and inasmuch as speakers prefer to express agents as grammatical subjects, animate nouns are used in subject NPs more frequently than are inanimate nouns. Relative to most inanimate nouns, animate nouns therefore tend to be more frequently preceded by nominative determiner forms of the given noun subclass.[181] And indeed, at least within each of the three subclasses of singular common nouns, animate nouns tend to be located towards the same edge of the corresponding cluster in the SCO vector space. Among feminine singular nouns and also masculine singular nouns, the animate nouns even have so consistent properties that they can be fully distinguished from non-nouns based on distributional information.[182]

This concludes the analysis of the large cluster of common nouns which represent the lion's share (88.8%) of all nouns in the target lexicon. In the remainder of this subsection, I discuss the distributional profiles of the two small subclasses of proper names in turn, by summarizing their commonalities with the profiles of common nouns, and by highlighting the ways in which they are special.

Proper names for individuals share many preferences with the four common noun subclasses (cf. Table 4-11 on p. 177). Their most salient properties are likewise to occupy the two utterance-final positions and to occur right after determiners, though these preferences are less pronounced than for common nouns. The co-occurrences with determiners mainly pertain to the feminine and masculine base form of the definite article, *die* and *der* (cf. Appendix F). In German CDS — as well as in various regional dialects of adult German — speakers tend to use first names with a definite article, the grammatical gender of which matches the sex of the individual referred to. And for the particular case of the *Leo* corpus, the child's immediate family members do this routinely even when referring to themselves (e.g., *"kriegt die mama auch 'n kuss ?"*; English: *"Does (the) Mommy also get a kiss?"*) which is not uncommon for CDS in general either.

However, this tendency of speakers to use first names with a determiner is clearly reduced when the name occurs in a PP. As a consequence, names for individuals are

[181] A second distributional consequence is an increased preference of animate nouns to occur close to the utterance beginning, due to the bias of subject NPs to occur utterance-initially while accusative and dative NPs occur mostly at the utterance end. This second consequence thus further contributes to a more consistent profile of animate nouns within a grammatical subclass of common nouns.

[182] Cues from grammatical case only single out animate nouns, and only to some extent — they do not actually predict the animate–inanimate distinction. There are many inanimate nouns which also make good agents for some verbs, while many other inanimate nouns are indeed unlikely to be used as an agent. For this reason, inanimate nouns are rather heterogeneous with respect to the relevant kinds of distributional preferences.

about equally likely to occur with a preposition one or two words to their left — a property which they share with plural nouns, but not with singular nouns.

Like all common noun subclasses, names for individuals are likely to co-occur with verbs and do so most frequently in context position [+1]. The most preferred individual verb form is once again *ist* (English: *is*), just as for the three subclasses of singular common nouns. However, this preference is more pronounced for the names than for the singular common nouns, and this is no accident. In fact, one sizable difference between these two groups that cannot be directly inferred from Table 4-11 and Appendix F, is that names for individuals are by far more likely to be immediately followed by a finite verb form (on average, 21.3%) than by a nonfinite verb form (6.2%), whereas singular common nouns have no clear bias towards either class of verbs (12.1% vs. 10.7%, respectively).[183] Presumably, when a noun is succeeded by a finite verb, it is most likely the head of the subject NP in a main clause declarative, whereas nonfinite verb forms follow a noun most typically when the noun occurs within a PP or object NP.[184] If this speculative statement is correct, then the divergent preferences with respect to finite verbs reflect that names for individuals are, on average, more likely to be the grammatical subject of a predicative utterance than are common nouns. This conclusion is further corroborated by the higher preference of names to occur in the first two utterance positions (compare these preferences with those of common noun subclasses, cf. Appendix F).

An increased preference of names for individuals to be used as the grammatical subject is highly plausible, because individuals are prime candidates for being the agent of an action talked about in CDS (cf. the earlier considerations for animate nouns, p. 182). And as was just argued, this increased preference at the syntactic level has consequences at the level of local distributional properties, which therefore differentiate the bulk of names for individuals from common nouns.[185]

Finally, proper names for individuals occur quite often right next to the single conjunction *und* (English: *and*), on either side. These preferences are not shared to the

[183] Plural common nouns do not have such a bias either; but even if they had the same bias towards being followed by finite verb forms, these would often be plural forms — due to agreement between a subject noun and the corresponding main verb — whereas names for individuals prefer to be followed by singular verb forms. This is therefore a case where agreement between a noun and a verb contributes to the internal substructure of the noun cluster.

[184] Of course, there are also other types of constructions in which nouns are used within a PP or object NP but are nonetheless succeeded by a finite verb form; e.g., in subclauses where the main verb occupies the final position. However, such constructions are rare relative to main clause declaratives.

[185] Recall that the low separation of names for individuals from some of the common noun subclasses arises from a few pronounced outliers whereas the majority of names cluster rather well (cf. p. 174).

same extent by common nouns such that they further carve out the main subcluster of names. But despite these differences between the distributional profiles of common nouns and names for individuals, their commonalities still predominate such that the subcluster of names is connected to the core cluster of common nouns.

To turn to the small subclass of proper names for places, their distributional commonalities with common nouns and names for individuals are not as abundant. Like these other subclasses, names for places occur very often in the utterance-final positions. But here, the differences already begin, since names for places on average occupy this last utterance position even much more frequently than do the members of any other noun subclass.

This high preference of names for places to occur utterance-finally coincides with an exceedingly high preference to immediately follow a preposition. These two observations are directly related as they both arise from the strong preference at the structural level: Names for places most typically occur in PPs — rather than in subject NPs — and therefore in the majority of cases close to the utterance end. The co-occurrences with prepositions (61.2%) are almost exclusively accounted for by the individual prepositions *nach* (36.1%; English: *to*), *in* (19.3%; English: *in*), and *von* (4.2%; English: *from, of*)

When names for places do not occur in the last utterance position, they are most frequently succeeded by a verb form; and in this preference, they do resemble singular common nouns to some extent because they, too, have no clear bias towards nonfinite or finite verb forms (12.6% and 15.3%, respectively).

Names for places are generally not preceded by a determiner, a fact which further distinguishes these names from common noun subclasses.[186] Interestingly, however, this general absence of determiners in context position [−1] and the frequent co-occurrences with prepositions in the same context position — both of which individually constitute a distributional difference between names for places and common nouns — together indirectly support some distributional commonalities between these classes in context position [−2]: Inasmuch as common nouns occur in object NPs more frequently than in PPs, and to the extent that PPs and object NPs tend to occur after the same kinds of words (e.g., adverbs and finite forms of verbs that can be used transitively and intransitively), common nouns and names for places are likely to occur with these kinds

[186] The fact that two determiner forms do show up among the salient preferences of names for places, arises solely from the outlier *ostsee* (English: *Baltic Sea*) which virtually always is preceded by a definite article (cf. footnote 170, p. 173).

of words in context position [−2] (compare *"der faehrt den bus ."* with *"der faehrt nach stuttgart ."*; English: *"He is driving the bus."* and *"He is driving to Stuttgart."*, respectively).

In sum, despite several unique co-occurrence preferences, proper names for places share some of their salient distributional properties with common nouns. But the fact that, in the SCO vector space, names for places are located in the common noun region, is not only a result of these commonalities alone but also of the fact that this region is fairly large, and that even within each particular common noun subclass, members are somewhat dispersed. This is because the probability of individual members to enter a particular co-occurrence relation can in some cases vary considerably around the subclass-wide average (i.e., its preference value). Seen in this light, even the more profound distributional differences between the common noun subclasses and proper names for places only entail that proper names are distant from the center of the common noun cluster, but not that the names are disconnected from this cluster altogether.

4.4.3 Implications: Verbs vs. nouns

The starting point of this section about nouns and verbs was a fundamental topographic difference between the SCO vector constellations of verbs and nouns. The subsequent investigations revealed that verbs essentially partition into two isolated clusters corresponding to nonfinite and finite verbs forms, whereas nouns essentially inhabit one coherent, though large, region in the SCO vector space. It was further found that the cluster of nonfinite verbs overlaps with some clusters of non-verbs. Although the cluster of finite verbs does not show such overlap, it is nonetheless too close to many non-verbs, given the size of the region that it occupies. Each of these two problems applies to some extent also to the noun cluster: It partly overlaps with some clusters of non-nouns, and it is fairly large, though not quite as large as the finite verb cluster.

The detailed analyses of distributional profiles uncovered three main factors — namely, syntactic structure, usage preferences, and grammatical agreement — that together cause the topography of the verb category. First, syntactic structure constrains the possible local contexts in which verbs can occur, and these constraints were found to be very inconsistent between finite and nonfinite verb forms. Second, within these constraints, the individual verb subclasses have different usage preferences for their possible contexts. And third, agreement between subject pronouns and finite verbs

creates further distributional differences among verb subclasses. In sum, all three factors support the formation of more or less distinct verb subclasses which make it difficult, if not impossible, to discover the entire verb category from distributional information alone.

The influence of these factors is somewhat different for the noun category. In fact, for the lion's share of nouns (viz., common nouns), the local constraints imposed by syntactic structure are essentially identical. Their usage preferences within these constraints may differ in some cases, but not to the same extent as they do for verbs. Both these factors thus create many distributional commonalities among common nouns which supports the coherence of the overall noun category. Only the third factor, agreement (here with determiners), actually results in systematic distributional contrasts between different subclasses of common nouns.[187]

In fact, the effects of agreement are even more problematic for the noun category than they are for the overall verb category. This follows from two distributional experiments which are summarized in the following. In the first experiment, all pronouns among the context words were collapsed to one single context item. More precisely, for each target word, all its co-occurrences with any pronoun form in a particular context position were summed up to one single co-occurrence count. Like this, the model had information as to how frequently the target word co-occurs with pronouns in the given context position, without further differentiating which particular pronoun it is. This manipulation removed all consequences that agreement with pronouns has on the distributional properties of verbs. The result was that, without this agreement information, Distributional Usefulness for the overall verb category rose from .38 to .47.

Of course, agreement with pronouns only concerns finite verbs; and it was argued in 4.4.1 that agreement is problematic mainly in that it causes the cluster of finite verbs to be overly spread out in the SCO vector space. This topographic effect of agreement is also confirmed by the experiment, as the separation of finite verbs from non-verbs improves from .72 to .84, whereas the separation of nonfinite verb forms from non-verbs changes only slightly from .48 to .53. At the same time, the mutual separation values between finites and nonfinites remain roughly constant (.78 and .86 for the default analysis, .82 and .82 on the experimental manipulation). Collapsing pronouns

[187] The interaction of the three factors is clearly more complex for proper names; but they nevertheless do create a sufficient amount of local distributional commonalities between proper names and common nouns.

thus renders the finite verb cluster more compact; but it does not change the fact that verbs partition into two highly isolated clusters.

However, collapsing all pronouns actually discards not only information from grammatical agreement but also information about the grammatical case of the pronoun, along with other grammatical contrasts, such as that between personal pronouns (e.g., *er*; English: *he*), reflexive pronouns (e.g., *sich*; English: *herself, himself, itself*), or indefinite pronouns (e.g., *jemand*; English: *someone*). Because only nominative (i.e., subject) pronoun forms agree with verbs, and because any distributional cues arising from the other grammatical contrasts would be independent of agreement as well, two additional variants of the first experiment were conducted. In the first variant, only nominative pronoun forms were collapsed, while in the second, this set was further restricted to personal pronouns. However, these variants only reduced the size of the experimental effects without changing their overall pattern. All three versions of this experiment therefore lead to essentially the same conclusion: Agreement is simply not the cause for the pronounced separation between finite and nonfinite verb forms.

The second experiment was targeted at agreement between determiners and nouns. In analogy to the previous experiment, any distributional consequences of this agreement relation were discarded by collapsing all determiners in the context lexicon to one single context item. The result of this manipulation is a profound improvement of Distributional Usefulness for the overall noun category (from .56 to .77). Since common nouns co-occur by far more systematically with determiners than do proper names, it is mainly the common noun cluster that becomes more compact (their separation from non-nouns increases from .59 to .83). By comparison, the changes for the two small clusters of proper names for individuals and proper names for places are marginal.

In sum, agreement with pronouns or determiners clearly affects the usefulness of distributional information for acquiring the overall verb or noun category, respectively. But the consequences are even more severe for the noun category such that agreement effects essentially do not contribute to the distributional advantage of the noun category over the verb category. The crucial factors explaining this advantage are thus syntactic structure and, to a smaller degree, usage preferences. Both factors conspire to make the local environment of common nouns very predictable: Particular kinds of words tend to occur in the same context position relative to the noun, and the noun tends to occupy either of the two final utterance positions. For verbs, by contrast, both factors generate a high degree of variation in the verb environment: In particular, neither the serial position of a verb form (utterance beginning vs. utterance end), nor the relative position

of a co-occurring subject pronoun (to the left vs. to the right of a finite verb form) is by itself very predictable.

In this sense, the verb category is distributionally more complex than the noun category, and it would therefore pose a much greater challenge to any appropriate category learning mechanism exploiting distributional regularities in the *Leo* corpus. Such a purely distributional learner is likely to discover the noun category, or at least an approximate notion of it; but it would almost necessarily fail to develop a global verb category from distributional information alone. Especially the two distant verb clusters are problematic, and if the distributional learner would indeed group finite verbs in the same category with nonfinite verbs, this category would probably contain almost all non-verbs as well und therefore not be of any use to the learner (cf. p. 160). Therefore, in a more realistic classification, the distributional learner might rather discover a category of finite verb forms and another category of nonfinite verb forms — in fact, an independent category of finite verb forms is discovered from distributional information even more easily than the overall noun category (cf. p. 176). But at least early on, the learner is yet more likely to discover distinct categories of second person singular forms, imperative singular forms, and so forth.

As was stated in 1.2.2, children most certainly do not solely rely on distributional information for developing their lexical categories. But if they use this information to any significant extent, especially at the onset of category formation, the findings in the current section would predict that (i) a reliable notion of noun emerges before the verb category, and (ii) the earliest abstractions across different verb forms are made within the same grammatical subclass rather than for different inflected forms of the same verb lexeme.

To my knowledge, this second prediction has not yet been tested in developmental studies. But there is some promising evidence which is at least consistent with the prediction. In detailed cross-linguistic studies about the developmental course of the verb *go* (or its respective counterparts in Dutch and German), Theakston, Lieven, Pine, and Rowland (2002; for English) together with Behrens (2003; for German and Dutch) found strong evidence indicating that children initially acquire semantic and structural knowledge separately for each of the inflected forms of a lexeme before they begin to link these forms to a unified concept for which they can then acquire more abstract knowledge. If prediction (ii) above is correct, then there would be an intermediate stage during this developmental course of a verb lexeme in which children discover abstractions across forms of the same inflectional subclass before building unified

concepts for the underlying verb lexemes. For the case of the particular verb lexeme *go*, this would mean that children start out to discover usage commonalities between verb forms like *going* and *trying* before they link *going* with *go* and *went*.

To decide about the first prediction — viz., that in German, the noun category is acquired prior to the verb category — the critical evidence from developmental studies is also still missing. At least for English, it was shown in controlled experiments that children develop a productive noun category well before a productive verb category (Tomasello & Olguin, 1993; Olguin & Tomasello, 1993). But for lack of corresponding experiments in German, it remains unclear whether the predicted order of acquisition also holds for children acquiring German as their first language. Recordings of spontaneous speech do not show more than a trend that German children begin to vary syntactic and morphological properties of nouns earlier than those of verbs (e.g., Clahsen, 1982; Mills, 1985; Behrens, 1993; Eisenbeiß, 2002). However, this trend is not found in all such corpora. For instance, Stern and Stern (1928/1965) interpret the data from their diary study to imply that all major categories develop simultaneously. Furthermore, even if all available corpora would consistently show that syntactic and morphological variation appears on nouns earlier than on verbs, this could in theory still be accounted for by rote-learned forms rather than systematic categorial knowledge.

More promising evidence is provided by a recent experimental study that Höhle, Weissenborn, Kiefer, Schulz, and Schmitz (2004) conducted. Using the head-turn preference paradigm, the authors found that German children as young as 14-16 months are more surprised to encounter novel words in verb contexts than in noun contexts, after having been familiarized with these words as occurring immediately after an indefinite article. This indicates that children at this age have developed at least a rudimentary notion of a noun category, defined in terms of some of its distributional properties, and that they can use this distributional knowledge for ad hoc category assignments. No corresponding categorization effect was found when children were familiarized with novel words occurring after a subject pronoun. This may seem to indicate that the children have not yet discovered a verb category, or at least not its distributional characterization. However, as the authors point out, an alternative explanation is that subject pronouns might simply not be a very reliable cue to the lexical category of the immediately succeeding word. The findings in the dissertation at hand support this second explanation because, first, infinitives do not tend to follow subject pronouns at all; and second, even for most finite verb forms, when they do take

a pronoun as their subject, it occurs more frequently *after* than before the verb (due to inversion, cf. p. 167).

Both predictions about the developmental course of verb and noun category therefore remain open research questions. For the prediction concerning the order of acquisition, it would seem that the only viable way to obtain the critical evidence is to conduct experiments along the lines of Tomasello and Olguin (1993; Olguin & Tomasello, 1993). The prediction that the earliest abstractions for verbs occur within grammatical subclasses is tricky to test, for it might be hard to devise an experimental design that is in principle capable of actually falsifying this prediction. In order to prove it right, however, one would have to demonstrate that some abstract (semantic or combinatorial) property that holds for a variety of verbs irrespective of inflection, is first discovered only for verb forms of one particular grammatical subclass but not yet for other inflected forms of the very same verb lexemes.

4.5 Links to development

The distributional explorations in the preceding section yielded two predictions about the course of acquisition for two major categories; and these predictions would have to be tested in empirical studies of language development. Conversely, one may also use the evidence from existing developmental studies to revise the current model and analyze distributional information in a more realistic way. Two such developmental perspectives were considered and tested for their distributional consequences. The first one concerns evidence that children's initial sensitivity to function words differs from their sensitivity to content words (subsection 4.5.1), whereas the second issue takes into account that categories do not appear to emerge simultaneously, and that the first categories might facilitate or hinder the later acquisition of other categories (subsection 4.5.2).

4.5.1 The early role of function words

It has long been noted for many languages, that in their first multi-word productions children tend to omit function morphemes, i.e., function words and bound inflectional morphemes (e.g., Brown & Fraser, 1963; Miller & Ervin, 1964; for German: Stern &

Stern, 1928/1965; Miller, 1976).[188] But despite the *telegraphic style* of their own productions, children are sensitive to function morphemes in their linguistic input, as indicated by a rich body of evidence. For instance, as was already mentioned earlier (p. 21), function morphemes can roughly be distinguished from content words by their phonological and acoustic properties in a typological variety of languages (Shi, Morgan, & Allopenna, 1998; the papers in Morgan & Demuth, 1996, Part III). Shi, Werker, and Morgan (1999) found that already newborns are sensitive to these kinds of perceptual cues, which is clearly long before they can even make use of these cues because infants typically do not begin to learn segmenting speech into words before they are six months old (for review see Jusczyk, 1999). Shafer, Shucard, Shucard, and Gerken (1998) demonstrated in an ERP study that 11-month old American infants can actually use such perceptual cues to distinguish English function words from nonsense words occurring in function word contexts in fluent speech. This is confirmed by Shady (1997) who found in a number of experiments using the head-turn preference paradigm that by 10.5 months, infants recognize the phonological properties of function words. Shady's work further suggests that by 16 months, children have learned where in an utterance the various function words typically occur.

This rapid course of acquisition raises the question about the mental representations that children entertain for function words — or, more generally, function morphemes — at different developmental points. By age two, children appear to have acquired rich representations for at least the most frequent function words (Gerken, Landau, & Remez, 1990; Gerken & McIntosh, 1993). But during an intermediate stage roughly between 11 and 16 months, they might treat all function words as one and the same item before differentiating this single concept into separate lexical representations for the individual function words.

There are two ways in which such a developmental stage may help the distributional learner. First, to the extent that perceptual cues already lead the learner to a sufficiently reliable distinction between content and function words, the learning task for the distributional learner is considerably simplified: This distinction in effect

[188] Later research suggests that this early omission of function morphemes may not be true of all children. Some children were found to produce *schwa* sounds and other phonological filler items between the content words. And these *filler syllables* have been interpreted as the children's attempt to imitate the prosodic structure of their input language such that the fillers serve as *prosodic placeholders* for function morphemes (e.g., Bloom, 1970; Peters, 1997, 2001; Lieven, 1997; for German: Lleó, 2001; Peters, 1997, interpreting data from Stern & Stern, 1928/1965). This certainly applies to the target child of the current study who frequently produced schwa sounds, especially in unmistakable determiner positions, until he was roughly 30 months old.

partitions the target lexicon into two separate lexica (one for content words, and one for function words) such that distributional information would only be needed to discover the fewer categorial distinctions among each of these lexica.

Of course, as long as individual function words in the input are not discriminated as separate lexical items by the learner, finding categorial distinctions among these function words is impossible. However, once children start to pay attention to where function words appear in an input utterance, regardless of which particular words these are, they might already begin to exploit co-occurrence relations that other words — viz., content words that are already differentiated — have with these function words. Of course, this co-occurrence information is not as rich as it would be with differentiated function words. But as we have seen repeatedly in the two previous sections for the case of parallel cues, differentiated co-occurrence relations may sometimes create distributional inconsistencies within a category that actually render this category more difficult to be acquired from distributional information. Therefore, children's early failure to differentiate individual function words might even serve to augment the distributional evidence for categorial distinctions among content words. And this possibility is the second way in which a distributional learner might benefit from the developmental stage postulated above.

Both these influences were assessed in two experiments. Starting from the default analysis, the first experiment simply removed all function words (i.e., all interrogative words, pronouns, determiners, prepositions, conjunctions, and particles) from the target lexicon.[189] This amounts to asking how useful distributional information is for discovering the distinctions only between the five categories of interjections and content words. This reduced target lexicon was retained in the second experiment, but additionally, all function words in the *context lexicon* were collapsed to one single context item such that all co-occurrences with any of these function words (as context words) were treated as co-occurrences with the new context item. This step models the distributional consequences of not discriminating individual function words.

Formally, the same type of experiment was already encountered in 4.4.3, where determiners and pronouns in the context lexicon were collapsed to one context item each. However, the underlying research questions were very different from the current

[189] Note that, for convenience, I here equate function words with closed-class items, and likewise, content words with open-class items. Although these two distinctions differ theoretically and do not yield the exact same partitioning of the lexicon, they are in practice sufficiently good approximations of each other to justify being equated for the purpose of the current experiments.

one. There, the experiments were carried out only to assess the distributional consequences of grammatical agreement; the collapsing was therefore simply pursued as a formal research strategy. In the current experiment, by contrast, the collapsing actually represents realistic modeling assumptions about the learning situation of the child.

The Distributional Usefulness scores resulting from both experiments are shown in Figure 4-20, together with the corresponding scores for the default analysis in which both target lexicon and context lexicon consisted of the 1,017 most frequent words in the corpus. It turns out that removing function words from the target lexicon (first experiment vs. default analysis) improves Distributional Usefulness for all five categories that remain after the removal. However, this increase is rather small for all categories, except for **adverbs** which benefit considerably (Distributional Usefulness rises from .22 for the default analysis to .34 for the first experiment). The reason, of course, is that adverbs are distributionally very confusable with particles which are now removed from the target lexicon (cf. Table 4-1 on p. 101). The other four categories are not remotely as confusable with any category of function words.

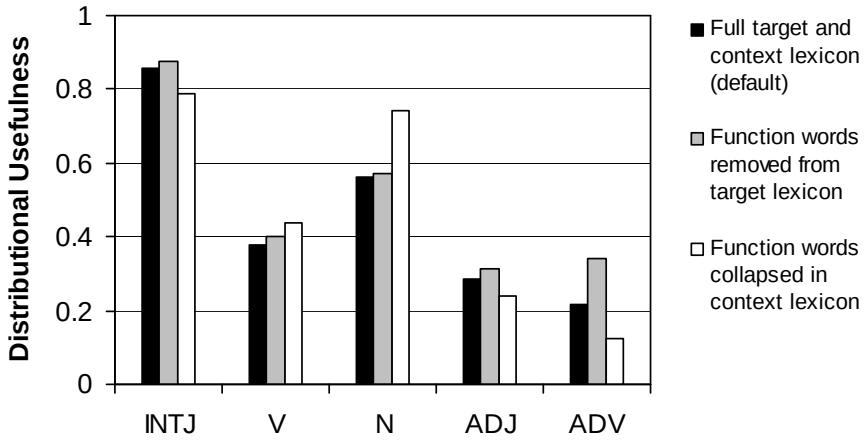

Figure 4-20: *Effects of the acquisition pattern of function words on other categories*

Distributional Usefulness for the default analysis (i.e., for the full target and context lexicon); when function words are removed from the target lexicon; and when, additionally, function words in the context lexicon are collapsed to one single context item.

When function words in the context lexicon are collapsed to one context item (second vs. first experiment), Distributional Usefulness clearly drops for **interjections**, **adjectives**, and **adverbs**, most substantially for the latter category. In fact, in all three cases, the drop clearly overcompensates for the earlier gain when function words were

removed from the target lexicon. This means that the information from co-occurrences with function words becomes less useful for discovering these categories when it is not clear which particular function words they co-occur with.

For **verbs**, Distributional Usefulness improves slightly (from .40 for the first experiment to .44 for this second experiment). A supplementary analysis in terms of Global Coherence (a minor increase from .27 to .30) and Local Coherence (a sizable leap from 78.6 to 104.5) indicates that the Hybrid Scenario characterizing the verb category becomes even more pronounced when individual function words are not differentiated. In other words, the two clusters of finite and nonfinite verb forms become even more compact in themselves, but they do not move closer together to fill the gap between them.

Probably the most remarkable finding in this second experiment is that the **noun** category benefits very much from the collapsed co-occurrence information (Distributional Usefulness rises from .57 in the first experiment to .74 in the second experiment). Moreover, both Global Coherence (from .64 to .78) and Local Coherence (from 78.2 to 122.8) increase in unison, which implies that the overall noun category becomes more compact and looks even more like a Clump Scenario.

Note that collapsing all function words in particular also removes the unfavorable distributional effects of grammatical agreement between determiners and nouns on the one hand, and pronouns and verbs on the other hand, that were explored earlier (cf. 4.4.3). Therefore it is not entirely surprising to find that both noun and verb category benefit in the current experiment. At the same time, collapsing words simultaneously from all function word categories might discard not only the agreement effects, but also some distributional cues that are critical for discriminating nouns or verbs from other categories. And for verbs, this indeed appears to be the case because the rise in Distributional Usefulness is smaller in the current experiment (from .40 to .44, see above) than it is when only pronouns are collapsed (from .38 to .47 when based on the full target lexicon; and from .40 to .50 when function words are removed from the target lexicon, to have comparable experimental conditions). The noun category, by contrast, benefits nearly as much from collapsing all function words (from .57 to .74, see above) as it does when only determiners are collapsed (from .56 to .77 on the full target lexicon; and from .57 to .78 on the target lexicon without function words).

This substantial improvement for the noun category suggests that if there is indeed a developmental stage in which children can already distinguish content from function words — even if only roughly — but not yet differentiate the individual function words

as independent lexical items, the discovery of the noun category would be facilitated considerably during this stage. In other words, a specific developmental phenomenon that might first have seemed to cause a loss of important information, turned out to potentially boost the acquisition of at least one category.[190]

At a general level, this result is in line with independent findings which suggest that early cognitive limitations might actually constitute an advantage for the learning infant because they initially reduce the learning space to a relatively small domain which draws the attention to a few relevant aspects to be acquired (e.g., Newport, 1990; Elman, 1993). Once these aspects have been mastered, they might in turn help the child to bootstrap into other aspects of his first language. In the current context, this suggests that the discovery of the noun category might facilitate the acquisition of some other categories; and this possibility is investigated in the next subsection.

4.5.2 The role of the noun category for other categories

Suppose that some particular category is mastered considerably before other categories; then children will acquire these other categories in the context of knowing about the first one. An obvious question is therefore to which extent this knowledge might facilitate the later discovery of the other categories. A distributional experiment was conducted to test this question for the particular case of the noun category. This is a reasonable candidate because, as was stated above, there is reason to assume that the noun category is generally acquired earlier than other lexical categories.

[190] Mintz et al. (2002) conducted a number of corresponding experiments and found that, overall, verbs either benefited from the collapsed co-occurrence information or remained unaffected, whereas nouns tended to cluster slightly worse. Because the *Purity* score used in their study most closely relates to Local Coherence, their findings for the verb category are consistent with those presented here. The divergent results for the noun category may stem from the slightly different analytical paradigms and the highly different sizes of data sets. But it is quite likely that they mostly reflect grammatical differences between German and English. For instance, collapsing all function words ameliorates the distributional basis for the noun category in German mainly by removing the unfavorable effects of grammatical agreement between nouns and determiners (cf. p. 194). But while this agreement involves number, case, and, crucially, gender, English only requires agreement for number, and some determiners do not even inflect for number. In some sense, relative to German, the English determiner system is in effect already collapsed, such that collapsing all determiners and function words provides little extra benefit. Redington et al. (1998) also experimented with collapsing function words to one symbol. They did not report category-specific figures, but overall, their measures of goodness decreased when co-occurrences with function words were collapsed. This is not immediately consistent with the current findings since the substantial improvement for the large category of nouns overcompensates for the decrease of Distributional Usefulness for some other categories. The different outcomes might therefore again be accounted for mainly by the grammatical agreement effects in German.

To provide the co-occurrence model with knowledge about the noun category, an experiment was conducted that is formally very similar to the one in the preceding subsection, yet, with one crucial difference. In that former experiment, all function words were collapsed to one new context item such that it replaced all individual function words in the context lexicon. This was meant to model the assumption that during some early stage, children may note when and where function words occur in their input, but do not differentiate individual function words. However, the assumption that is made for nouns in the current experiment is fundamentally different, for children do not cease to differentiate individual nouns after having acquired a noun category. They simply have access to the category information *in addition* to the individual noun.

This different situation was modeled as follows. Like before, a new context item was introduced, this time representing the noun category. Collapsed co-occurrences with nouns (as context words) were computed and treated as co-occurrences with the new context item. What is different from the previous experiment is that the new context item does not replace the individual nouns in the context lexicon; rather it is included in addition to all existing context words. Like this, the model extracts the same co-occurrence information as in the default analysis, only extended by information about the target words' preference to co-occur with the noun category.[191] The distributional consequences of this additional noun cue on the other 10 benchmark categories can be viewed in Figure 4-21 below.

Relative to the default analysis, Distributional Usefulness does not change very much for most of these categories which implies that these categories do not benefit from the noun cue. The scores for **interrogative words** and **interjections** even drop slightly, indicating that the additional information is not only of little use for these categories but that it even obstructs the more informative cues from other kinds of co-occurrences. However, Distributional Usefulness increases substantially for the categories of **determiners** (from .38 to .52) and **prepositions** (from .48 to .56), and to rather negligible degrees also for **verbs** (from .38 to .40) and **adjectives** (from .28 to .30).[192] Having discovered the noun category would therefore potentially facilitate the

[191] It should be noted that collapsing co-occurrence counts may not be the best way to integrate category information with distributional information, as Redington et al. (1998) point out. And it is possible that more appropriate implementations might yield partly different results.

[192] Interestingly, when the individual nouns are indeed removed from the context lexicon and actually replaced by the new context item (as in the previous experiment on function words), the Distributional Usefulness scores are essentially the same as they are for the added noun cue. The only measurable difference is that determiners and prepositions would benefit even more than they do already from the added noun cue.

acquisition of the categories of determiners and prepositions.[193] And this constitutes another specific prediction about development that calls for experimental investigation.

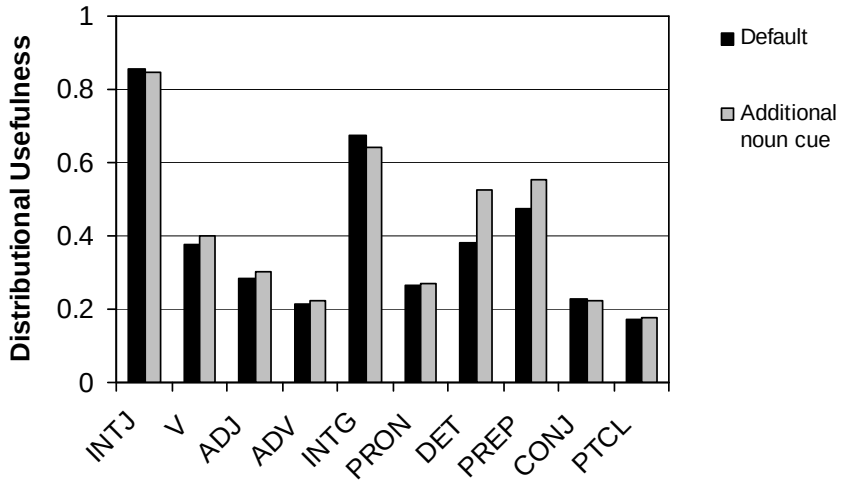

Figure 4-21: *Effects of adding a noun cue*

Distributional Usefulness for the default analysis and when the context lexicon is extended by an additional context item representing the full noun category. Effects on Distributional Usefulness are shown for all target word categories except nouns. Note, however, that nouns were still included in the target lexicon.

With respect to determiners, a plausible explanation for the improved distributional evidence involves grammatical agreement between determiners and nouns. The agreement dependencies are unfavorable not only from the perspective of the noun category (cf. pp. 139f, 179f, 186f), but just as well for the determiner category, because they result in distributional contrasts among determiners which render the overall determiner category more difficult to be acquired from distributional information. Beyond grammatical agreement, there are other ways in which co-occurrence relations of determiners with individual nouns might generate such unfavorable contrasts; for instance, they may arise from the distinction between count nouns and mass nouns, or from semantic and pragmatic factors.[194] The added noun cue ignores any of these

[193] In a similar experiment, Redington et al. (1998) found the overall goodness of co-occurrence information to drop slightly when noun cues were incorporated. The authors did not report results for individual categories; but at the level of the global category system, their result is not immediately consistent with the findings in the current study. At least some of the difference may be accounted for by the fact that grammatical agreement between determiners and nouns, which in the current study is responsible for most for the improvement for the determiner category (see below), but which plays a much smaller role in English (cf. footnote 190, p. 195).

[194] The count–mass noun distinction, coupled with pragmatic factors, influences the co-occurrence preferences of some determiners in the following example, independently of grammatical agreement. All requirements of grammatical agreement are met in the NPs "*keine ahnung*" (English: "*no clue*")

problematic influences but still fully captures the highly informative fact that most determiners in the corpus are frequently followed by some noun. Like this, the added noun cue profoundly ameliorates the distributional information about the overall determiner category; and this can explain the higher Distributional Usefulness score.

Similar considerations also seem to apply to the preposition category and its strong co-occurrence relation with nouns, or rather common nouns, in context position [+2] (relative to the preposition). Grammatical agreement plays virtually no role between prepositions and nouns.[195] It is mainly because of semantic and pragmatic factors, and also because of the count–mass noun distinction, that co-occurrences with nouns may give rise to distributional contrasts among prepositions.[196] Therefore, just as was the case for determiners, the co-occurrence relation of prepositions with nouns is captured by the additional noun cue in such a fashion that all these problematic influences are removed. This accounts for the increase of Distributional Usefulness for the preposition category.

The fact that the increase is even more substantial for determiners presumably arises for at least two reasons. First, the problems that the added noun cue overcomes are more substantial ones for the determiner category, since grammatical agreement effectively does not apply to the relation between prepositions and nouns. Second, the co-occurrence relation of determiners with nouns in context position [+1] simply is a stronger and more reliable cue in itself, compared to the co-occurrence relation between determiners and nouns in [+2] (cf. Table 4-6 on p. 144).

Integrating these findings with the results from previous experiments, a consistent picture emerges that credits the noun category with a special role in the course of

---

and *"die ahnung"* (English: *"the clue"*), but whereas the probability of *keine* to enter this co-occurrence relation with *ahnung* is very high (15.3%), not a single instance of *"die ahnung"* was found in the corpus. A second example illustrates how semantic and pragmatic aspects pertaining to the noun can influence the co-occurrence preferences of certain determiners. Some nouns denote entities for which it is unlikely, at least in CDS, to talk about as belonging to somebody. Therefore, possessive pronouns — they were classified as determiners in the benchmark category system — occur hardly ever with any of these nouns, irrespective of grammatical agreement, whereas other types of determiners might do so very well. A particular example from the corpus is *idee* (English: *idea*) which does not occur after any possessive pronoun form at all while it does appear after an indefinite article (*eine* and its enclitic form *'ne*, English: $a_{\text{fem.sg.:nom.+acc.}}$; $one_{\text{fem.sg.:nom.+acc.}}$), after a demonstrative pronoun (*diese*; English: $this_{\text{fem.sg.:nom.+acc}}$), and after a negated indefinite numeral (*keine*; English: *no*, in the sense of $not\_any_{\text{fem.sg./fem.+masc.+neut.:pl.}}$).

[195] In theory, each preposition requires the dependent NP to take a particular grammatical case. But case marking on nouns is generally rather poor in German, and at least nouns in singular number do not have different forms for dative and accusative, which are the only grammatical cases that are relevant for the prepositions in the target lexicon.

[196] For instance, locative prepositions such as *hinter* (English: *behind*) are relatively unlikely to co-occur with almost any noun that does not denote a physical object, e.g. *abend* (English: *evening*), *aufnahme* (English: *recording*).

category development. First, it is probably acquired earlier than most, or possibly all, other categories; and its discovery is most likely based on various sources of information in the input, including semantic, perceptual, morphological, and lexical-distributional cues. Second, this discovery might be further facilitated by children's early failure to differentiate individual function words. Third, once the noun category is mastered, it might in turn facilitate the acquisition of other categories, most of all determiners and prepositions. A reliable notion of the determiner and the preposition category would then in turn presumably help to extend and refine the noun category.

Chapter 5
General discussion

The main objective of this dissertation was to analyze the distributional evidence that children acquiring German as their first language might find in their input about the major lexical categories underlying this language. A deliberate constraint was to address this question without committing to a particular learning mechanism. Therefore, by contrast to previous work, this approach did not systematically generate hypotheses about the word classes that children might induce from their input. Instead, the focus was on characterizing in detail the distributional information that is available in the input, and to assess how useful this information would be if lexical categories were to be acquired from this very data sample.

To this end, it was crucial to analyze the input to individual children separately, and to use input samples that are as representative as possible of the input that the children actually encountered. To my knowledge, size and sampling rate of the corpus that was used in this study is unprecedented among language acquisition corpora for German. For this reason, it was decided to conduct these analyses as a case study, which in turn raised the issue of how representative any findings might be of the input to German children in general. Some of the following conclusions should therefore be treated with caution until they are verified for the input to other children.

As the most basic result, the current study confirmed for the German input to one child the general findings of earlier studies for English (Redington et al., 1998; Mintz et al., 2002) that distributional information in terms of highly local lexical contexts is potentially useful for acquiring lexical categories, even though German has fewer word order restrictions and more complex inflectional morphology than English. Moreover, it was shown from several complementary perspectives that the distributional evidence is fairly robust even when the underlying data or the computational methods are significantly reduced in size or power. Various useful cues were identified for each category, but categories were found to differ considerably with respect to the number

and usefulness of these cues. For the example of noun and verb category and their relevant subclasses, these cues were analyzed at greater detail and in many cases traced back to the underlying grammatical regularities and usage frequencies at the constructional level. These links underscore the great advantage of using large corpora, as they allow for assessing such interplay between theoretical rules and actual usage.

In particular, these close links offer an explanation why local distributional regularities are at all informative about lexical categories, despite some influential claims to the contrary. These claims are addressed in the following. Next, I make a few preliminary remarks about how children might actually exploit the kind of highly local distributional cues that were assessed here. The dissertation concludes with a list of limitations of the current study, some of which lead to specific proposals for future research.

5.1 The effectiveness of highly local distributional cues

Various concerns can be found in the literature questioning the informativeness of overt lexical-distributional cues to lexical categories, and the feasibility to extract them from the input. The most prominent ones are the five objections brought forward by Pinker (1984; also see Pinker, 1979) that were summarized earlier (pp. 25f). Because these objections were formulated in response to the early pioneering work in the distributional modeling field (in particular, to Maratsos & Chalkley, 1980), it is not surprising that two of them — viz., those concerning the computational and theoretical consequences of a blind search for correlated distributional patterns — do not apply to the more recent family of models that the present study is based on. First, these models do not look for correlations among distributional cues; rather, such correlations emerge in the form of coinciding salient properties each of which sticks out by itself.[197] Note, however, that this question may not depend so much on the distributional information per se but rather on the particular learning mechanism extracting this information from the input (cf. 5.2). Second, and more directly relevant for the available information, the newer models only assess a finite set of co-occurrence relations (viz., a limited number of

[197] The models are sensitive to such salience by virtue of the formal similarity measure since the similarity between two target words — or rather between their SCO vectors — crucially depends on the preferred co-occurrence relations of both words.

context positions and context words), rather than considering all possible distributional cues.

Of course, a finite set of 4,076 context dimensions may still seem quite large. If it would turn out that the computational complexity of the distributional information investigated in this study exceeds the early cognitive capacities of children, this complexity could be reduced without much loss of informativeness as some of the experiments suggest (see subsections 4.2.3, 4.2.4, and 4.3.1). But this might be unnecessary if certain constraints on the relevant learning mechanisms are considered (cf. section 5.2).

Pinker's other three arguments do indeed point to serious challenges to any distributional approach that exploits overt word order regularities. However, these arguments are based entirely on theoretical considerations whereas it is primarily an empirical question what the distributional regularities in real input to children actually look like, and whether these regularities can be extracted in a way that overcomes these challenges. And as will become clear in the following discussion, these claims point in fact to some of the strengths of the distributional paradigm (for related discussions see Mintz et al., 2002; Redington et al., 1998; Finch & Chater, 1994).

The first of these arguments asserts that "most linguistically relevant properties are abstract, pertaining to phrase structure, syntactic categories, grammatical relations, and so on" (Pinker, 1984:49). In Pinker's view, a mechanism that has only access to surface word order is bound to miss these properties. He illustrates this point by the example that "there are grammatically relevant consequences of a word's appearing in the subject noun phrase of a sentence, not of its being in the first serial position of a sentence" (ibid.:49f).

Whether or not one subscribes to this statement, depends crucially on what type of relation is meant by "consequences". While it is undoubtedly true that surface word order does not predict phrase structure in any deterministic way, these two levels are not entirely independent either. Indeed, their relation becomes quite constrained if facts about actual language usage are taken into account, because samples of spoken language are no random subsets of all grammatically possible sentences. Such constraints may be particularly strong for child-directed speech (CDS) which is predominated by very short and structurally rather simple utterances (Snow, 1972; Phillips, 1973; Newport, 1977).

Furthermore, even though CDS utterances are distributed across a fairly broad range of abstract construction types such as declaratives, imperatives, or questions

(Newport, 1977; Newport et al., 1977), there is remarkably little variability at the level of item-specific construction frames (Cameron-Faulkner et al., 2003). This results in a high degree of lexical repetitiveness particularly at the utterance onset, which directly takes up Pinker's example above. Cameron-Faulkner and colleagues found that across their 12 English CDS corpora, 45% of all utterances (excluding fixed performatives) begin with one of just 17 words. The single German corpus used in the current study yields very similar figures: 57% of all multiword utterances (comprising at least three word tokens) begin with one of 20 target words.[198]

More immediately relevant for the challenge of category acquisition, categories turned out to differ substantially with regard to their preference to take the first position in a multiword utterance in the *Leo* corpus. When the corpus is restricted to multiword utterances, the probability of the 17 interrogative words in the target lexicon to occur in this position ranges from 24% to 74%, whereas there is not one noun, determiner, or preposition that occupies this position with more than 22% of its instances. This contrast becomes even more obvious when category medians are compared: The median preference to occur utterance-initially in multiword utterances is 48% for interrogative words, in comparison to 2% for nouns, 6% for determiners, and 9% for prepositions. This single cue would thus in principle suffice to fully discriminate the category of interrogative words from those of nouns, determiners, and prepositions — much in contrast to Pinker's claim.

The low percentages for nouns reflect syntactic structure since at least singular common noun forms, which constitute the lion's share of nouns in the target lexicon, require some kind of determiner word to their left and thus do not occur utterance-initially when the utterance is a complete and grammatically correct sentence. The same is not true of prepositions and determiners, however, since these can appear in the first position for a number of complete sentential structures. The fact that most determiners and prepositions do not do this very often in the corpus can therefore only be explained in terms of speakers' preferences to use constructions (whether they are complete sentences or not) in which these words occupy other positions in the utterance. In the previous chapter, various other examples were encountered for how usage preferences constrain syntactic variation in the input data and thereby result in useful distributional

[198] There is much greater lexical variation at the utterance's end. It requires as many as 200 word form types to account for 57% of all tokens occupying the final position of multiword utterances. One reason is that the majority of utterances end on an open-class word (most frequently adverbs, nouns, and nonfinite verb forms), coupled with the fact that individual open-class words do not occur very frequently in general.

cues in terms of overt word order regularities, though it is not always the entire category that benefits from these cues (as was observed for the verb category and its subclasses, cf. 4.4.1). But even syntactic structure itself can impose strong constraints on the local lexical contexts that a particular word might occur in (as was observed for the case of the noun category and its subclasses, cf. 4.4.2).

The way in which grammatical rules (such as syntactic regularities and agreement relations) and usage preferences interact in shaping the local distributional profiles of words and their categories can be visualized as follows. Suppose that the set of co-occurrence relations that were considered in this study (i.e., the 4,076 context dimensions) is represented as a dense grid of sockets for light bulbs, with each socket corresponding to one particular co-occurrence relation (i.e., one particular context word in one particular context position). If there is one such grid for each target word, then the constraints of grammatical rules for a particular target word have the effect that not all sockets in the corresponding grid actually carry a light bulb. The usage preferences correspond to the wiring and the electrical resistance of each wire such that they define how the available electric current is distributed among the existing light bulbs, with only a few of them shining brightly while the majority are considerably dimmed or switched off entirely. The resulting pattern of light on the grid characterizes the given word's distributional properties. Correspondingly, the current model is only sensitive to where the light is, and especially sensitive to the brighter sources of light. It cannot, however, tell between empty sockets and sockets with light bulbs; but inasmuch as empty sockets do not shine, their locations still have an indirect influence on the overall pattern of light detected by the model.

Of course, not all CDS utterances conform to grammatical rules identified by linguistic description — even though the degree of well-formed utterances appears to be higher for CDS than it is for speech among adults (e.g., Snow, 1972; Phillips, 1973; Newport, 1977). Deviant utterances might generate any conceivable local context for a given target word. Therefore, to stay in the picture, the empty sockets would actually have to contain some light bulbs, but the wires to these bulbs have a very high resistance and are connected by some loose contact such that these bulbs only light up by accident and at best radiate some weak background light — assuming that deviations from grammatical rules do not occur systematically across utterances. When such deviations are indeed systematic, they might simply reflect a linguistic rule that has not yet been acknowledged by linguists — perhaps because it is a rule that is idiosyncratic to the speakers in the child's environment.

The light bulb metaphor illustrates how local distributional cues which are solely based on overt word ordering can provide relevant information about lexical categories, even though these may not be defined in terms of local contexts by linguists. However, this conclusion does not rule out the possibility that some relevant linguistic distinctions might nevertheless be missed by the model, without having access to the phrase structure underlying the individual utterances. And this is precisely the subject of Pinker's second central objection against distributional approaches to category learning. Consider the following sentences, taken from Pinker (1984:49).[199]

(a) John eats meat.
(b) John eats slowly.
(c) The meat is good.
(d) *The slowly is good.

Sentences (a) and (b) are parallel from the perspective of overt word ordering, even though the underlying syntactic structures are fundamentally different. Pinker argues that a distributional learner solely utilizing overt word order regularities might infer from these sentences that *meat* and *slowly* belong to the same lexical category, and therefore, that both can be used in the same kinds of contexts in general. Thus, after encountering *meat* in a sentence such as (c), the learner would erroneously take (d) also to be a valid sentence.

A learning mechanism drawing conclusions from single observations might indeed fall for this trap; but, as Redington et al. (1998:432) point out, this is not an intrinsic problem of distributional approaches per se. In particular, one of the main assumptions underlying the frequency-based model presented here is that the learner bases any inferences on the full spectrum of lexical contexts a given word is encountered in. Therefore, evidence of the types (a) and (b) should prompt the learner to infer that *meat* and *slowly* share some combinatorial properties. At the same time, evidence of type (c) and consistently missing, or at least relatively rare, evidence of type (d) should indicate to the learner that their combinatorial properties differ in systematic ways. Whether or not the learner takes the two words to be members of the same category primarily depends on whether their distributional commonalities outweigh the differences. But

[199] Translating these four sentences in a word-by-word fashion would establish the same example for German.

this is an empirical question, and precisely the kind of question that was investigated extensively in this study.

For the particular case of nouns and adverbs — to return to Pinker's example of *meat* vs. *slowly* — nouns and adverbs do indeed share several of their local distributional properties in the *Leo* corpus (cf. pp. 136f and also Table 4-4 on p. 134). For instance, members of both categories generally prefer to be followed by a verb — as in (c) and (d) above — or to occur in the last two serial positions of an utterance. But at the same time, they differ largely with respect to other properties. Most notably, adverbs do not share the nouns' high preference to be immediately preceded by the various determiners (cumulative preference 3.5% for adverbs vs. 54.9% for nouns) which also reflects that sentences of type (c) are likely to be encountered while sequences of the ungrammatical type (d) are not.[200] As a consequence of the combination of shared and distinct distributional properties, nouns and adverbs separate fairly well from each other in the SCO vector space though they could still separate much better space (separation values are .56 and .50, respectively; cf. Table 4-1 on p. 101). They occupy essentially different regions which, however, overlap.

The *meat* vs. *slowly* example also points to a more fundamental issue which constitutes Pinker's third relevant argument, questioning the feasibility of distributional approaches to category learning. The challenge of category acquisition requires children to generalize beyond their input; but at the same time, this process has to be constrained in order to prevent children from ending up with too general categories. According to Pinker (1984), the input could theoretically provide suitable constraints only in the form of *explicit negative evidence*, that is, by identifying to the child the kinds of sentence-level constructions in which a particular word cannot be used; e.g., adverbs cannot be used in utterances such as (d). And inasmuch as the input is generally lacking such overt negative information, an acquisition mechanism solely relying on the input would essentially be unconstrained and should therefore be bound to result in overly general categories.[201] This position has been attacked by a number of counterarguments (cf. pp. 18f). Most relevant in the given context is the alternative view that the child

[200] Finch and Chater (1994) found in an unspecified corpus of English that the sequence *DET ADV is* — abstracted from the lexical sequence *"the slowly is"* in Pinker's example (d) — does indeed occur but is by far less likely to be encountered than would be expected from mere chance.

[201] The specific example that Pinker (1984:48f) offers, only applies to Maratsos and Chalkley's (1980) equation of the distributional properties that drive the acquisition of lexical categories and the properties that drive their later usage in language production and comprehension. By contrast, this equation is generally not made for the type of distributional cues to which models such as the current one are sensitive. Nevertheless, the general issue of overgeneralization potentially applies to current models as well.

might also utilize the systematic absence of utterances of type (d) from his input as a form of *implicit negative evidence* that these are ungrammatical (e.g., Finch & Chater, 1994).

This view was rejected by Pinker (1984:48). He argues that the child could never know for sure whether the absence of utterances such as (d) indicates that (d) is truly ungrammatical (reflecting what Elman, 2002, called a *systematic gap*) or is in fact grammatical but did simply not yet occur in the input (reflecting an *accidental gap*). Given speakers' highly skewed usage preferences across the set of possible structures, accidental gaps are a very common phenomenon in children's input, such that distinguishing them from systematic gaps constitutes indeed a critical challenge for the acquisition of categories, and of grammar in general. Nevertheless, it was demonstrated in connectionist simulations that this distinction can in principle be discovered from overt distributional information, as a function of the learning mechanism's distributional experience over time (Elman, 2002).

From the perspective of strictly local lexical contexts, however, the situation is even more complex, because at this level, grammatical judgments and explicit negative evidence are not only unavailable to the child, but these concepts do not even exist. Without considering the overall utterance, it would in most cases simply not be meaningful to ask whether or not it is *grammatical* for a particular word to occur in a particular local context. Instead what matters within the current model is how *likely* the word does so, and the relative co-occurrence frequencies observed in children's input provide estimates for these probabilities. As the earlier example for adverbs and nouns illustrated, quantitative differences in these relative frequencies can serve to constrain generalization in a learning mechanism that exploits local distributional information. In this sense, at the level of local contexts, such differences are the closest equivalent to (implicit) negative evidence. And it is in this way that observed frequencies serve a very similar function for the model as do grammaticality judgments for linguists (cf. pp. 28f).

Nevertheless, inasmuch as distributional cues are no perfect predictors of lexical category, they can also lead the learner to temporary overgeneralizations (e.g., particles and adverbs are distributionally indistinguishable), and at the same time cause the learner to miss some other generalizations that would actually be appropriate (e.g., the fact that nonfinite and finite verb forms all belong to the verb category). A variety of possible causes for these deficiencies of local distributional information are discussed in section 5.3.

5.2 Statistical learning

One of the main contributions of this dissertation is to have empirically determined a detailed profile of distributional cues to each major category, and for nouns and verbs also to some relevant subclasses. Provided that these profiles will prove to be representative of German CDS in general, one of their benefits is that they suggest some specific predictions about development which need to be tested experimentally (e.g., the two predictions formulated in 4.4.3, and the prediction in 4.5.2). A more fundamental benefit from the profiles is that they specify at the level of individual context words which co-occurrence relations constitute strong positive or negative cues to a particular category. And further experimental studies will have to determine the extent to which children actually rely on these cues in the process of category development.

But in practice, it is often difficult to design such experiments in a way that allows for testing individual cues in isolation. The study by Höhle et al. (2004) that was described earlier (p. 189) is an example in which even very young children were successfully shown to use one specific cue to the noun category (or rather to the subclasses of masculine and neuter singular nouns) for early categorization. However, this finding is only relevant for the second subtask of category acquisition (p. 19), namely the mapping of novel words onto some previously acquired proto-category of noun; but it does not demonstrate that the children relied on this cue for discovering this proto-category in the first place — and the authors did not aim to provide such a demonstration. In fact, tracking the influence of specific cues through the process of category development in first language acquisition might even be impossible because one would have to control for all the experience that a child has with the cues in question, a requirement that is unworkable not only for ethical reasons.

A common solution to this dilemma therefore is to test children's language learning mechanisms on artificial toy languages. Because children come to the experiment entirely inexperienced with the artificial language, the investigator can fully control the experience on which any learning from this grammar is based. The draw-back is that such artificial language learning experiments can only demonstrate that children are capable of exploiting certain types of cues in principle, but they cannot show that children actually do so in acquiring their first language. In particular, they cannot assess the specific contribution of the individual cues that were uncovered by the current study.

Nevertheless, the findings that this experimental paradigm yielded so far are more than reassuring. From very early on, infants appear to be very sensitive to overt and highly local co-occurrence regularities in their input; moreover, they appear to be able to employ this sensitivity in a variety of acquisition tasks, including the formation of abstract lexical categories (for review, see Gómez & Gerken, 2000).

Conflicting evidence might seem to follow from artificial language experiments by Smith (1966, 1969). After having been trained on two-word sentences of the form MN and PQ, where M, N, P, and Q denote four non-overlapping classes of words, adult participants readily learned which words can occur in the first position (M and P) and which in the second (N and Q); but they failed to discover the class distinctions M vs. P and N vs. Q such that they also considered sequences of the form MQ or PN as admissible sentences of the language.[202] Thus, even adults have great difficulty to learn this particular type of co-occurrence relations between abstract word classes.[203]

However, such "MN/PQ structures" (Braine, 1987) do not describe the type of distributional cues investigated in the present study, for these cues are defined as co-occurrence relations with individual context words, but not with classes of words. In Smith's experimental setting, an idealized cue of this kind would correspond to restricting M and P to one single word each, such that they function as two fully reliable lexical markers to categories N and Q. And with this modification, both adults and infants would most certainly have no difficulty to discover the distinction N vs. Q after having been trained the same way as in Smith's experiment.

Of course, realistic cues are not as reliable as such an idealized cue. But the results of Smith's study imply that assessing the informativeness of cues in terms of individual context words, as done by the current model, is the appropriate level of analysis. In particular, this means that the detailed lists of distributional cues given in Appendices D through F are the relevant profiles to be analyzed, whereas the corresponding profiles at the cumulative level (cf. Table 4-4, Table 4-6, Table 4-9, and Table 4-11) can only serve to structure this analysis in a linguistically meaningful way. And this describes precisely how the analyses in sections 4.3 and 4.4 proceeded.

But beyond justifying the research strategies taken in the current study, Smith's findings also provide experimental support for the conclusion of this study that the phenomenon of grammatical agreement places a burden on the acquisition of noun and

[202] In the actual experiments, Smith used single letters instead of words.
[203] Monaghan, Chater, and Christiansen (2005) question this conclusion, arguing that it is not warranted by the testing procedures that Smith applied.

verb category from distributional information, because agreement complicates the co-occurrence relations of (common) nouns with determiners, and likewise of (finite) verbs with (subject) pronouns (cf. 4.4). At the same time, other artificial language learning studies indicate realistic ways by which infants might overcome distributional difficulties of this kind (e.g., Braine, 1987; Gerken, Wilson, & Lewis, 2005). And these studies in turn also constitute reassuring — though not sufficient — evidence concerning another prediction derived from the current study, namely, that the acquisition of the noun category should help children to subsequently discover the categories of determiners and prepositions (cf. 4.5.2).

In sum, existing evidence from artificial language learning experiments suggests that children possess cognitive mechanisms by which they can in principle develop word classes from precisely the type of distributional cues that were investigated in this dissertation. And it is therefore entirely plausible that they capitalize on such cues in acquiring the categories of their first language. Nevertheless, more experimental studies are needed to assess the extent to which children are able to base their early categories on specific types of cues as identified in this dissertation, as far as this is possible within existing experimental paradigms.

As was just argued, the *type* of information extracted by the present distributional model appears to involve no unrealistic assumptions about the child. But the extracted *amount* of information might very well exceed children's cognitive capacities. At least early on, it seems unlikely that children would utilize information from more than 4,000 different co-occurrence relations as does the model (more than 1,000 context words in four context positions). However, the analyses showed from several perspectives that the usefulness of the extracted information does not crucially depend on this large number. Recall that the dominant (positive and negative) cues were found only among a fairly small set of highly frequent — and therefore potentially salient — context words and utterance boundary markers (p. 151). In consequence, confining the model to co-occurrences with these few items yielded nearly as useful distributional information as did the full model (subsection 4.2.3). And the computational complexity could even be further reduced without much loss in informativeness, by moving from exact co-occurrence counts to sloppy counting (subsection 4.2.4). The most useful cues and their informativeness are simply not affected very much by these modifications.

Positing that children substantially capitalize on distributional cues to lexical categories would therefore not appear to involve unrealistic assumptions about the amount of relevant information that children would have to extract from the input. But

making claims about the information utilized by children does not entail any claims about the cognitive mechanisms by which they do so. In particular, the specific model used in this study only served as an investigative tool; but it is not meant to suggest that children are literally counting co-occurrences between words, no matter whether it is by exact or sloppy counting. Instead, I propose to think of co-occurrence vectors as summarizing some crucial aspects of a child's distributional experience with individual words. This experience may guide learning in the child such that distributional patterns gradually shape the way the child processes and uses language.

This notion corresponds to what Elman (2002) termed *statistically driven learning*. The inherent statistics of the input have an effect on learning and future processing, but the statistics themselves need not be stored by the child: Any input utterance can be forgotten the very next moment. By contrast, the formal co-occurrence model used in this study performs what Elman called *learning of statistics*, for it accumulates all relevant co-occurrences that it encounters for a given target word, and in this way, it gradually builds up an explicit statistical representation (viz., the co-occurrence vector) for the distributional properties of this word. The fundamental distinction between these two notions of probabilistic learning is critical for putting the current study in proper perspective.

But how might a cognitively plausible category acquisition mechanism look like that is based on statistically driven learning? One type of formal learning mechanisms that comes to mind as a first approximation is the general framework of connectionist networks, because these models implement statistically driven learning by definition. And a possible avenue for future research is to devise a plausible network architecture that can exploit co-occurrence patterns to develop lexical categories from CDS samples. A good starting point would be the *simple recurrent network* (SRN; Elman, 1990) which was demonstrated to have these capacities in principle (cf. p. 30). As one of its great advantages for current purposes, the model only discovers co-occurrence regularities that are useful to achieve its particular learning task. After successful learning, detailed inspection of the specific regularities that the model did pick up can therefore provide insights into the kinds of distributional cues that may play a crucial role in category acquisition. However, until now most SRN simulations investigating category acquisition have only involved artificial languages with small vocabularies and few grammatical rules. It us thus an open question whether SRNs will scale up readily to learn from a full-blown corpus of spoken language of the size as used in this study (for similar concerns, see Redington et al., 1998:434).

5.3 Limitations of the current approach

Even if statistically driven learning is plausible and children make extensive use of it, there obviously are limits to learning from the lexical co-occurrence patterns that were identified in this study. Distributional Usefulness did not reach the maximal score 1.0 for any single benchmark category; in fact, some of them achieved rather low scores. The kind of distributional regularities that the model extracted from the *Leo* corpus provides useful cues to all categories, but these cues clearly do not suffice for accurately acquiring the full benchmark category system.

There are a number of potential explanations for why the observed Distributional Usefulness levels were not higher. In a nutshell, they concern the four general possibilities (i) that this particular model fails to pick up distributional information that is available in the corpus, (ii) that the corpus is not fully representative of the linguistic evidence that the child is exposed to and that he can actually discern, (iii) that the benchmark category system is not composed of the right categories, and (iv) that distributional information per se is in some ways intrinsically insufficient. Explanations of all four types identify possible limitations of the current study and in most cases directly lead to specific proposals for how these limitations might be overcome. The most relevant limitations of all four types are given in the following.

Potentially available distributional information missed by the model

One clear drawback of the current model and the associated evaluation scheme is that categorial ambiguity is not taken into account. Because each target word was assigned to only one benchmark category, ambiguous words blur the distributional separation between the various categories they can instantiate. This became obvious in the extended confusability analyses in subsection 4.1.2 where separation scores improved when ambiguous words were removed from the analysis. But this removal served only as a means of investigation and constitutes no solution to the issue because children have to deal with ambiguity as well. Nevertheless, these findings strongly suggest that the usefulness of distributional information would improve if both model and evaluation scheme were modified to accommodate categorial ambiguity in a more appropriate way. But this is no trivial challenge. One distributional approach that can in principle deal with ambiguity is the incremental learning mechanism proposed by Cartwright and

Brent (1997) which implements a notion of distributional information that differs considerably from the one analyzed in the current study (cf. p. 30).

A distributional solution more closely related to the current model was developed by Schütze (1998). Although it is targeted primarily at ambiguity with respect to word sense, it could be modified slightly to be more sensitive to categorial ambiguity. The key idea to this approach is that it applies *second-order co-occurrence statistics* to determine a distributional representation for each individual target word token. The various tokens of a given target word are then subjected to cluster analysis by which tokens with similar distributional properties are assigned to the same cluster. Different clusters represent different lemmas (i.e., senses) of the given target word. Extending Schütze's work, these unlabeled lemmas could then be fed, as separate target words, into the model used in the current study such that an independent SCO vector would be derived for each of them. However, while these computational steps are fairly easy to implement, the real challenge would be to build a reasonable benchmark classification for the automatically derived lemmas against which these vectors could be evaluated.

A second explanation for the nonmaximal Distributional Usefulness scores is that the current model might blur some distributional structure by considering local contexts anywhere within an utterance. Braine (1987) proposed to apply distributional approaches to co-occurrences only within salient phrasal units rather than within full utterances. It was already noted earlier that the linguistic input to children contains a variety of reasonably reliable prosodic cues to phrase boundaries (cf. p. 22); and at least for English, it was shown that children as young as 9 months are sensitive to such cues (Jusczyk et al., 1992). And from around 16 months onwards, the prosodic cues are augmented by children's ability to detect function words which may then serve as markers to phrase boundaries (Shady, 1997). Braine's proposal therefore constitutes a realistic constraint by which statistical noise from co-occurrences across phrase boundaries would be removed from the current model. On the other hand, some cross-phrasal co-occurrence relations might be very informative about lexical categories, and these would be removed as well. Therefore, it is an open question to which extent Distributional Usefulness would increase for the various categories. Promising explorations in this direction by Mintz et al. (2002) demonstrated for English that even a rough notion of phrase boundary ameliorates the distributional information at least about the categories noun and verb; but especially for the case of verbs, it is unclear how well these results would transfer to German because the internal structure of VPs

differ in crucial ways between both languages (also see p. 168). Therefore, corresponding explorations for German CDS would be worthwhile to pursue.

A third way in which the current model might miss distributional information that is actually available in the corpus concerns the fact that it extracts distributional cues separately for each of the four context positions. In consequence, any systematic dependencies between cues from different context positions are ignored.[204] For instance, both subject pronouns and finite forms of transitive verbs frequently occur in utterance-initial position, and they also share a preference to be followed by an accusative determiner form (such as *den*; English: *the*$_{\text{masc.+neut:sg.acc./fem.+masc.+neut.:pl.dat.}}$) or an accusative pronoun form (such as *ihn*; English: *him*$_{\text{acc.}}$). However, for the finite verbs (esp. imperative singular forms), these are two statistically independent co-occurrence events: How likely it is for such a verb to be followed by an accusative determiner, does essentially not depend on whether it occurs utterance-initially or not, and vice versa. By contrast, for subject pronouns, there is a strong dependency between both kinds of co-occurrence relations for they practically never enter both of these relations simultaneously, that is, for the same pronoun tokens. When they occur utterance-initially, they are almost exclusively followed by a verb and thus virtually never by determiners or other pronouns. Conversely, they can be followed by an accusative determiner or pronoun in inverted declaratives, in questions, and in subclauses, but in each of these cases, they do not occupy the utterance-initial position.

The distributional separation between finite verbs and (subject) pronouns would therefore be expected to improve if the model were given access to the dependency between these co-occurrence relations. This could be achieved, for instance, by assessing all four context positions in combination. That is, distributional cues would consist of simultaneous co-occurrences with fixed four-word contexts.

This possibility was explored empirically by Mintz (2003), except that he worked with a smaller context window, comprising only the two context positions [−1] and [+1]. Distributional cues were thus defined as simultaneous co-occurrences with fixed pairs of words in these two context positions. For statistical and cognitive considerations, it is reasonable to restrict the analysis of such cues to only those word pairs that occur fairly frequently in these two context positions of any word — just as only frequent word forms were considered as potential cues in the current study. Word

[204] The model already does benefit from the statistical interaction between cues in different context positions (cf. pp. 130f, 156). But the systematic dependencies across context positions that are referred to here constitute additional information beyond simple interaction.

pairs that qualify by such a frequency criterion constitute what Mintz termed *frequent frames*. His findings demonstrate that grouping together all words that co-occur with the same frequent frames (e.g., all words that are observed to fill the empty slot of the frame *"the __ is"*) results in very accurate word classes; that is, words in the same frame-based class generally belong to the same lexical category. On the other hand, the degree of generalization achieved in this way is rather limited as each lexical category corresponds to several frame-based classes, and only a small portion of category members appears at all in any of these classes (cf. Monaghan & Christiansen, 2004). Distributional cues from frequent frames might therefore be most useful at the onset of category acquisition, contributing to the discovery of early proto-categories around which the adult categories can gradually develop when other distributional and nondistributional evidence is also exploited.

Different learning situations for child and model

The corpus data underlying this study and the fashion of how they were fed into the model might differ in important ways from the actual learning situation that the child is faced with. One obvious difference concerns the amount of available evidence. Even though the *Leo* corpus has an exceedingly greater sampling rate than any corpus of German CDS covering a comparable period of time (three years), it still does not capture the full input that the child was exposed to during the same period. The findings in subsection 4.2.2 indicate that usefulness of distributional information would improve for all categories if this information were extracted from larger samples of this full input.

On the other hand, children do not wait three years to take in all input and only then begin to learn something from it. As experimental evidence suggests (e.g., Tomasello & Olguin, 1993; Höhle et al., 2004), children's first categories are based on the input of a shorter period of time; and, at least as importantly, this period begins well before their second birthday.[205] A substantial portion of the relevant developmental period is therefore not covered at all by the Leo corpus. This constitutes a limitation only if one assumes that the distributional regularities available in the input change over time. And it seems quite likely that this is the case. As the child's language skills develop, caretaker's utterances tend to involve a larger vocabulary, and to become syntactically more complex — a superficial indicator of the syntactic changes is the

[205] For similar considerations, Mintz et al. (2002) confined their analyses to children's input before age 2;6.

increasing average length of input utterances as the child gets older (cf. Figure 2-3 on p. 42). The distributional consequences and the time course of such changes are not clear, but it seems quite likely that a simpler early input would provide the child with more reliable distributional cues at least for some categories, which might further aid the child to break into the category system. In any case, an appropriate empirical assessment of this issue would have to be based on input corpora that cover several years and begin only a few months after the children are born.

Another limitation concerns not so much the available input but rather how well the child can discern this input. For instance, recall from 4.5.1 that there might be a developmental stage in which children are sensitive to the occurrence of function words in their input but do not yet differentiate between individual function words. Additionally, there might also be many content words in their input that children do not readily acquire as lexical items. Intuitively, a child's sensitivity to the distributional properties of a particular word should be very limited until he has developed a lexical representation for this word. Therefore, in order to assess the relevant distributional information that a child is able to discern at any developmental point, the co-occurrence model should be restricted to the child's (productive or receptive) vocabulary at that point. It is a realistic possibility, that this would improve the distributional situation for some categories. A likely candidate is the noun category because nouns tend to predominate in children's early vocabularies.

Limitations concerning the benchmark category system

The distributional structure was not found to completely reflect the benchmark category system, and all explanations offered so far concerned the data from which this structure was derived, and the particular mechanism by which it was derived. Additionally, it is also reasonable to question the benchmark categories as the approximated endpoint of category acquisition. They were used in the current study only as a preliminary heuristic, and it is very likely that they do not accurately describe the categories that are used by adults (cf. 2.2.4). The benchmark categories and classification criteria were derived from standard dictionaries that essentially capture the classical category distinctions. However, there continues to be considerable debate on the nature and partitioning of lexical categories, both from a linguistic and a cognitive perspective.

For instance, a well-established observation from typological research is that for languages like English and German, nonfinite verb forms display various properties of *nouniness* whereas finite verb forms generally do not (e.g., Ross, 1972; Sasse, 1993).

And in Langacker's semantic-conceptual characterization of major lexical categories, nonfinite verb forms are not even considered verbs at all (Langacker, 1987:75-78). The current findings line up nicely with this research in that nonfinite verb forms are distributionally very different from finite verb forms and more similar to nouns and adjectives. This example suggests that some unexpected distributional substructure may in fact be linguistically meaningful. Earlier, we also observed an example for the complementary case in which distributional structure fails to reflect some benchmark category distinctions that are linguistically debatable (adverbs vs. particles, cf. pp. 107f). Seen in this light, the nonmaximal Distributional Usefulness scores might partly express that distributional structure constitutes in some ways a better approximation of the real categories than do the benchmark classifications.

Intrinsic insufficiency of highly local distributional information

Suppose the model, the corpus, and the benchmark category system could be modified in such a way that all limiting factors considered so far would disappear. Then the overall usefulness of the extracted distributional information should improve considerably, but it is unlikely that it would rise as high as to fully predict the (modified) benchmark categories. And this deviation from full prediction can be attributed to local distributional information per se. The cues that children can extract from overt and local word order regularities in their input do presumably not suffice for acquiring the adult categories.

As the debates about the linguistic criteria for defining lexical categories illustrate (cf. 1.1.1), the relevant category distinctions cannot be made solely in terms of the words' combinatorial properties — morphological and semantic contrasts are needed as well (cf. Maratsos, 1990; Behrens, 2005). It would therefore not be surprising if children have to utilize information at all these linguistic levels as well, in order to acquire lexical categories. And, taking the argument even further, there is no a-priori reason why children should refrain from also using other kinds of available information, such as perceptual and pragmatic cues. Any kinds of regularities in their input that may help them to become more successful in understanding and using their first language would seem to be realistic candidates for contributing to the process of category formation, provided that children have the cognitive prerequisites to exploit these regularities.

This assumption characterizes most of the more recent empirical work concerned with category acquisition (cf. p. 26). Several research groups have begun to determine

more systematically at each of these different linguistic levels (lexical-distributional, morphological-distributional, pragmatic-semantic, and perceptual) the kinds of cues that children can find in their input, and the extent to which children can and do exploit these cues for category acquisition.[206] The currently available findings are more than reassuring — it seems indeed a viable possibility that the cues from these four levels would together suffice for acquiring the lexical categories of the target language.

The current study by itself already provides examples for how the combination, and especially the interaction, of cues from different sources can potentially facilitate the acquisition of some categories beyond the contribution of either source in isolation. One such example is the experiment in subsection 4.5.1 where perceptual cues to the distinction between content and function words were indirectly integrated into the co-occurrence model which resulted in a considerable improvement of the distributional evidence for the noun category. A second example that concerns the verb category is more speculative. Lexical-distributional cues were found to constitute very reliable evidence about the class of finite verb forms, and even more reliable evidence about some of the inflectional subclasses. While morphological-distributional cues are likely to further support the acquisition of these classes as separate categories, semantic and perceptual cues would presumably be a good basis for discovering the links between the different inflected forms of a verb lexeme, most reliably for regular verbs. By these links, the isolated verb subclasses could then be joined to one unified verb category.

However, in both these examples, cues from two different sources were combined in a two-stage fashion, by importing the outcome of learning from one source into the learning procedures that are sensitive to the other source. A similar two-stage approach was taken by Cartwright and Brent (1997) in an experiment where their distributional model gradually builds up categories around some pre-existing semantically-based word classes. These two-stage approaches necessarily only consider unidirectional interactions between cues from different sources. Clearly, more progress would be made by systematically assessing the bi- and multidirectional interactions between cues from different sources.

Promising advances in this direction are, for instance, the studies by Christiansen and Dale (2001) and by Monaghan, Chater, and Christiansen (2005) who directly

[206] For the case of perceptual cues, subsection 1.2.2 already summarized empirical evidence, indicating for several languages that a combination of such cues alone can provide a reliable basis for discriminating function words from content words, and nouns from verbs. Furthermore, it was shown that both adults and children are very sensitive to such correlations between a word's category and its perceptual properties (cf. Kelly, 1992, 1996; Farmer, Christiansen, & Monaghan, 2006).

integrated distributional and perceptual cues into a single computational model. Interestingly, the latter study suggests that the two sources of information are complementary in the sense that distributional cues are more reliable for highly frequent words whereas phonological cues are more reliable for less frequent words.

But this does not imply a general *division of labor* between the different sources of information in the input. In fact, there is a considerable degree of redundancy — that is, correlation — between cues from different sources. This becomes apparent even within the current study, for although the model is exclusively based on distributional information, the resulting SCO vector space nonetheless also displays some degree of morphological and semantic organization (for similar observations see Redington et al., 1998). For instance, verbs were found to form clusters corresponding to their inflectional subclasses; proper names clustered into names for places and names for individuals; and also temporal adverbs, locative adverbs, and color adjectives each turned out to form three fairly tight clusters by themselves.

Redundancy might first seem to limit the benefit from combining cues because it implies less additional information. But the exact opposite is the case, for redundancy is in several ways a crucial and very useful feature of the input. First, correlated cues (within and across different sources) presumably make the available information more salient. Second, they are likely to promote the robustness of the acquisition process. Third and maybe most importantly, experimental evidence indicates that such correlated cues — provided that the correlations are only partial — can help children to also exploit information from the otherwise difficult MN/PQ structures which were briefly discussed above, and furthermore, that children indeed heavily capitalize on such correlations (cf. Gómez & Gerken, 2000; Gerken et al., 2005). And the current study further suggests that such partial correlations arise almost inevitably if negative cues are taken into account (cf. 4.3.4).

5.4 Conclusion

Despite considerable advances over the past few decades, our current knowledge of how children acquire the lexical categories underlying their first language is still fairly limited. It remains unclear on what specific evidence in their input children rely, at which time course they exploit the various types of evidence, and which learning

mechanisms they employ in doing so. Whereas earlier research has generated specific and contrasting hypotheses about these issues, the field has more recently converged on the view that all available sources of information might play a role at any developmental point, and that progress can be made by studying these sources both in isolation and in combination. This research strategy can lead to hypotheses about the available cues that children might actually use, and testing these hypotheses in controlled experiments in turn is likely to provide insights into the underlying learning mechanisms. With this dissertation, I hope to have made a contribution to this overall research program.

das hat viel spass gemacht ne leo ?

(taken from corpus)

References

Altmann, G. T. M. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences, 2* (4), 146-152.

Altmann, G. T. M., Nice, K. Y. van, Garnham, A., & Henstra, J.-A. (1998). Late closure in context. *Journal of Memory and Language, 38* (4), 459-484.

Bates, E., & Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes, 12* (5/6), 507-584.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 173-218). Cambridge: Cambridge University Press.

Behrens, H. (1993). *Temporal reference in German child language: Form and function in early verb use*. Doctoral dissertation, Universiteit van Amsterdam, The Netherlands.

Behrens, H. (2002). Learning multiple regularities: Evidence from overgeneralization errors in the German plural. In B. Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development* (BUCLD 26, Vol. 1, pp. 72-83). Somerville, Mass.: Cascadilla Press.

Behrens, H. (2003). Bedeutungserwerb, Grammatikalisierung und Polysemie: Zum Erwerb von ‘gehen’ im Deutschen, Niederländischen und Englischen. In S. Haberzettl & H. Wegener (Eds.), *Spracherwerb und Konzeptualisierung* (pp. 161-181). Frankfurt/Main: Lang.

Behrens, H. (2005). Wortarten-Erwerb durch Induktion. In C. Knobloch & B. Schaeder (Eds.), *Wortarten und Grammatikalisierung: Perspektiven in System und Erwerb* (pp. 177-198). Berlin: de Gruyter.

Belew, R. K. (2001). *Finding out about: A cognitive perspective on search engine technology and the WWW*. Cambridge: Cambridge University Press.

Berman, R. A. (1991). In defense of development [Open peer commentary of S. Crain's article "Language acquisition in the absence of experience" in BBS 14, 597-612]. *Behavioral and Brain Sciences, 14* (4), 612-613.

Bierwisch, M. (1981). Linguistics and language error. In *Linguistics* (A shorter German version was published 1970 in Linguistic Inquiry, 1 (4), 397-414).

Bloom, L. (1970). *Language development: Form and function in emerging grammars.* Cambridge, Mass.: MIT Press.

Bohannon, J. N., III, MacWhinney, B., & Snow, C. (1990). No negative evidence revisited: Beyond learnability or who has to prove what to whom. *Developmental Psychology, 26* (2), 221-226.

Borovsky, A., & Elman, J. L. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of Child Language, 33* (4), 759-790.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. A. Hawkins (Ed.), *Explaining language universals* (pp. 73-101). Oxford: Basil Blackwell.

Braine, M. D. S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65-87). Hillsdale, New Jersey: Erlbaum.

Braine, M. D. S. (1988). Modeling the acquisition of linguistic structure. In Y. Levy, I. M. Schlesinger, & M. D. S. Braine (Eds.), *Categories and processes in language acquisition* (pp. 217-259). Hillsdale, New Jersey: Erlbaum.

Braine, M. D. S. (1992). What sort of innate structure is needed to "bootstrap" into syntax? *Cognition, 45*, 77-100.

Brill, E. (1993). A corpus-based approach to language learning (Doctoral dissertation 1993, University of Pennsylvania, IRCS Report 93-44). *Dissertation Abstracts International, 54* (06), 3177-B.

Brown, R., & Fraser, C. (1963). The acquisition of syntax. In C. N. Cofer & B. S. Musgrave (Eds.), *Verbal behaviour and learning: Problems and processes* (pp. 158-201). New York: McGraw-Hill.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5* (4), 325-337.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12* (2/3), 177-210.

Bußmann, H. (1990). *Lexikon der Sprachwissenschaft* (2nd ed.). Stuttgart: Kröner.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27* (6), 843-873.

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

Chater, N., & Vitányi, P. (2004). A simplicity principle for language acquisition: Re-evaluating what can be learned from positive evidence. Manuscript in preparation.

Chen, S., & Bates, E. (1998). The dissociation between nouns and verbs in Broca's and Wernicke's aphasia: Findings from Chinese. *Aphasiology*, *12* (1), 5-36.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures.* Dordrecht: Foris.

Christiansen, M. H., & Dale, R. A. C. (2001). Integrating distributional, prosodic, and phonological information in a connectionist model of language acquisition. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 220-225). Mahwah, New Jersey: Erlbaum.

Clahsen, H. (1982). *Spracherwerb in der Kindheit: Eine Untersuchung zur Entwicklung der Syntax bei Kleinkindern.* Tübingen: Narr (Doctoral Dissertation 1981, Universität-Gesamthochschule Wuppertal, Germany).

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14* (4), 597-612.

Croft, W. (1991). *Syntactic categories and grammatical relations: The cognitive organization of information.* Chicago: University of Chicago Press.

Croft, W. (2001). *Radical Construction Grammar: Syntactic theory in typological perspective.* Oxford: Oxford University Press.

Croft, W. (2005). Logical and typological arguments for Radical Construction Grammar. In J.-O. Östman & M. Fried (Eds.), *Construction Grammars: Cognitive grounding and theoretical extensions* (pp. 273-314). Amsterdam: Benjamins.

Culicover, P. W. (1999). *Syntactic nuts: Hard cases, syntactic theory, and language acquisition.* Oxford: Oxford University Press.

Dagan, I., Lee, L., & Pereira, F. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (ACL). [Simultaneously the 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL)].

Dagan, I., Lee, L., & Pereira, F. C. N. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, *34*, 43-69.

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), *90*, 4957-4960.

Daniele, A., Giustolisi, L., Silveri, M. C., Colosimo, C., & Gainotti, G. (1994). Evidence for a possible neuroanatomical basis for lexical processing of nouns and verbs. *Neuropsychologia, 32* (11), 1325-1341.

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore: The Johns Hopkins Press.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20* (6), 611-629.

Duden: Deutsches Universalwörterbuch (5th ed.). (2003). [CD-ROM]. Mannheim: Dudenverlag.

Duden: Grammatik der deutschen Gegenwartssprach (6th ed.). (1998). Mannheim: Dudenverlag.

Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1, pp. 189-229). Amsterdam: Benjamins.

Eisenbeiß, S. (2002). *Merkmalsgesteuerter Grammatikerwerb: Eine Untersuchung zum Erwerb der Struktur und Flexion der Nominalphrase.* Unpublished doctoral dissertation, Universität Düsseldorf, Germany.

Eisenberg, P. (2000). *Grundriß der deutschen Grammatik* (Rev. ed.). Stuttgart: Metzler.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14* (2), 179-211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7* (2/3), 195-224.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48* (1), 71-99.

Elman, J. L. (2002). Generalization from sparse input. In M. Andronis, E. Debenport, A. Pycha, & K. Yoshimura (Eds.), *Proceedings of the 38th Annual Meeting of the Chicago Linguistic Society. Vol. 2: The panels* (pp. 175-200). Chicago: CLS.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8* (7), 301-306.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, Mass.: MIT Press.

Engel, U. (1996). *Deutsche Grammatik* (3rd ed.). Heidelberg: Groos.

Ervin, S. M. (1961). Changes with age in the verbal determinants of word-association. *American Journal of Psychology, 74*, 361-372.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103 (32), 12203-12208.

Federmeier, K. D., & Bates, E. (1997). Contexts that pack a punch: Lexical class priming of picture naming. In *Center for Research in Language Newsletter* (Vol. 11, pp (2)). La Jolla, Cal.: University of California, San Diego. Retrieved from ftp://ftp.crl.ucsd.edu/pub/newsletter/pdf/11-2.pdf.

Federmeier, K. D., Segal, J. B., Lombrozo, T., & Kutas, M. (2000). Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*, *123* (12), 2552-2566.

Finch, S., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society* (pp. 301-306). Hillsdale, New Jersey: Erlbaum.

Finch, S. P. (1993). *Finding structure in language.* Unpublished doctoral dissertation, Center for Cognitive Science, University of Edinburgh, Scotland.

Fisher, C., & Tokura, H. (1996). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 343-363). Mahwah, New Jersey: Erlbaum.

Friederici, A. D., & Saddy, D. (1993). Disorders of word class processing in aphasia. In G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, C.-W. Wallesch (Ed.), *Handbooks of linguistics and communication science. Vol. 8: Linguistic disorders and pathologies: An international handbook* (pp. 169-181). Berlin: de Gruyter.

Fries, C. C. (1957). *The structure of English: An introduction to the construction of English sentences.* London: Longmans, Green and Co (Original work published 1952, New York: Harcourt, Brace and Co.).

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47* (1), 27-52.

Garrett, M. F. (1975). The analysis of sentence production. *The psychology of learning and motivation: Advances in research and theory*, *9*, 133-177.

Gerken, LA., Landau, B., & Remez, R. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology, 26* (2), 204-216.

Gerken, LA., & McIntosh, B. J. (1993). The interplay of function morphemes and prosody in early language. *Developmental Psychology, 29* (3), 448-457.

Gerken, LA., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32* (2), 249-268.

Gibson, E. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language, 54* (3), 363-388.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10* (5), 447-474.

Gómez, R. L., & Gerken, LA. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4* (5), 178-186.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 165-182). Cambridge, Mass.: MIT Press.

Harris, Z. S. (1946). From morpheme to utterance. *Language, 22* (3), 161-183.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, Maryland: Brookes.

Haspelmath, M. (2001). Word classes and parts of speech. In *International encyclopedia of the social and behavioral sciences* (Vol. 24, pp. 16538-16545). Oxford: Elsevier.

Helbig, G. (1994). *Lexikon deutscher Partikeln* (3rd ed.). Leipzig: Langenscheidt.

Hentschel, E., & Weydt, H. (1994). *Handbuch der deutschen Grammatik* (2nd ed.). Berlin: de Gruyter.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy, 5* (3), 341-353.

Hopper, P. J., & Thompson, S. A. (1984). The discourse basis for lexical categories in universal grammar. *Language, 60* (4), 703-752.

Hughes, J., & Atwell, E. (1994). The automated evaluation of inferred word classifications. In Tony Cohn (Ed.), *Proceedings of the 11th European Conference on Artificial Intelligence* (ECAI, pp. 535-539). Chichester: Wiley.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford: Oxford University Press.

Jusczyk, P. W. (1997). *The discovery of spoken language.* Cambridge, Mass.: MIT Press.

Jusczyk, P. W. (1998). Dividing and conquering linguistic input. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Proceedings of the 34th Annual Meeting of the Chicago Linguistic Society. Vol. 2: The panels* (pp. 293-310). Chicago: CLS.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3* (9), 323-328.

Jusczyk, P. W., Hirsh-Pasek, K., Kemler Nelson, D. G., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24* (2), 252-293.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99* (2), 349-364.

Kelly, M. H. (1996). The role of phonology in grammatical category assignments. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 249-262). Mahwah, New Jersey: Erlbaum.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *The psychology of learning and motivation: Advances in research and theory, 7,* 1-41.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104* (2), 211-240.

Langacker, R. W. (1987). Nouns and verbs. *Language, 63* (1), 53-94.

Lee, L. (1999). Measures of distributional similarity. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL), 25-32.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22* (1), 1-38.

Lieven, E. V. M. (1997). Variation in a crosslinguistic context. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition. Vol. 5: Expanding the contexts* (pp. 199-263). Mahwah, New Jersey: Erlbaum.

Liu, H. (1997). Lexical access and differential processing in nouns and verbs in a second language. (Doctoral dissertation 1996, University of California, San Diego). *Dissertation Abstracts International, 57* (10), 6606-B.

Lleó, C. (2001). The interface of phonology and morphology: The emergence of the article in the early acquisition of Spanish and German. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 2, pp. 23-44). Amsterdam: Benjamins.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101* (4), 676-703.

Macnamara, J. (1982). *Names for things: A study of human learning.* Cambridge, Mass.: MIT Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Vol. I: Transcription Format and Programs* (3rd ed.). Mahwah, New Jersey: Erlbaum.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, Mass.s: MIT Press.

Maratsos, M. P. (1990). Are actions to verbs as objects to nouns? On the differential semantic bases of form, class, category. *Linguistics, 28*, 1351-1379.

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

McNicol, D. (1972). *A Primer of Signal Detection Theory.* London: Allen and Unwin.

Meringer, R., & Mayer, K. (1895). *Versprechen und Verlesen: Eine psychologisch-linguistische Studie.* Stuttgart, Germany: Göschen.

Miller, M. H. (1976). *Zur Logik der frühkindlichen Sprachentwicklung: Empirische Untersuchungen und Theoriediskussion.* Stuttgart: Klett (An English translation by R. T. King with the title 'The logic of language development in early childhood' was published 1979 by Springer).

Miller, W., & Ervin, S. (1964). The development of grammar in child language. In U. Bellugi & R. Brown (Eds.), *The Acquisition of Language* (Monographs of the Society for Research in Child Development, 29:1).

Mills, A. E. (1985). The acquisition of German. In D. I. Slobin (Ed.), *The cross-linguistic study of language acquisition. Vol. 1: The data* (pp. 141-254). Hillsdale, New Jersey: Erlbaum.

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

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

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96* (2), 143-182.

Monaghan, P., & Christiansen, M. H. (2004). What distributional information is useful and usable for language acquisition? In T. Regier, D. Gentner, & K. Forbus (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 963-968). Mahwah, New Jersey: Erlbaum.

Monaghan, P., Christiansen, M. H., & Chater, N. (2004). *How language-specific is language acquisition? A cross-linguistic analysis of cues for syntactic categorisation* (Poster presented at the 10th Annual Conference on Architectures and Mechanisms of Language Processing (AMLaP 2004), Aix-en-Provence, France). Abstract available at http://www.lpl.univ-aix.fr/~AMLaP2004/booklet.pdf.

Morgan, J. L., & Demuth, K. (Eds.). (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Mahwah, New Jersey: Erlbaum.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 263-283). Mahwah, New Jersey: Erlbaum.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading times in sentence processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* (NIPS 2001, Vol. 14). Cambridge, Mass.: MIT Press.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. *Psychological Bulletin, 84* (1), 93-116.

Newport, E. L. (1977). Motherese: The speech of mothers to young children. In N. J. Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 177-217). Hillsdale, New Jersey: Erlbaum.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14* (1), 11-28.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 109-149). Cambridge: Cambridge University Press.

Nübling, D. (2004). Die prototypische Interjektion: Ein Definitionsvorschlag. *Zeitschrift für Semiotik, 26* (1-2), 11-45.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development, 8* (3), 245-272.

Perani, D., Cappa, S. F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., & Fazio, F. (1999). The neural correlates of verb and noun processing: A PET study. *Brain, 122* (12), 2337-2344.

Peters, A. M. (1997). Language typology, prosody, and the acquisition of grammatical morphemes. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition. Vol. 5: Expanding the contexts* (pp. 135-197). Mahwah, New Jersey: Erlbaum.

Peters, A. M. (2001). Filler syllables: What is their status in emerging grammar? *Journal of Child Language, 28* (1), 229-242.

Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development, 44* (1), 182-185.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7* (3), 217-283.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, Mass.: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In Brian MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, New Jersey: Erlbaum.

Postman, L., & Keppel, G. (Eds.). (1970). *Norms of word association.* New York: Academic Press.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19* (1-2), 9-50.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22* (2), 253-279.

Pulvermüller, F., Lutzenberger, W., & Preissl, H. (1999). Nouns and verbs in the intact brain: Evidence from event-related potentials and high-frequency cortical responses. *Cerebral Cortex, 9* (5), 497-506.

Quasthoff, U. (1998). Deutscher Wortschatz im Internet. *LDV-Forum* (GLDV-Journal for Computational Linguistics and Language Technology), *15* (2), 4-23. Queries were made between March and June, 2003, from http://wortschatz.uni-leipzig.de/.

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

Redington, M., Chater, N., Huang, C., Chang, L.-P., Finch, S., & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Chinese. *Proceedings of the 4th International Conference on the Cognitive Science of Natural Language Processing, Dublin* (CSNLP).

Ross, J. R. (1972). The category squish: Endstation Hauptwort. In P. M. Peranteau, J. N. Levi, & G. C. Phares (Eds.), *Proceedings of the 8th Annual Meeting of the Chicago Linguistic Society* (pp. 316-328). Chicago: CLS.

Sasse, H.-J. (1993). Syntactic Categories and Subcategories. In J. Jacobs, A. Stechow, W. Sternefeld, & T. Vennemann (Eds.), *Syntax: An international handbook of contemporary research* (Vol. 1, pp. 646-686). Berlin: de Gruyter.

Schlesinger, I. M. (1988). The origin of relational categories. In Y. Levy, I. M. Schlesinger, & M. D. S. Braine (Eds.), *Categories and processes in language acquisition* (pp. 121-178). Hillsdale, New Jersey: Erlbaum.

Schlesinger, I. M. (1991). Innate universals do not solve the negative feedback problem. [Open peer commentary of S. Crain's article "Language acquisition in the absence of experience" in BBS 14, 597-612]. *Behavioral and Brain Sciences*, *14* (4), 633.

Scholz, B. C., & Pullum, G. K. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review*, *19* (1-2), 185-223.

Schütze, H. (1995). Distributional part-of-speech tagging. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics* (EACL95), 141-148.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24* (1), 97-123.

Shady, M. E. (1997). Infants' sensitivity to function morphemes (Doctoral dissertation 1996, State University of New York at Buffalo). *Dissertation Abstracts International*, *58* (01), 441-B.

Shafer, V. L., Shucard, D. W., Shucard, J. L., & Gerken LA. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language and Hearing Research*, *41* (4), 874-886.

Shapiro, K. A., Pascual-Leone, A., Mottaghy, F. M., Gangitano, M., & Caramazza, A. (2001). Grammatical distinctions in the left frontal cortex. *Journal of Cognitive Neuroscience, 13* (6), 713-720.f

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language, 25* (1), 169-201.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72* (2), B11-B21.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology, 72* (4), 580-588.

Smith, K. H. (1969). Learning co-occurrence restrictions: Rule learning or rote learning? *Journal of Verbal Learning and Verbal Behavior, 8* (2), 319-321.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development, 43* (2), 549-565.

Sokolov, J. L., & Snow, C. E. (1994). The changing role of negative evidence in theories of language development. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 38-55). Cambridge: Cambridge University Press.

Stern, C., & Stern, W. (1965). *Die Kindersprache: Eine psychologische und sprachtheoretische Untersuchung.* Darmstadt: Wissenschaftliche Buchgesellschaft (Original work published 1928, 4th ed., Leipzig: Barth).

Taylor, J. R. (1989). *Linguistic categorization: Prototypes in linguistic theory*. Oxford: Clarendon.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2002). Going, going, gone: The acquisition of the verb 'go'. *Journal of Child Language, 29* (4), 783-811.

Thomas, M. (2002). Development of the concept of "the poverty of the stimulus". *The Linguistic Review, 19* (1-2), 51-71.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development.* Cambridge: Cambridge University Press.

Tomasello, M. (1995). Language is not an instinct [Review of the book *The language instinct: How the mind creates language*]. *Cognitive Development, 10* (1), 131-156.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74* (3), 209-253.

Tomasello, M. (2003a). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, Mass.: Harvard University Press.

Tomasello, M. (2003b). Introduction: Some surprises for psychologists. In M. Tomasello (Ed.), *The new psychology of language: Cognitive and functional approaches to language structure* (Vol. 2, pp. 1-14). Mahwah, New Jersey: Erlbaum.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development, 8* (4), 451-464.

Tracy, R. (1990). Spracherwerb trotz Input. In M. Rothweiler (Ed.), *Spracherwerb und Grammatik: Linguistische Untersuchungen zum Erwerb von Syntax und Morphologie* (pp. 22-49). Opladen: Westdeutscher Verlag.

Trask, R. L. (1999). Parts of speech. In K. Brown & J. Miller (Eds.), *Concise encyclopedia of grammatical categories* (pp. 278-284). Oxford: Elsevier.

Trueswell, J. C., & Tanenhaus, M. K. (1994). Towards a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 155-179). Hillsdale, New Jersey: Erlbaum.

Vigliocco, G. (2001). Tip-of-the-tongue, psychology of. In *International encyclopedia of the social and behavioral sciences* (Vol. 23, pp. 15759-15762). Oxford: Elsevier.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science, 8*, 314-317.

Vinson, D. P., & Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: Semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics, 15* (3-5), 317-351.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition.* Cambridge, Mass.: MIT Press.

Wright, B., & Garrett, M. (1984). Lexical decision in sentences: Effects of syntactic structure. *Memory and Cognition, 12* (1), 31-45.

Zavrel, J. (1996). *Lexical Space: Learning and using continuous linguistic representations.* Unpublished doctoral dissertation, Department of Philosophy, Utrecht University, The Netherlands.

Appendices

Appendix A Target words by benchmark category

Below, each of the 11 benchmark categories is specified extensionally, i.e., by listing the target words that were assigned to it by the procedures documented in section 2.2. Category members are sorted by their frequency in the corpus, with the more frequently occurring words listed first. Note that, deviating from standard orthography, some target words carry the plus sign + or the apostrophe ‘. Adhering to CHAT format, the transcribers used the plus sign to link fixed expressions (such as *oh+gott* and *was+für+ein*; English: *oh god* and *what* $a_{\text{masc.sg.nom./neut.sg.:nom.+acc.}}$, respectively) to which they attributed a lexical status similar to that of compounds (such as *flughafen* and *tonbandaufnahme*; English: *airport* and *tape recording*, respectively). The apostrophe was used to mark contracted forms (such as *‘nen* as a contracted variant of *einen*; English: $a_{\text{masc.sg.acc.}}$).

Interjections [INTJ; 77 members]:

ja, hm, oh, na, genau, nee, nein, okay, ach, ah, aha, ach+so, bitte, naja, huch, halt, hey, moment, aeh, na+gut, och, he, hallo, upps, oje, ha, vorsichtig, aua, uah, huh, na+sowas, hmhm, noe, boah, eh, vorsicht, jawohl, oh+ja, brumm, schwupp, hilfe, oh+mann, aehm, augenblick, brr, tsch, uh, ui, oh+gott, tja, bitteschoen, wunderbar, mensch, guten+tag, wow, uppsa, bumm, puh, hach, ah+ja, danke, upp, huhu, au, hupps, kuckuck, bumms, hui, ho, gott, hopps, mm, oi, blupp, hoi, entschuldige, nja

Verbs [V; 288 members]:

ist, guck, hat, hast, haben, machen, sind, kann, muss, war, muessen, soll, macht, glaub, hab, sollen, komm, geht, musst, kommt, koennen, bist, weiss, gucken, gemacht, weisst, faehrt, wird, mach, wollen, fahren, sieht, kannst, machst, essen, willst, gibt, sehen, meinst, glaube, sein, warte, stimmt, will, bauen, bin, werden, lass, gehen, pass, malen, waren, nehmen, gesehen, heisst, sag, sagen, steht, passt, sagt, brauchen, darf, spielen, hol, zeig, wart, lassen, siehst, gegessen, kommen, holen, trinken, koennte, magst, passiert, gehoert, habt, kriegt, isst, isses, tun, gesagt, gefahren, schau, tut, weisste, stehen, schauen, kannste, moechtest, find, waer, sitzt, angucken, seh, haste, kriegen, nimm, moechte, fliegen, schmeckt, ziehen, liegt, zieh, dachte, waere, fehlt, anziehen,

reparieren, geh, koennten, gedacht, braucht, kriegst, gib, habe, hatte, siehste, probier, darfst, finden, denk, sitzen, warst, gebaut, liegen, tu, findest, wollte, musste, erzaehl, willste, ueberlegen, wuerde, isser, mag, gefunden, gekriegt, faellt, schlafen, bleiben, setz, guckt, wolltest, probieren, warten, leg, brauchst, legen, kaufen, geben, suchen, fliegt, haelt, halten, gehabt, aufpassen, klappt, laufen, hatten, wollt, einkaufen, stell, helfen, setzen, sagst, bau, brauch, aufmachen, gab, wissen, malst, aussteigen, vorlesen, erzaehlt, laeuft, stellen, gefallen, hoer, kommst, nimmst, geworden, bleibt, unterbrechen, drehen, baust, mitnehmen, bring, gemalt, haette, wuerd, haettest, komme, aufbauen, krieg, finde, vergessen, druecken, sollte, wuerdest, nehm, werd, gibst, versuch, heissen, koenntest, bringt, laesst, hattest, einsteigen, umdrehen, angeln, bringen, reicht, mitgebracht, reintun, passen, angeguckt, hoert, schnaeuzen, hoeren, aussieht, fragen, schlaeft, kennst, runtergefallen, wirst, findet, gewesen, denke, erinnern, erzaehlen, pusten, versuchen, gekauft, gespielt, ging, nimmt, mitfahren, verloren, packen, repariert, tust, kleben, gekommen, schwimmen, anmalen, ausziehen, vorstellen, zeigen, funktioniert, wollten, gehst, baden, genommen, wuerfeln, gefangen, sei, such, schneiden, versteckt, durchfahren, sollten, setzt, ausschneiden, meinste, versteh, wohnt, hilft, bekommen, putzen, scheint, umgefallen, dreh, fahr, fangen, schieben, haengen, dranmachen, kam, abmachen, hinsetzen, kochen, seid, gefaellt, gucke, koennt, wartet

Nouns [N; 268 members]:

leo, wilhelmine, zug, auto, papa, mama, wasser, tunnel, flugzeug, i+c+e, eichi, bus, ernie, fisch, leute, maus, recht, seite, hand, eule, buch, stueck, kuh, eisenbahn, schiff, haus, sachen, tuer, bett, tiere, nase, bahnhof, mund, farbe, karte, kindergarten, bobo, zoo, autos, fische, beispiel, oma, idee, teil, bruecke, bild, mechthild, ente, kopf, billi, pipi, kinder, osterhase, zeit, tisch, hose, kuchen, katze, karten, s+bahn, huhn, schornstein, knete, schweinchen, krokodil, augen, raeder, meer, platz, hunger, schienen, fenster, boden, finger, junge, ball, hubschrauber, schaf, spiel, dach, lok, vivien, flughafen, frosch, hase, klo, mond, baum, glueck, eichhoernchen, maeuschen, trambahn, apfel, bauernhof, gans, elefant, hund, bagger, giraffe, tag, stuhl, teile, mann, wuerstchen, leipzig, ordnung, leuchtturm, nudeln, sonne, berlin, gertrud, haende, aufnahme, tier, brot, schwein, paris, schokolade, zimmer, bauch, uhr, stuttgart, rauch, tueren, eis, muscheln, turm, stadt, schere, fuss, schnecke, papier, schranke, strasse, lego, lokomotive, abend, laster, gondel, teppich, licht, tiger, gegend, kueche, gleise, loch, v+w+kaefer, enten, swantje, ei, schatz, kiste, tonbandaufnahme, geburtstag, weihnachten, baer, boot, kind, socken, deckel, zuege, urlaub, china, loeffel, stift, geschichte, opa, kaertchen, teller, bauer, laerm, menge, see, decke, auge, fahrrad, ding, minuten, zaehne, runde, kartoffeln, lust, loewe, fuesse, steine, u+bahn, kleber, pizza, eier, wagen, buttermilch, kaese, pferd, schaffner, berg, milch, aepfel, problem, bilder, farben, pilot, schuhe, gesicht, fruehstueck, dinge, haare, angst, feuerwehr, haenger, arzt, angel, hause, nacht, woche, mittagessen, rakete, blatt, blumen, aal, beine, mittag, fluegel, hauptbahnhof, mist, rad, becher, gleis, schluss, schluessel, pflaster, ahnung, glas, kaffee, lokfuehrer, puzzle, bein, kirche, schnee, arm, haeschen, ohren, polizei+auto, l+k+w, stern, baeren, berni, eimer, hausschuhe, nuesse, schrank, strassenbahn, richtung, spass, traudel, anhaenger, ferkel, luft, paar, berge, ostsee, spielplatz, blume, butter, froesche, peter, tee, weiche, leo+hartwig

Adjectives [ADJ; 96 members]:

gut, kleine, kleinen, toll, super, grosse, rot, fertig, prima, gruen, kleiner, kaputt, grossen, ehrlich, gross, kleines, gelb, blau, lang, schlecht, ganzen, gute, weit, ganze, lustig,

muede, kalt, wahr, grosses, grosser, rote, fein, lila, klein, klar, heiss, schwarz, langsam, voll, schoene, nass, schwierig, neue, lecker, schlimm, naechste, roten, gruene, klasse, warm, nett, tolle, gruenen, schwer, gelbe, alte, alt, sauer, blaue, laut, tolles, orange, krank, schoenes, letzte, schick, arme, neuen, gelben, traurig, wach, dunkel, alter, letzten, neues, rosa, weisse, neu, naechsten, schoenen, komisch, geschickt, langen, unglaublich, leer, schrecklich, raffiniert, schoener, dreckig, toller, interessant, blauen, doof, gesund, huebsch, groesser

Adverbs [ADV; 94 members]:

da, so, noch, hier, jetzt, wieder, schon, gerade, heute, gleich, drauf, nochmal, immer, nur, drin, rein, hin, her, dran, erstmal, kurz, schnell, oben, gerne, richtig, naemlich, zusammen, weiter, nachher, hinten, unten, runter, erst, morgen, dazu, vorne, lieber, gestern, lange, rum, dabei, irgendwie, irgendwo, genug, nun, kraeftig, ordentlich, dafuer, selber, draussen, manchmal, zurueck, nie, alleine, vorhin, sofort, fast, davon, vorbei, hinein, drunter, bald, auseinander, davor, rauf, dahin, vorwaerts, trotzdem, anders, dort, fuerchterlich, ueberall, hierhin, dahinten, drueber, danach, herum, ran, daraus, doll, irgendwann, vor+allem, dazwischen, uebrigens, zuhause, obendrauf, zwar, daneben, etwa, frueher, direkt, unterwegs, zwischen, darauf

Interrogative words [INTG; 17 members]:

was, wo, wie, wer, warum, welche, wieso, wohin, wann, wen, wem, welches, was+fuer+eine, was+fuer+ein, wieviel, woher, welchen

Pronouns [PRON; 35 members]:

das, du, ich, wir, ‘s, es, sie, man, dir, er, mir, sich, alles, dich, alle, ihr, mich, uns, nichts, ihn, sowas, eines, ihm, eins, beiden, irgendwas, jemand, beide, euch, nix, mer, denen, de, ihnen, dies

Determiners [DET; 61 members]:

die, der, ein, den, eine, einen, dem, im, ‘ne, zum, ‘m, zwei, keine, deine, viel, am, andere, einem, viele, kein, dein, anderen, mein, einer, vom, ein+paar, zur, des, beim, deinen, diese, drei, meine, deinem, keinen, unser, anderes, deiner, unsere, dieses, vier, dieser, meinen, fuenf, seine, ‘nen, sechs, diesen, diesem, meinem, vielen, seinen, ihre, ‘ner, acht, lauter, seinem, zehn, meiner, ihren, solche

Prepositions [PREP; 15 members]:

mit, auf, in, fuer, von, nach, bei, durch, vor, ueber, unter, bis, ohne, hinter, gegen

Conjunctions [CONJ; 13 members]:

und, dann, oder, wenn, dass, weil, ob, damit, als, um, sonst, deshalb, sondern

Particles [PTCL; 53 members]:

mal, nicht, auch, ‘n, denn, ganz, aber, doch, zu, aus, ein+bisschen, an, vielleicht, mehr, ne, gar, schoen, eigentlich, hae, sehr, echt, weg, ziemlich, raus, also, einfach, einmal, wirklich, gell, ab, natuerlich, besser, hoch, los, wohl, ueberhaupt, bestimmt, etwas, bisschen, eher, tatsaechlich, bloss, sogar, fest, weh, wahrscheinlich, sicher, rueber, eben, leider, voellig, hinterher, ruhig

Appendix B Deriving the expectation of Average Precision

Let Γ denote a category comprising C out of a total of L target words. Let t be a particular member of Γ. If the $L-1$ other target words are ranked randomly (using a uniform probability distribution), the Average Precision AP_t of the resulting rank list has the expected value

$$\mu_{AP_t} = \frac{1}{(C-1)\binom{L-1}{C-1}} \sum_{i=1}^{C-1} \sum_{r=i}^{L-C+i} \binom{r-1}{i-1} \binom{L-1-r}{C-1-i} \frac{i}{r} \quad .$$

Proof

To recap the relevant symbols introduced in subsection 3.3.3, let

$$\text{rank}_t(1) < \text{rank}_t(2) < \ldots < \text{rank}_t(C-1)$$

denote the particular ranks occupied by the $C-1$ members of Γ (other than t). The rank list's Average Precision AP_t is then given by

$$AP_t = \frac{1}{C-1} \sum_{i=1}^{C-1} P_t(i) \ , \quad \text{where} \quad P_t(i) = \frac{i}{\text{rank}_t(i)} \ .$$

Because of

$$\mu_{AP_t} = \frac{1}{C-1} \sum_{i=1}^{C-1} \mu_{P_t(i)} \ ,$$

it suffices to show that for the i-th member in the rank list, the expected value of $P_t(i)$ is

$$\mu_{P_t(i)} = \frac{1}{\binom{L-1}{C-1}} \sum_{r=i}^{L-C+i} \binom{r-1}{i-1} \binom{L-1-r}{C-1-i} \frac{i}{r} \ . \tag{19}$$

To prove this equation, let Δ denote the set of all possible rank lists $\text{rank}^*(\cdot)$ with $1 \le \text{rank}^*(1) < \text{rank}^*(2) < \ldots < \text{rank}^*(C-1) \le L-1$. Let further $D = |\Delta|$ denote the number of possible rank lists. Assuming a uniform probability distribution across Δ — i.e., assuming that each possible rank list is equally likely to occur by chance —, the expected value of $P_t(i)$ is computed as

$$\mu_{P_t(i)} = \frac{1}{D} \sum_{\text{rank}^* \in \Delta} \frac{i}{\text{rank}^*(i)} \ . \tag{20}$$

From this, equation (19) can be derived by exploiting the well-known fact that there are

$$\binom{N}{K} = \frac{N!}{K!\,(N-K)!}$$

different ways of assigning K (undistinguished) objects to N different slots such that no two objects occupy the same slot (resulting in an *ordered partition* of these slots). Since building a rank list for target word t corresponds to assigning $K = C-1$ members (objects) to $N = L-1$ ranks (slots), there are a total of

$$D = \binom{L-1}{C-1} \tag{21}$$

possible rank lists. Let $D_{i,r}$ denote the number of possible rank lists in which the i-th member occupies rank r. Within this subset of rank lists, there are $i-1$ other members to be assigned to the $r-1$ higher ranks (i.e., above member i on rank r) and $C-1-i$ other members to be independently assigned to the $L-1-r$ lower ranks (i.e., below member i). Applying the object–slot metaphor twice, one obtains

$$D_{i,r} = \binom{r-1}{i-1}\binom{L-1-r}{C-1-i} \ . \tag{22}$$

For each of these rank lists, the Precision at the i-th member is

$$P_t(i) = \frac{i}{r} \ .$$

Because of the fixed number of other members being ranked above and below it ($i-1$ and $C-1-i$, respectively), the i-th member can only occupy the ranks $i, i+1, i+2, \ldots, L-\mathrm{C}+i$. Therefore, equation (20) becomes

$$\mu_{P_t(i)} = \frac{1}{D} \sum_{r=i}^{L-C+i} D_{i,r} \frac{i}{r} \ . \tag{23}$$

Together with equations (21) and (22), this yields equation (19) and proves the assertion.

Appendix C Interpreting L_1 distance ranges

The L_1 distance between any two SCO vectors lies within the interval from 0.0 (identical vectors) to 2.0 (orthogonal vectors). When the default settings of the co-occurrence method (as specified in section 3.1) are applied to the *Leo* corpus, the L_1 distances actually observed range from .42 to 1.97. In order to investigate the robustness of distributional information (cf. section 4.2), it is useful to get a linguistic sense of how similar any two target words with a particular L_1 distance actually are. To this end, a large number of target word pairs of various distance levels were inspected, and this pilot work suggests the following interpretations:

- $\boldsymbol{L_1 < .6}$: Any two target words whose SCO vectors have an L_1 distance of less than .6 can be regarded as being as similar as synonyms. They thus tend to be *fully replaceable* in most contexts, preserving the syntactic structure and essentially without altering the overall semantic content.
- $\boldsymbol{L_1 < .8}$: Any two target words whose SCO vectors have an L_1 distance of less than .8 can be regarded as being similar like semantically related words of the same lexical category and with the same grammatical features (such as inflection). They thus tend to be *well replaceable* in most contexts, preserving the syntactic structure and with only minor changes to the overall semantic content.
- $\boldsymbol{L_1 < 1.0}$: Any two target words whose SCO vectors have an L_1 distance of less than 1.0 can be regarded as being similar like words of the same lexical category which typically have the same grammatical features but need not be semantically related. They thus tend to be *syntactically replaceable* in most contexts, preserving the syntactic structure.

Being based on semi-formal pilot work, the choice of these particular distance ranges is essentially subjective, though not arbitrary. The specific observations that motivated them are summarized below.

- $\boldsymbol{L_1 < .6}$: First, of the 516,636 possible pairs of target words, there are only 118 (0.02%) that have an L_1 distance below .6. Second, for each of these 118 target word pairs, both words belong to the same category. While in the vast majority of cases (101 word pairs), this concerns the interjection category, four categories (PRON, PREP, CONJ, PTCL) do not contain a single word pair closer than .6.

Third, where applicable, any two words with L_1 distance below .6 have identical grammatical features such as gender, number, and person. Fourth, for almost all of the 118 target word pairs with L_1 distance below .6, both words are semantically very closely related, and almost exclusively by one of the following relations: They are synonyms, or one is the hyponym of the other, or both are different forms of the same lexeme.

- ***L*₁ < .8**: First, of the 516,636 possible pairs of target words, only 1,030 (0.2%) have an L_1 distance below .8. Second, for virtually all (1,008) of these word pairs, both words belong to the same category. Third, where applicable, words with L_1 distance below .8 tend to have identical grammatical features. Fourth, for most target word pairs with L_1 distance below .8, both words are semantically closely related, and generally by one of the following relations: They are synonyms, or one is the hyponym of the other, or both are different forms of the same lexeme, or both belong to the same semantic field (such as colors, spatial relations, and means of transportation).
- ***L*₁ < 1.0**: First, of the 516,636 possible pairs of target words, only 4,629 (0.9%) have an L_1 distance below 1.0. Second, for the vast majority (4,283) of these word pairs, both words belong to the same category. Third, where applicable, words with L_1 distance below 1.0 tend to have identical grammatical features. Fourth, for many target word pairs with L_1 distance below 1.0, both words are semantically related by one of the relations specified in the preceding paragraph, while for others, there is no obvious semantic relation. Fifth, even for the most compact category (INTJ), both the mean and median of all L_1 distances between members of this category are slightly above 1.0 (and clearly above 1.0 for all other categories). This is relevant here because from a purely syntactic point of view, almost all interjections are replaceable for each other since they generally do not interact syntactically with other word tokens in an utterance (cf. 2.2.1).

It should be emphasized that these L_1 ranges are by no means a *natural law* or an inherent property of the Manhattan metric L_1. They only arise for the combination of a particular co-occurrence model and a particular corpus (here: the *Leo* corpus and the default settings for the co-occurrence model). Presumably, the most influential factors in this dependency are (i) the size and composition of the context lexicon, and (ii) the base frequencies of target words and context words. For instance, everything else being

equal, raising the base frequency of all target words tends to diminish the L_1 ranges. By contrast, raising the context lexicon size typically causes the L_1 ranges to increase.

In consequence, the L_1 ranges obtained above cannot readily be extrapolated to other corpora or other model settings. Yet, such extrapolation is precisely the purpose of these ranges: In section 4.2, they are used to study various aspects of robustness by means of certain manipulations of both data and model. The solution is to study each of these manipulations in isolation, and to address for each manipulation the issue of how it influences the L_1 ranges. But rough estimates will do since nothing crucial hinges on these L_1 ranges. They merely serve to provide an intuition as to whether two particular SCO vectors with, say, an L_1 distance of .55, are surprising to be as similar to each other as was otherwise observed for synonyms.

Appendix D Individual preferences and discriminators

On the following pages, the most important distributional properties and discriminators the 11 benchmark categories are listed at the level of individual context words. There are three tables for each category. The first one presents all distributional properties of the given category that have a preference value of at least 2% (sorted by descending preference). The second table lists all positive discriminators for the given category that have a relative preference value of +1.5% or greater (sorted by descending relative preference), together with their discriminative power. The third table does the same for all negative discriminators with a relative preference value of −1.5% or less.

Preference values and relative preference values are given as percentages. Context words are always listed together with their category (according to the benchmark classification) to facilitate relating these tables to Table 4-4, Table 4-6, and Table 4-7 in section 4.3. The four utterance boundary markers (as virtual context words) are represented by the same symbols that were introduced earlier (cf. p. 60): the symbol _<_ as the pre-utterance marker, and the symbols _._, _?_, and _!_ as the three post-utterance markers (matching the three possibilities of utterance-terminal punctuation). Like in section 4.3, the symbol <Bnd> is used to provide these four virtual context words with a category specification.

Interjections

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-1	_<_	<Bnd>	73.9
+1	_._	<Bnd>	43.8
-2	_<_	<Bnd>	40.5
+2	_._	<Bnd>	19.0
+2	ist	V	7.1
+1	_!_	<Bnd>	4.8
+1	leo	N	3.9
+1	das	PRON	3.7
+1	_?_	<Bnd>	3.0
+2	das	PRON	2.9
+2	mal	PTCL	2.5
+2	du	PRON	2.4
-2	der	DET	2.2
+2	_!_	<Bnd>	2.1

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	_<_	<Bnd>	63.2	+.414
+1	_._	<Bnd>	28.8	+.059
-2	_<_	<Bnd>	22.7	+.080
+2	ist	V	6.0	+.011
+1	_!_	<Bnd>	4.2	-.007
+2	_._	<Bnd>	2.2	-.001
+1	leo	N	2.2	-.012
+1	das	PRON	2.0	-.001

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+2	_?_	<Bnd>	-9.6	+.078
+1	_?_	<Bnd>	-7.6	+.055
-1	die	DET	-5.2	+.050
-1	das	PRON	-3.6	+.034
-1	der	DET	-3.2	+.030
-1	ich	PRON	-2.2	+.021
+1	mal	PTCL	-1.9	+.018
-1	du	PRON	-1.7	+.015
-1	da	ADV	-1.6	+.015
-2	ist	V	-1.6	+.012
-1	ein	DET	-1.6	+.015
-1	was	INTG	-1.5	+.015

Verbs

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-2	_<_	<Bnd>	28.8
-1	_<_	<Bnd>	13.5
+1	_._	<Bnd>	13.0
+2	_._	<Bnd>	12.2
+1	_?_	<Bnd>	10.1
+2	_?_	<Bnd>	8.4
+1	du	PRON	7.4
-1	ich	PRON	6.2
-1	das	PRON	5.4
-2	und	CONJ	4.8
+1	ich	PRON	4.7
+1	mal	PTCL	4.6
+1	wir	PRON	4.3
-1	was	INTG	4.2
-1	dann	CONJ	4.0
+2	mal	PTCL	3.7
-1	da	ADV	3.6

Context position	Context word	Benchmark category	Preference (in %)
<continued from left>			
-1	du	PRON	3.4
+1	die	DET	3.3
-2	die	DET	3.1
+1	das	PRON	3.0
+2	das	PRON	2.9
-1	die	DET	2.9
+2	die	DET	2.7
+2	nicht	PTCL	2.5
-2	ja	INTJ	2.5
-1	jetzt	ADV	2.4
+2	auch	PTCL	2.3
-1	mal	PTCL	2.3
+2	noch	ADV	2.2
-1	wir	PRON	2.1
+1	leo	N	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-2	_<_	<Bnd>	12.9	+.010
+1	du	PRON	6.7	-.053
-1	ich	PRON	5.5	-.041
+1	ich	PRON	4.2	-.033
+1	wir	PRON	4.1	-.031
-1	dann	CONJ	3.8	-.012
-1	was	INTG	3.7	-.016
+1	mal	PTCL	3.6	-.030
-2	und	CONJ	3.1	-.004
-1	da	ADV	2.7	-.005
+2	mal	PTCL	2.6	-.016
-1	das	PRON	2.6	-.007
+1	die	DET	2.0	-.004
-1	du	PRON	2.0	-.017
-1	jetzt	ADV	2.0	-.006
+1	das	PRON	1.7	-.007
+1	's	PRON	1.7	-.012
+2	nicht	PTCL	1.7	-.007
+2	auch	PTCL	1.5	-.004

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+2	_._	<Bnd>	-6.6	+.010
+1	_._	<Bnd>	-5.9	+.026
-2	ist	V	-3.5	+.031
+2	_?_	<Bnd>	-3.2	+.003
-1	die	DET	-2.9	+.034
-1	_<_	<Bnd>	-2.7	+.033
-1	der	DET	-2.6	+.023
-1	ein	DET	-2.4	+.024
+1	ist	V	-1.9	+.017
-1	'n	PTCL	-1.9	+.019
-1	ist	V	-1.7	+.016

Nouns

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
+1	_._	<Bnd>	18.4
+2	_._	<Bnd>	18.4
+2	_?_	<Bnd>	15.7
+1	_?_	<Bnd>	15.5
-1	die	DET	12.5
-2	_<_	<Bnd>	12.1
-1	der	DET	8.9
-1	_<_	<Bnd>	6.2
-1	das	PRON	5.0
-1	ein	DET	4.6
-2	auf	PREP	4.0
-1	'n	PTCL	3.8
-2	ist	V	3.8
-2	noch	ADV	3.4
	<continued from left>		
-2	in	PREP	3.3
-1	den	DET	3.3
-2	mit	PREP	3.0
+1	ist	V	2.8
+1	und	CONJ	2.7
-1	dem	DET	2.6
-2	mal	PTCL	2.4
+2	leo	N	2.2
-2	die	DET	2.1
-1	eine	DET	2.1
-2	du	PRON	2.0
-2	und	CONJ	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	die	DET	10.2	-.047
-1	der	DET	7.7	-.035
+1	_?_	<Bnd>	7.4	+.049
+2	_?_	<Bnd>	6.9	+.027
-1	ein	DET	4.0	-.018
-2	auf	PREP	3.7	-.026
-1	'n	PTCL	3.3	-.014
-2	in	PREP	3.0	-.021
-1	den	DET	2.7	-.016
-1	dem	DET	2.4	-.010
-2	mit	PREP	2.4	-.007
-2	noch	ADV	2.1	-.004
+2	_._	<Bnd>	1.9	+.025
-1	das	PRON	1.9	-.018
-1	eine	DET	1.7	-.013
+1	_._	<Bnd>	1.7	+.062
-1	zum	DET	1.7	-.015
-1	im	DET	1.7	-.014
+1	ist	V	1.6	+.003
+1	und	CONJ	1.5	.000
-1	'm	DET	1.5	-.011

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	_<_	<Bnd>	-12.6	+.105
-2	_<_	<Bnd>	-10.0	+.083
-1	ich	PRON	-3.0	+.029
+1	du	PRON	-3.0	+.029
+1	ich	PRON	-2.0	+.019
-1	da	ADV	-2.0	+.018
-1	was	INTG	-1.9	+.019
+1	mal	PTCL	-1.9	+.019
-1	du	PRON	-1.9	+.016
+1	wir	PRON	-1.8	+.018
-1	mal	PTCL	-1.6	+.014
-1	dann	CONJ	-1.6	+.016

Adjectives

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
+2	_._	<Bnd>	26.3
+1	_._	<Bnd>	23.5
+2	_?_	<Bnd>	12.6
-2	_<_	<Bnd>	12.5
-2	ist	V	10.5
+1	_?_	<Bnd>	9.3
-1	_<_	<Bnd>	8.6
-1	ganz	PTCL	6.4
-2	das	PRON	4.9
-1	ist	V	4.9
-1	so	ADV	4.5
-1	ein	DET	4.4
-1	die	DET	4.3
-1	das	PRON	3.7
-1	der	DET	3.2
-2	ja	INTJ	3.1
-1	ja	INTJ	2.9
<continued from left>			
-2	nicht	PTCL	2.7
-1	'n	PTCL	2.7
-1	nicht	PTCL	2.6
-2	noch	ADV	2.5
+1	aus	PTCL	2.4
+1	ist	V	2.4
-1	eine	DET	2.3
+2	leo	N	2.3
+2	du	PRON	2.3
+1	und	CONJ	2.2
-2	die	DET	2.2
+1	leo	N	2.1
-2	auch	PTCL	2.1
+2	die	DET	2.1

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+2	_._	<Bnd>	10.3	+.042
-2	ist	V	8.3	+.009
+1	_._	<Bnd>	7.0	-.003
-1	ganz	PTCL	6.0	-.003
-1	so	ADV	3.7	-.019
-1	ist	V	3.6	-.015
-1	ein	DET	2.9	-.025
-2	das	PRON	2.6	-.016
+1	aus	PTCL	2.1	-.017
+2	_?_	<Bnd>	2.1	+.014
-1	ja	INTJ	1.7	-.012
-1	ziemlich	PTCL	1.7	-.008
-1	sehr	PTCL	1.6	-.010
-2	nicht	PTCL	1.6	-.011
-1	eine	DET	1.6	-.013

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-2	_<_	<Bnd>	-7.7	+.061
-1	_<_	<Bnd>	-7.6	+.059
+1	du	PRON	-2.2	+.022
-1	ich	PRON	-2.0	+.020
+1	ich	PRON	-1.6	+.015
-1	da	ADV	-1.5	+.014

Adverbs

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
+2	_._	<Bnd>	19.4
+1	_._	<Bnd>	17.6
+2	_?_	<Bnd>	11.1
+1	_?_	<Bnd>	9.5
-2	_<_	<Bnd>	9.3
-1	_<_	<Bnd>	7.7
-2	die	DET	4.2
-1	mal	PTCL	4.2
-1	da	ADV	4.0
-2	ich	PRON	3.5
-2	das	PRON	3.4
-2	du	PRON	3.2
-2	wir	PRON	3.1
-1	du	PRON	3.0
+2	die	DET	2.8
-1	hier	ADV	2.6
-1	wir	PRON	2.6
-1	ja	INTJ	2.6
-1	noch	ADV	2.6
-1	und	CONJ	2.5
-2	ist	V	2.4
-1	nicht	PTCL	2.4
+1	noch	ADV	2.3
-2	der	DET	2.3
+1	ist	V	2.3
-1	ist	V	2.2
-1	auch	PTCL	2.2
-1	so	ADV	2.2
+2	leo	N	2.1

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	mal	PTCL	2.8	-.011
+2	_._	<Bnd>	2.7	+.006
-1	da	ADV	2.6	-.018
-1	hier	ADV	2.1	-.010
-2	ich	PRON	2.0	-.006
-2	die	DET	1.7	+.008
-2	wir	PRON	1.6	-.002
+1	noch	ADV	1.6	-.008

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-2	_<_	<Bnd>	-11.3	+.077
-1	_<_	<Bnd>	-8.6	+.071
-1	die	DET	-4.1	+.042
-1	der	DET	-3.0	+.030
+1	du	PRON	-2.3	+.023
-1	das	PRON	-2.2	+.025
-1	ein	DET	-1.8	+.018
+1	ich	PRON	-1.6	+.016

Interrogative words

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-1	_<_	<Bnd>	51.0
-2	_<_	<Bnd>	48.6
-1	und	CONJ	11.7
+2	_?_	<Bnd>	11.6
+2	du	PRON	11.5
+1	ist	V	7.9
+1	_?_	<Bnd>	7.7
+2	die	DET	6.3
+2	das	PRON	6.2
+1	farbe	N	5.5
+2	der	DET	5.4
+2	wir	PRON	5.3
+2	denn	PTCL	5.0
<continued from left>			
+2	'n	PTCL	3.8
-2	und	CONJ	3.4
-2	mal	PTCL	3.4
+1	hast	V	3.2
+2	_._	<Bnd>	2.9
+1	haben	V	2.8
+2	hat	V	2.6
+1	hat	V	2.5
-1	mit	PREP	2.2
-1	mal	PTCL	2.0
+1	kommt	V	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	_<_	<Bnd>	36.1	+.247
-2	_<_	<Bnd>	29.6	+.159
-1	und	CONJ	10.1	+.025
+2	du	PRON	9.8	-.002
+1	ist	V	6.3	-.025
+1	farbe	N	5.5	-.052
+2	denn	PTCL	4.6	+.008
+2	wir	PRON	4.3	-.014
+2	das	PRON	4.1	-.002
+2	die	DET	4.1	-.004
+2	der	DET	3.9	-.010
+2	'n	PTCL	3.3	-.002
+1	hast	V	3.0	-.009
+1	haben	V	2.4	-.017
+2	hat	V	2.3	-.019
+1	hat	V	2.0	-.007
+1	kommt	V	1.9	-.014
-2	weisst	V	1.8	-.005
+1	faehrt	V	1.6	-.013
+1	soll	V	1.6	-.003
-2	guck	V	1.6	-.008
+1	sind	V	1.5	-.002

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	_._	<Bnd>	-16.4	+.151
+2	_._	<Bnd>	-14.3	+.112
-1	die	DET	-5.0	+.049
-1	das	PRON	-3.5	+.035
-1	der	DET	-3.2	+.032
-2	ist	V	-2.7	+.025
+1	_?_	<Bnd>	-2.3	+.022
-2	die	DET	-2.1	+.017
-1	ich	PRON	-2.0	+.020
+1	mal	PTCL	-1.9	+.019
+1	leo	N	-1.8	+.016
+2	mal	PTCL	-1.7	+.016

Pronouns

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-2	_<_	<Bnd>	16.5
+2	_._	<Bnd>	12.4
-1	_<_	<Bnd>	8.5
+2	_?_	<Bnd>	8.1
+1	_._	<Bnd>	7.0
+1	_?_	<Bnd>	4.5
+1	mal	PTCL	4.1
-1	du	PRON	4.0
-2	ich	PRON	3.8
-2	da	ADV	3.6
-2	das	PRON	3.5
-1	ich	PRON	3.4
-1	wir	PRON	3.4
-2	was	INTG	3.2
-2	dann	CONJ	3.2
-2	du	PRON	3.1
+1	nicht	PTCL	2.9
-1	die	DET	2.8

Context position	Context word	Benchmark category	Preference (in %)
	<continued from left>		
-1	ist	V	2.7
-2	die	DET	2.5
+1	da	ADV	2.4
+1	das	PRON	2.4
-2	wir	PRON	2.4
+2	mal	PTCL	2.4
-1	mit	PREP	2.3
-2	jetzt	ADV	2.3
+1	ist	V	2.2
-1	noch	ADV	2.1
-2	und	CONJ	2.1
+2	die	DET	2.1
+1	auch	PTCL	2.0
+2	nicht	PTCL	2.0
+1	noch	ADV	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-2	dann	CONJ	2.9	-.004
-2	da	ADV	2.7	-.008
-2	was	INTG	2.4	-.009
-2	ich	PRON	2.3	-.007
+1	mal	PTCL	2.2	.000
-1	wir	PRON	2.2	-.012
-1	du	PRON	2.1	-.005
+1	nicht	PTCL	1.9	+.007
-2	jetzt	ADV	1.6	-.003
-1	hat	V	1.6	-.008
-1	wenn	CONJ	1.6	-.009
-1	bei	PREP	1.6	-.011

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	_._	<Bnd>	-10.6	+.075
-1	_<_	<Bnd>	-7.2	+.050
+1	_?_	<Bnd>	-5.7	+.046
+2	_._	<Bnd>	-4.7	+.029
-2	_<_	<Bnd>	-3.2	+.041
-1	das	PRON	-3.1	+.028
-1	der	DET	-3.0	+.029
+2	_?_	<Bnd>	-2.6	+.029
+1	du	PRON	-2.4	+.023
-1	die	DET	-2.3	+.019
-1	ein	DET	-1.8	+.018
+1	ich	PRON	-1.6	+.016
-2	noch	ADV	-1.5	+.013

Determiners

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
+2	_._	<Bnd>	19.7
-2	_<_	<Bnd>	12.6
+2	_?_	<Bnd>	12.1
-1	_<_	<Bnd>	9.2
-1	mit	PREP	6.5
+1	_._	<Bnd>	5.9
-1	in	PREP	5.2
-1	auf	PREP	4.7
-1	ist	V	3.5
-1	noch	ADV	3.4
-2	ist	V	3.1
-2	das	PRON	3.0
+1	_?_	<Bnd>	2.8
-2	du	PRON	2.7
-2	ich	PRON	2.6
-1	die	DET	2.5
-2	wir	PRON	2.4
-2	mal	PTCL	2.3
-1	mal	PTCL	2.3
+2	und	CONJ	2.2
-1	und	CONJ	2.2
+2	leo	N	2.1
-2	die	DET	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	mit	PREP	6.0	-.034
-1	in	PREP	4.7	-.018
-1	auf	PREP	4.5	-.020
+2	_._	<Bnd>	2.9	+.032
-1	noch	ADV	2.2	-.007
-1	ist	V	2.0	-.010
+1	kleinen	ADJ	2.0	-.010
-1	von	PREP	1.7	-.011
-1	an	PTCL	1.7	-.008
+1	seite	N	1.6	-.014
+2	_?_	<Bnd>	1.5	+.020

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	_._	<Bnd>	-12.0	+.076
+1	_?_	<Bnd>	-7.7	+.060
-2	_<_	<Bnd>	-7.4	+.078
-1	_<_	<Bnd>	-6.6	+.068
-1	das	PRON	-2.8	+.026
+1	du	PRON	-2.6	+.026
-1	die	DET	-2.6	+.021
-1	der	DET	-2.5	+.023
+1	das	PRON	-1.7	+.015
+1	mal	PTCL	-1.6	+.015
+1	die	DET	-1.6	+.013
-1	ein	DET	-1.5	+.015

Prepositions

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-1	_<_	<Bnd>	13.9
+2	_._	<Bnd>	12.8
+1	die	DET	9.6
+2	_?_	<Bnd>	9.5
-2	_<_	<Bnd>	9.5
+1	der	DET	8.9
+1	den	DET	7.3
+1	dem	DET	7.2
-2	die	DET	5.5
+1	_._	<Bnd>	4.4
+1	'm	DET	4.2
-1	mal	PTCL	3.6
-2	der	DET	3.5

Context position	Context word	Benchmark category	Preference (in %)
	<continued from left>		
+1	dir	PRON	3.1
+1	'n	PTCL	3.1
-1	nicht	PTCL	3.0
-2	das	PRON	3.0
-2	du	PRON	2.9
-2	ist	V	2.6
+1	_?_	<Bnd>	2.5
-1	hier	ADV	2.3
+1	's	PRON	2.2
+2	bett	N	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	der	DET	8.0	-.012
+1	die	DET	7.9	-.032
+1	dem	DET	7.2	-.010
+1	den	DET	6.8	+.012
+1	'm	DET	4.2	-.021
+1	dir	PRON	2.9	-.020
+1	'n	PTCL	2.8	-.007
-2	die	DET	2.8	+.013
+2	bett	N	2.0	-.015
-1	mal	PTCL	2.0	+.010
-2	der	DET	1.9	+.007
-1	nicht	PTCL	1.8	+.010
+1	's	PRON	1.6	-.003
+2	gegend	N	1.6	-.016
-1	hier	ADV	1.6	+.007

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	_._	<Bnd>	-12.9	+.092
-2	_<_	<Bnd>	-10.2	+.096
+1	_?_	<Bnd>	-7.7	+.059
+2	_._	<Bnd>	-4.2	+.035
-1	die	DET	-3.9	+.044
+1	du	PRON	-2.3	+.021
-1	der	DET	-2.2	+.027
+1	mal	PTCL	-2.0	+.020
-1	das	PRON	-1.9	+.019
-1	ein	DET	-1.7	+.016
+2	das	PRON	-1.6	+.014
-1	_<_	<Bnd>	-1.6	+.078
+1	ist	V	-1.6	+.015
+1	ich	PRON	-1.5	+.015
+2	du	PRON	-1.5	+.013

Conjunctions

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
-1	_<_	<Bnd>	29.5
-2	_<_	<Bnd>	20.2
+1	du	PRON	6.9
+1	die	DET	6.3
+1	_?_	<Bnd>	6.0
+2	_._	<Bnd>	5.1
+1	das	PRON	5.1
+1	wir	PRON	5.0
-1	und	CONJ	4.6
+2	das	PRON	4.6
-2	mal	PTCL	4.3
+2	die	DET	4.1
+2	_?_	<Bnd>	4.1
+1	ich	PRON	3.8

Context position	Context word	Benchmark category	Preference (in %)
<continued from left>			
+2	wir	PRON	3.6
+1	_._	<Bnd>	3.6
+2	ich	PRON	3.1
-2	nicht	PTCL	2.9
+1	der	DET	2.8
-1	nicht	PTCL	2.8
+2	du	PRON	2.6
+1	sie	PRON	2.4
-2	die	DET	2.4
+2	nicht	PTCL	2.3
+2	der	DET	2.1

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	_<_	<Bnd>	14.2	+.047
+1	die	DET	4.5	-.002
+1	du	PRON	4.3	-.013
+1	wir	PRON	3.7	-.009
+1	das	PRON	3.3	-.008
-1	und	CONJ	2.9	-.019
+2	wir	PRON	2.5	-.017
+2	das	PRON	2.5	-.005
-2	mal	PTCL	2.4	-.024
+1	ich	PRON	2.1	.000
+1	sie	PRON	2.0	-.005
+2	die	DET	1.8	+.001
+2	ich	PRON	1.8	-.014
+1	der	DET	1.7	+.001
-1	gucken	V	1.7	-.017
-2	nicht	PTCL	1.7	-.010
+1	man	PRON	1.6	-.010
+1	er	PRON	1.6	-.004
-1	na	INTJ	1.5	-.007
-1	nicht	PTCL	1.5	-.008

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
+1	_._	<Bnd>	-13.8	+.100
+2	_._	<Bnd>	-12.0	+.073
+2	_?_	<Bnd>	-6.7	+.039
-1	die	DET	-4.6	+.045
+1	_?_	<Bnd>	-4.1	+.012
-1	das	PRON	-3.2	+.032
-1	der	DET	-3.1	+.030
-1	ich	PRON	-1.9	+.019
+1	mal	PTCL	-1.8	+.018
-1	ein	DET	-1.6	+.017
+1	leo	N	-1.6	+.014

Particles

Most salient distributional properties

Context position	Context word	Benchmark category	Preference (in %)
+2	_._	<Bnd>	19.8
+1	_._	<Bnd>	14.8
-1	_<_	<Bnd>	11.4
-2	_<_	<Bnd>	10.5
+1	_?_	<Bnd>	10.3
+2	_?_	<Bnd>	7.7
-2	das	PRON	6.5
-1	ist	V	5.6
+1	nicht	PTCL	4.7
-2	ist	V	4.1
-2	die	DET	4.0
-1	ja	INTJ	3.9
-2	ich	PRON	3.4
-2	du	PRON	3.0
-2	der	DET	2.7
-1	du	PRON	2.6
-1	nicht	PTCL	2.5
-1	mal	PTCL	2.4
-1	noch	ADV	2.3
-2	wir	PRON	2.3
+1	noch	ADV	2.2
-1	das	PRON	2.2
+2	die	DET	2.1
-1	wir	PRON	2.0

Dominant positive discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-1	ist	V	4.2	-.001
-2	das	PRON	4.2	-.002
+1	nicht	PTCL	3.9	-.027
+2	_._	<Bnd>	3.0	+.003
-1	ja	INTJ	2.7	-.004
-2	ich	PRON	1.9	-.007

Dominant negative discriminators

Context position	Context word	Benchmark category	Relative preference	Discriminative power
-2	_<_	<Bnd>	-9.5	+.069
-1	_<_	<Bnd>	-4.3	+.039
-1	die	DET	-3.9	+.040
+2	_?_	<Bnd>	-3.1	+.021
-1	der	DET	-2.7	+.027
+1	_._	<Bnd>	-2.5	+.010
+1	du	PRON	-2.1	+.021
-2	und	CONJ	-1.8	+.014

Appendix E Individual preferences of verb subclasses

The following two pages list for each grammatical subclass of the verb category that was assigned at least 10 member target words, its 20 most salient distributional properties at the level of individual context words (sorted by descending preference). Preference values are given as percentages. Context words are always listed together with their category (according to the benchmark classification) to facilitate relating these tables to Table 4-9 in subsection 4.4.1. The four utterance boundary markers (as virtual context words) are represented by the same symbols that were introduced earlier (cf. p. 60): the symbol _<_ as the pre-utterance marker, and the symbols _._, _?_, and _!_ as the three post-utterance markers (matching the three possibilities of utterance-terminal punctuation). As in Table 4-9, the symbol <Bnd> is used to provide these four virtual context words with a category specification.

Infinitives

Context position	Context word	Benchmark category	Preference (in %)
+1	_._	<Bnd>	24.5
+1	_?_	<Bnd>	20.5
+2	_._	<Bnd>	18.5
+2	_?_	<Bnd>	13.2
-2	_<_	<Bnd>	11.4
+1	wir	PRON	6.4
-1	mal	PTCL	6.3
-1	_<_	<Bnd>	5.6
-2	die	DET	5.3
+1	leo	N	4.2
-1	die	DET	4.0
-1	wir	PRON	3.7
-1	noch	ADV	3.1
-1	nicht	PTCL	3.1
+1	die	DET	3.0
-2	wir	PRON	3.0
+2	die	DET	3.0
+2	das	PRON	2.9
-1	zu	PTCL	2.9
+2	du	PRON	2.8

Past participles

Context position	Context word	Benchmark category	Preference (in %)
+1	_?_	<Bnd>	27.6
+2	_._	<Bnd>	24.3
+1	_._	<Bnd>	22.7
+2	_?_	<Bnd>	20.0
+1	leo	N	4.5
-2	die	DET	4.2
-2	du	PRON	3.8
-1	nicht	PTCL	3.6
-2	_<_	<Bnd>	3.6
-1	das	PRON	3.4
-1	du	PRON	3.3
-2	ich	PRON	3.1
-1	da	ADV	3.1
-2	der	DET	3.0
-2	hast	V	3.0
-2	ist	V	2.9
-2	hat	V	2.7
-1	ist	V	2.7
-1	was	INTG	2.6
-2	das	PRON	2.5

Imperatives singular

Context position	Context word	Benchmark category	Preference (in %)
-2	_<_	<Bnd>	46.5
+1	mal	PTCL	38.4
-1	_<_	<Bnd>	38.1
+2	mal	PTCL	15.4
-1	dann	CONJ	13.3
+2	_._	<Bnd>	10.7
-1	ich	PRON	10.1
+1	ich	PRON	8.3
-1	leo	N	5.7
+1	's	PRON	4.9
-2	na	INTJ	4.6
-1	jetzt	ADV	4.3
-2	ja	INTJ	4.3
+1	auf	PREP	3.4
+1	die	DET	3.4
-2	und	CONJ	3.4
+1	dich	PRON	3.2
+2	_!_	<Bnd>	3.2
+2	die	DET	3.2
+1	mir	PRON	3.1

First person singular verbs forms

Context position	Context word	Benchmark category	Preference (in %)
-1	ich	PRON	48.7
-2	_<_	<Bnd>	42.6
+1	ich	PRON	37.4
-1	_<_	<Bnd>	11.1
+2	_._	<Bnd>	9.6
-1	das	PRON	7.5
+2	nicht	PTCL	6.9
-2	ja	INTJ	6.9
+2	auch	PTCL	5.5
+1	das	PRON	5.4
+1	_._	<Bnd>	4.8
+2	mal	PTCL	4.0
+1	mal	PTCL	3.9
-2	und	CONJ	3.8
-1	dann	CONJ	3.6
+1	die	DET	3.0
-2	leo	N	3.0
-1	jetzt	ADV	2.9
+2	das	PRON	2.7
+2	ja	INTJ	2.7

Second person singular verb forms

Context position	Context word	Benchmark category	Preference (in %)
+1	du	PRON	60.6
-2	_<_	<Bnd>	44.5
-1	_<_	<Bnd>	24.9
-1	du	PRON	22.6
-1	was	INTG	10.1
-2	und	CONJ	7.2
+1	_._	<Bnd>	4.8
+2	das	PRON	4.6
+2	_?_	<Bnd>	4.5
+2	die	DET	3.9
+2	noch	ADV	3.8
-1	das	PRON	3.8
+2	auch	PTCL	3.7
-2	du	PRON	3.7
+2	nicht	PTCL	3.4
-1	da	ADV	3.3
+2	denn	PTCL	3.2
+2	mal	PTCL	3.1
+2	_._	<Bnd>	3.1
-2	leo	N	3.0

Third person singular verb forms

Context position	Context word	Benchmark category	Preference (in %)
-2	_<_	<Bnd>	38.9
-1	das	PRON	14.2
-1	_<_	<Bnd>	10.8
-2	und	CONJ	7.9
+1	_._	<Bnd>	7.3
+2	_._	<Bnd>	7.0
+1	das	PRON	6.5
-1	was	INTG	6.3
+1	der	DET	5.8
-1	da	ADV	5.8
-1	der	DET	5.5
+2	_?_	<Bnd>	5.5
+1	's	PRON	4.4
+1	die	DET	4.0
-1	ich	PRON	3.6
-1	die	DET	3.6
+2	auch	PTCL	3.6
-1	dann	CONJ	3.5
+2	nicht	PTCL	3.5
-2	der	DET	3.5

First person plural verb forms

Context position	Context word	Benchmark category	Preference (in %)
+1	wir	PRON	47.3
-2	_<_	<Bnd>	36.3
-1	wir	PRON	19.2
-1	_<_	<Bnd>	15.6
+2	mal	PTCL	8.8
-1	dann	CONJ	6.7
-1	die	DET	6.7
-2	und	CONJ	6.6
-1	da	ADV	5.9
+2	noch	ADV	5.9
+1	_._	<Bnd>	5.5
+1	die	DET	5.5
+2	die	DET	4.7
-1	das	PRON	4.2
+2	auch	PTCL	3.8
-1	jetzt	ADV	3.7
-1	was	INTG	3.7
+2	das	PRON	3.6
-2	die	DET	3.2
+1	mal	PTCL	3.1

Appendix F Individual preferences of noun subclasses

This and the following page present for each grammatical subclass of the noun category its 20 most salient distributional properties at the level of individual context words (sorted by descending preference). Preference values are given as percentages. Context words are always listed together with their category (according to the benchmark classification) to facilitate relating these tables to Table 4-11 in subsection 4.4.2. The four utterance boundary markers (as virtual context words) are represented by the same symbols that were introduced earlier (cf. p. 60): the symbol _<_ as the pre-utterance marker, and the symbols _._, _?_, and _!_ as the three post-utterance markers (matching the three possibilities of utterance-terminal punctuation). As in Table 4-11, the symbol <Bnd> is used to provide these four virtual context words with a category specification.

Feminine singular nouns

Context position	Context word	Benchmark category	Preference (in %)
-1	die	DET	31.2
+2	_._	<Bnd>	19.5
+1	_._	<Bnd>	18.6
+1	_?_	<Bnd>	15.1
+2	_?_	<Bnd>	14.7
-2	_<_	<Bnd>	12.4
-1	der	DET	12.1
-1	eine	DET	8.8
-2	in	PREP	6.8
-1	'ne	DET	5.6
-2	ist	V	4.5
-2	die	DET	4.3
-1	keine	DET	4.2
-1	_<_	<Bnd>	3.8
+1	ist	V	3.6
-2	mit	PREP	3.6
-2	eine	DET	3.1
-2	noch	ADV	2.9
-2	durch	PREP	2.6
-2	auf	PREP	2.5

Masculine singular nouns

Context position	Context word	Benchmark category	Preference (in %)
+2	_._	<Bnd>	19.3
+1	_._	<Bnd>	18.8
+2	_?_	<Bnd>	17.3
+1	_?_	<Bnd>	16.9
-1	der	DET	16.6
-2	_<_	<Bnd>	11.4
-1	den	DET	9.3
-1	'n	PTCL	8.4
-2	auf	PREP	7.1
-1	_<_	<Bnd>	5.6
-1	ein	DET	5.4
-2	ist	V	4.8
-1	dem	DET	4.7
-1	einen	DET	4.5
-1	im	DET	3.7
-1	'm	DET	3.5
-2	noch	ADV	3.3
-2	mit	PREP	3.3
-1	zum	DET	3.2
+1	ist	V	2.7

Neuter singular nouns

Context position	Context word	Benchmark category	Preference (in %)
-1	das	PRON	22.3
+1	_._	<Bnd>	20.6
+2	_._	<Bnd>	18.4
+2	_?_	<Bnd>	16.3
+1	_?_	<Bnd>	15.1
-1	ein	DET	13.7
-2	_<_	<Bnd>	10.6
-1	'n	PTCL	5.5
-1	dem	DET	4.9
-1	_<_	<Bnd>	4.8
-2	auf	PREP	4.8
-2	ist	V	4.7
-1	's	PRON	4.5
-2	noch	ADV	3.9
-2	das	PRON	3.7
-2	ein	DET	3.6
-1	zum	DET	3.6
-2	mit	PREP	3.2
-2	in	PREP	3.1
+1	ist	V	3.0

Plural noun forms

Context position	Context word	Benchmark category	Preference (in %)
-1	die	DET	27.7
+2	_._	<Bnd>	19.5
+2	_?_	<Bnd>	15.0
+1	_._	<Bnd>	14.6
+1	_?_	<Bnd>	14.0
-2	_<_	<Bnd>	10.7
-1	_<_	<Bnd>	7.8
-2	noch	ADV	5.6
-2	die	DET	4.7
-1	deine	DET	4.0
-2	sind	V	4.0
-2	mal	PTCL	3.1
-1	ein+paar	DET	3.0
+2	die	DET	2.8
+1	sind	V	2.8
-2	auch	PTCL	2.5
-2	du	PRON	2.5
-1	keine	DET	2.4
+2	leo	N	2.4
-1	viele	DET	2.4

Proper names for individuals

Context position	Context word	Benchmark category	Preference (in %)
-2	_<_	<Bnd>	19.4
+1	_._	<Bnd>	16.6
-1	_<_	<Bnd>	15.5
+1	_?_	<Bnd>	13.1
+2	_?_	<Bnd>	12.0
-1	der	DET	10.9
-1	die	DET	10.1
+2	_._	<Bnd>	9.3
-1	und	CONJ	7.0
+1	und	CONJ	5.9
-2	was	INTG	4.1
+1	ist	V	4.0
+1	hat	V	3.7
-1	mit	PREP	3.6
-2	du	PRON	2.7
-2	ist	V	2.6
-2	mal	PTCL	2.6
-2	mit	PREP	2.6
+2	auch	PTCL	2.5
+2	mal	PTCL	2.5

Proper names for places

Context position	Context word	Benchmark category	Preference (in %)
-1	nach	PREP	36.1
+1	_._	<Bnd>	22.7
+1	_?_	<Bnd>	21.8
+2	_._	<Bnd>	21.4
-1	in	PREP	19.3
+2	_?_	<Bnd>	17.6
-2	_<_	<Bnd>	14.5
-1	ist	V	6.3
-1	der	DET	5.7
-1	_<_	<Bnd>	5.7
+1	ist	V	5.7
-2	an	PTCL	5.6
+1	fahren	V	4.6
-2	da	ADV	4.2
-1	die	DET	4.2
-1	von	PREP	4.2
-2	ist	V	3.7
-2	wieder	ADV	3.5
-2	faehrt	V	3.4
-1	zur	DET	3.2

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Frage, wie Kinder die Wortarten ihrer Muttersprache erwerben und welchen Einfluss ihre Erfahrung mit Sprache dabei hat. Im Mittelpunkt stehen einfache distributionelle Regelmäßigkeiten, die Kinder auch ohne syntaktisches Vorwissen in ihrem Sprachinput vorfinden. Mithilfe automatischer Verfahren wird untersucht, wie informativ diese Regelmäßigkeiten hinsichtlich der verschiedenen Wortarten sind. Die Datengrundlage hierfür liefert ein umfangreiches Korpus, das die sprachliche Interaktion eines deutschsprachig aufwachsenden Kindes mit seiner Umgebung über einen Zeitraum von drei Jahren mit einer außergewöhnlich hohen Stichprobendichte dokumentiert.

Kapitel 1 umreißt den Hintergrund der Fragestellung. Zunächst werden der linguistische und kognitive Status von Wortarten in der Erwachsenensprache skizziert und zwei zentrale Debatten um die Rolle von Biologie und Erfahrung beim Kategorienerwerb zusammengefasst, die insbesondere das Interesse an distributionellen Regelmäßigkeiten begründen. Anschließend wird das hier benutzte Konzept von distributioneller Information gegen alternative Definitionen abgegrenzt und die Zielsetzung der vorliegenden Arbeit mit früheren Studien verglichen, die einen ähnlichen Distributionsbegriff verwenden.

Das zweite Kapitel beschreibt das o.g. Korpus sowie ein System von elf *Benchmark-Kategorien*, die als grobe Approximationen der tatsächlich erworbenen Wortarten konzipiert sind und hier als Heuristiken verwendet wurden, um den potentiellen Nutzen der vom Korpus extrahierten distributionellen Regelmäßigkeiten für den Wortartenerwerb zu bewerten.

In Kapitel 3 werden die formalen Methoden vorgestellt, mit denen distributionelle Regelmäßigkeiten aus dem Input extrahiert und ausgewertet wurden. Im Mittelpunkt steht ein kookkurrenz-statistisches Modell, welches für jedes Wort seine distributionellen Eigenschaften im Korpus bestimmt: Diese erfassen Spektrum und Häufigkeit der lokalen lexikalischen Kontexte, in denen dieses Wort im Korpus auftritt.

Dem Kookkurrenz-Ansatz liegt die Hypothese zugrunde, dass Wörter, die mit hoher Frequenz in ähnlichen lokalen lexikalischen Kontexten auftreten und somit ähnliche distributionelle Eigenschaften haben, wahrscheinlich zu derselben Wortart gehören. Um diese Hypothese zu formalisieren, wird eine mathematische Metrik beschrieben, die für zwei beliebige Wörter quantifiziert, wie sehr sich ihre distributionellen Eigenschaften ähneln. Abschließend werden mehrere Auswertungsmaße eingeführt, die formal bewerten, zu welchem Grad die Hypothese für die verschiedenen Wortarten (genauer: Benchmark-Kategorien) zutrifft.

Kapitel 4 präsentiert umfangreiche Analysen der aus dem Korpus extrahierten distributionellen Information. Zunächst wird ihr Informationsgehalt für den Erwerb der einzelnen Wortarten untersucht und durch Kontrollanalysen gegen mögliche Artefakte abgesichert. Die Ergebnisse zeigen, dass der deutschsprachige Input insgesamt sehr informative distributionelle Regelmäßigkeiten enthält. Alle elf Benchmark-Kategorien profitieren von dieser Information, v.a. jedoch Interjektionen, Fragewörter und Substantive. Adverbien und Partikeln hingegen lassen sich aufgrund ihrer distributionellen Eigenschaften praktisch nicht voneinander unterscheiden.

Ausgehend von Überlegungen zur Erwerbssituation des Kindes wird anschließend der Umfang bzw. die Qualität der extrahierten distributionellen Information in mehreren Experimenten sukzessive vermindert. Zusammenfassend lässt sich festhalten, dass ihr Nutzen für den Wortartenerwerb dabei nur graduell abnimmt und sich in Einzelfällen sogar verbessert. Dies unterstreicht, dass die Ergebnisse nur in geringem Maße von bestimmten Eigenschaften des Kookkurrenz-Modells abhängen und somit zuverlässige Rückschlüsse auf den Input zumindest dieses einen Kindes zulassen.

Dies ermöglicht nun, für jede Wortart ihre charakteristischen distributionellen Eigenschaften zu bestimmen. Anhand dieser detaillierten *distributionellen Profile* lassen sich die Gemeinsamkeiten und Unterschiede zwischen den einzelnen Wortarten benennen. Für Verben und Substantive werden diese Analysen noch weiter vertieft. Das bedeutendste Resultat hierbei ist, dass finite und infinite Verbformen sehr unterschiedliche distributionelle Eigenschaften aufweisen, Substantive aller Art dagegen eine Reihe distributioneller Eigenschaften teilen. Es wird angedeutet, wie sich dieser grundsätzliche distributionelle Unterschied zwischen Substantiven und Verben sehr wahrscheinlich auf syntaktische Strukturen und stark ausgeprägte Gebrauchs-präferenzen zurückführen lässt. Ferner ergeben sich aus dem Befund zwei testbare Vorhersagen über den Verlauf des Wortartenerwerbs im Deutschen.

In zwei abschließenden Experimenten werden psycholinguistische Befunde zum Wortartenerwerb in das Kookkurrenz-Modell integriert. Die Ergebnisse deuten u.a. darauf hin, dass die frühe Lernsituation des Kindes den Erwerb der Substantivkategorie und anschließend der Kategorien Artikelwort und Präposition unterstützt.

Im Schlusskapitel wird diskutiert, welchen Beitrag eine Fallstudie, die zudem nur den Input untersucht, für die Erforschung des Wortartenerwerbs leisten kann. Der wichtigste Beitrag dieser Studie ist, dass erstmals systematisch für das Deutsche gezeigt wurde, dass der Sprachinput eines Kindes informative distributionelle Hinweise auf die meisten Wortarten enthält. Theoretische Einwände, die gegen den grundsätzlichen Nutzen von distributionellen Ansätzen für den Wortartenerwerb vorgebracht wurden, können aufgrund der empirischen Ergebnisse aus dieser Arbeit weitgehend entkräftet werden, was einige Stärken des hier verwendeten Distributionsbegriffs unterstreicht.

Psycholinguistische Evidenzen sprechen dafür, dass sich Kinder beim Wortartenerwerb tatsächlich die hier untersuchte distributionelle Information zunutze machen. Daher werden anschließend wesentliche Eigenschaften eines möglichen Lernmechanismus charakterisiert, und es wird angedeutet, wie die gewonnenen Ergebnisse indirekt zu neuen Erkenntnissen über diesen Lernmechanismus beitragen können.

Abschließend werden die Grenzen dieser Arbeit benannt. Sie betreffen neben Schwächen des verwendeten Distributionsbegriffs auch wichtige Unterschiede hinsichtlich der Lernsituation von Kind und Modell, Mängel der Benchmark-Kategorien als Approximationen für die Wortarten erwachsener Sprecher und schließlich prinzipielle Beschränkungen von distributioneller Information an sich. Zuletzt werden Forschungsansätze diskutiert, mit denen sich diese Schwierigkeiten möglicherweise minimieren lassen.